Environ Res Tec, Vol. 8, Issue. 4, pp. 809-822, December 2025

Environmental Research and Technology
https://dergipark.org.tr/en/pub/ert
DOI: https://doi.org/10.35208/ert.1587308

Environmental Research & Technology

Research Article

Spatiotemporal analysis and machine learning-based prediction of air quality in Indian urban cities

Sitesh Kumar SINGH[1], Rituraj JAIN[2*], Damodharan PALANIAPPAN[2], Kumar PARMAR[2], Premavathi T.[3], Jaishri GOTHANIA[4]

[1]National University of Science and Technology, Department of Civil and Environmental Engineering, College of Engineering, Muscat, Sultanate of Oman
[2] Marwadi University, Department of Information Technology, Rajkot, Gujarat, India
[3] Marwadi University, Department of Computer Engineering - AI, Rajkot, Gujarat, India
[4] Lingaya's Vidyapeeth, Department of Computer Science Engineering, Faridabad, India

ARTICLE INFO

Article history
Received: 18 November 2024
Revised: 09 December 2024
Accepted: 12 December 2024

Key words:
Air quality prediction, gradient boosting, machine learning models, particulate matter (PM2.5), random forest, spatiotemporal analysis

ABSTRACT

Air pollution, more specifically Particulate Matter (PM2.5 - particulate matter with diameter less than 2.5 micrometers), threatens the public health most critically in urban Indian cities, and Delhi, among them, presents the most acute challenge. This study predicts the concentrations of PM2.5 using machine learning models using data ranging from 2010 to 2023 and assessing model fit via R^2, RMSE, MAE, and MAPE metrics. Models tested: Random Forest, Gradient Boosting, AdaBoost, Histogram-Based Gradient Boosting, XGBoost. The Random Forest model is extremely effective for the training set (R^2 = 0.99) but shows the highest degree of overfitting, with R^2 of 0.35 for the test set. Gradient Boosting has a more balanced result, with R^2 0.54 and 0.48, respectively on the training and test set, as well as fewer errors (RMSE: 56.46, MAE: 39.60, MAPE: 0.50). Hence, it is a good predictor. AdaBoost performs the worst with an R^2 of 0.28 on the test set and the highest errors in terms of RMSE: 66.86, MAE: 52.34, MAPE: 0.94. Histogram Gradient Boosting and XGBoost: both of these models yield an average accuracy value, but the Gradient Boosting model is still a tad better than the former ones in terms of RMSE and MAE. Thus, Gradient Boosting happens to be the most accurate model in light of generalization as well as accuracy for the prediction of the concentration of PM2.5. These results will be highly beneficial to policymakers to adopt machine learning-based air quality forecasting for better environmental management and the protection of public health.

Cite this article as: Singh SK, Jain R, Palaniappan D, Parmar K, Premavathi T, Gothania J. Spatiotemporal analysis and machine learning-based prediction of air quality in Indian urban cities. Environ Res Tec 2025;8(4) 809-822.

INTRODUCTION

One of the strongest concerns associated with urban environments is its air quality, especially in developing countries like India. Intensification of industrial and urbanization processes yields massive increases in levels of air pollution. According to reports from the World Health Organization, air pollution takes millions of lives due to premature deaths, and cities in India often feature prominently in rankings as most polluted. It is important to watch for this trend by predicting the quality of air in order to safeguard public health and inform policy action. Analysis of spatiotemporal data is important for understanding the dynamic traits of air quality within the Indian urban domain. This provides spatial and temporal information with respect to explaining the prevailing trends, patterns, and hotspots of pollution in relation to variables such as seasonal flux and urban expansion. It inte-

***Corresponding author.**
*E-mail address: jainrituraj@yahoo.com

grates data sources to create comprehensive maps and time series that illustrate pollutant spreading, enables the identification of suboptimal areas of air quality, and assesses management strategies. This analysis delved into the relationship between air quality and a chain of environmental and anthropogenic variables, thus providing much-needed knowledge for targeted interventions and evidence-based policymaking in rapidly urbanizing Indian cities. Combining these analytical methods with monitoring efforts will help policymakers and urban planners develop effective models of prediction for air quality and management strategies toward more sustainable urban environments that center on public health.

Recent advancements in machine learning (ML) and deep learning (DL) techniques have thrown open new avenues in air quality analysis and prediction. Different studies are proving the efficiency of this approach in the medium-range prediction of concentrations of different air pollutants, including PM2.5, PM10 (particulate matter less than 10 micrometers in diameter), NOx, and O3, by using past data and meteorological parameters. For instance, Support Vector Regression has been utilized promisingly to make high-accuracy forecasts of air quality indices (AQI). Furthermore, Long Short-Term Memory networks have been employed in an effort to successfully capture the time dependencies between air quality data, thus enhancing the predictive capabilities of traditional statistical models [1-4].

Although there have been great strides in the very recent past toward harnessing machine learning for air-quality prediction, gaps still exist in the current landscape of research. Most of the studies focused on short-term predictions or relied on data from a few monitoring stations, limiting their generalizability. This narrow focus usually misses the much-needed interplay between geographical and temporal variations in air pollution, thereby stalling the development of comprehensive predictive models [1, 5]. However, big cities continue facing hazardous levels of air pollution, notably during winter when adverse meteorological conditions trap pollutants near the ground. Delhi takes one of the top places among other polluted cities in the world's list, so determining the spatiotemporal dynamics of air pollution and making predictions about the expected trends is very important. This challenge requires innovative approaches that combine historical data with predictive modeling techniques to provide actionable insights [5-7].

This paper seeks to utilize spatiotemporal analysis and machine learning models in the prediction of air quality based on a robust dataset from 2010 to 2023. Advanced ML techniques such as Random Forest, Gradient Boosting, AdaBoost, and XGBoost are used for predicting future concentration of the key pollutants in particular PM2.5 as it possesses the highest adverse health effects. The suggested model encompasses feature engineering with lag variables in combination with meteorological parameters to extract seasonal and long-term patterns of pollution in the air. Using cross-validation and an optimization procedure for the hyperparameters, this system expects to provide a more accurate prediction regarding the trend of air quality for the year 2024. The results of the present research complement the available body of knowledge about air quality management in India. The results of this study will be useful in setting up a predictive framework for future research and arm policymakers with data-driven insights to mitigate the public health impacts of air pollution in Indian urban cities.

LITERATURE REVIEW

Existing studies on the prediction of air quality applied various machine learning methods with no exception to unearth their applicability in dealing well with the specifics in this problem. Furthermore, deep learning technologies have demonstrated impressive performance for predicting urban air quality models. Here, both Convolutional Neural Networks and Long Short-Term Memory achieved the state-of-the-art results [8]. These models can better determine spatial and temporal dependencies in air pollution data, hence making it possible to obtain higher predictive accuracy.

The hybrid approach was proposed by [3], which includes the combination of Discretized Regression and Least Square Support Vector Machines for air pollution prediction. Their method sought to handle the problem of missing data and provides the necessary accuracy in estimating the quality of the air. The study of [9] proposed an extensive review of applications of machine learning algorithms, especially regression techniques and deep learning models, for air quality forecasting. Authors brought out several methodologies in terms of strengths and weaknesses, thus giving an idea of which models are there based on a particular requirement concerning the problem at hand [2]. Many researchers in the Indian scenario have studied the use of machine learning for predicting air quality in cities such as Delhi, Kolkata, and Bangalore.

Spatiotemporal Analysis of Air Pollution

Spatiotemporal analysis becomes very important in observing how variations of air pollution occur in both space and time. Many works have implemented spatial and temporal models that aimed to capture changes in pollutant concentrations and pinpoint high-risk zones in urban areas. Such as [10], which integrated a spatial interpolation method with temporal models for mapping levels of PM2.5 and NOx across various cities in China. This approach provided an overall view of pollution patterns and helped identify the sources of pollution. Similar to this, [11] reviewed seasonal air quality trends in Indian cities, showing the vital need to understand the spatiotemporal variations in designing interventions.

Air pollution in Indian cities and seasons varies significantly due to climate, industrialized activities, and transport patterns. Studies like [12] and [13] have pointed out how pollution peaks during the winter months in cities like Delhi often due to crop burning and lower wind speed and higher vehicle emission. These spatiotemporal dynamics highlight the need for predictive models sensitive to seasonal and location-based variations in air quality to invoke pro-active measures.

Machine Learning and Air Quality Prediction

Machine learning (ML) is the latest tool used in making

air-quality forecasts because conventional statistical models fail to adequately explain the complex, nonlinear effects of air pollutants caused by other factors, including atmospheric conditions. Increasing environmental databases would facilitate research where ML could be trained in both forecasted pollutant concentrations and on the identification of the driving factors behind the air quality. In [3] exposed that, particularly, models like Support Vector Regression and Random Forests are capable of precise predictions of air pollutant levels compared to the traditional models in accuracy and speed. In a similar fashion, [14] also used LSTM networks in dependency analysis along the temporal lines found in the air quality data. In this manner, deep learning has presented some interesting results for long-term forecasting. The LSTM networks fall under the category of the subtypes of recurrent neural networks and have emerged as a significant tool in the domain of researching predictions regarding air quality. These networks excel in extracting temporal dependencies within air quality datasets, thereby particularly showing excellence in long-term forecasting efforts. In addition, LSTMs have been proven to enhance the predictive capability of statistical models by accurately modeling complicated time-related dependencies involved in air quality information.

Many machine learning algorithms have been applied to air quality prediction; each has its merits and demerits. These include decision trees, gradient boosting, and also ensemble techniques like Random Forests and XGBoost, generally much stronger at handling high-dimensional data and nonlinear relations among variables. Of late, a number of studies by [5] and [15] have depicted that these models help in predictions for the accurate concentration level of PM2.5-PM10, NOx, and O_3. These models utilize historical data related to air pollution, besides meteorological variables such as temperature, humidity, and wind speed to gain a detailed view of pollution dynamics.

In the Indian context, several studies have concentrated on air quality prediction using machine learning. [16] implemented an in-depth analysis of pollution data from 23 Indian cities by using regression techniques and feature selection methods to improve model performance. Their findings, in fact, show that the models could predict AQI well and illustrated the influence of external factors such as the increase in traffic emissions and weather patterns on the levels of pollution. In addition, usage of IoT sensors has also enhanced the ability for real-time monitoring of air quality. IoT sensor data has been found effective when used for integrating with machine learning models, because such integration can enhance timeliness and accuracy of the predictions made [15].

Air Pollution in Indian Cities

Indian cities, especially Delhi, Mumbai, and Kolkata, are ranked to be some of the most polluted cities in the world; in fact, many extreme pollution events are a result of seasonal trends. Delhi severely degrades air quality during winters due to vehicular emissions, industrial activities, and agricultural residue burning in neighbouring states [17]. Studies have found the repeated violations of WHO-promoted safe air standards all over Delhi, making it an appropriate site for sophisticated modeling techniques intended to predict and control peaking levels of pollution.

However, one recent study reported machine learning models capable of predicting air quality trends in Indian cities using historical pollution data and meteorological variables as well as emission sources. For instance, [18] and [19] applied Random Forest and Gradient Boosting algorithms to predict the level of PM2.5 in Delhi, indicating a high accuracy and the presence of meteorological conditions influencing the dynamics of air quality. These studies give an importance to the application of machine learning techniques to air quality issues in Indian urban cities because they form a platform for unique challenges.

Gaps and Future Directions

Although there is significant advancement in the application of machine learning to air quality prediction, an important gap still exists in the literature. For one, most current studies focus on short-term predictions or use data from a few monitoring stations only, with generalizability highly compromised. For another, hardly any study has made it possible to integrate spatiotemporal analysis with machine learning, giving less attention to simultaneous geographic and temporal variations in air pollution. Finally, although promising results are seen by the machine learning models in capturing extreme events of pollution going critical to public health interventions, there are scopes to improve model accuracy.

To fill these gaps, this study combines spatiotemporal analysis with advanced machine learning techniques for predicting air quality in urban areas of India. It focuses on one of the most polluted cities in the world, Delhi. Proposed system attempts to create a framework of prediction such that it can predict the levels of key pollutants at high accuracy and incorporates temporal patterns as well as spatial variations. It will, therefore, include meteorological data and external factors for an integrated approach towards the pollution dynamics giving way to actionable insights for policy direction. This will not only bridge gaps in the existing literature but also offer a scalable solution towards future studies on air quality management in other Indian cities.

METHODOLOGY

This research is a spatiotemporal study combined with machine learning for the prediction of air quality, integrating location and time. The study is based on the urban cities of India from the year 2010 to 2023. The whole methodology involves key steps such as collection of data, preprocessing, feature engineering, model selection, and training and evaluation with subsequent predictions of air quality, as shown in Figure 1. The main goal is to develop robust models of PM2.5 and other important pollutants for more informed decision-making regarding air quality management.

Mathematical Formulation for the Proposed System

Data preprocessing and feature engineering

The proposed spatiotemporal air quality predictive model starts with data preprocessing and feature engineering. The

time stamp extracted in the dataset provides relevant temporal features, such as hour, day, month, and year. Lag features, which include pollutant levels from one and two years ago, have been established to set up the level of dependency between these temporal features. Next, data groupings based on different time steps have been made (daily, monthly, yearly) to analyze the trends and correlations with target pollutants, especially PM2.5.

Let the dataset be represented as:

$$D = \{(X_i, y_i) | i = 1, 2, \ldots, n\} \quad (1)$$

Where:

$X_i \in R^m$ is the feature vector for each observation.

y_i is the target variable (e.g., air pollutant concentration, such as PM2.5)

n is the total number of observations.

m is the number of features.

Datetime Features: To capture the temporal dynamics, various datetime-related features are extracted from the timestamp T:

$$X_i = \{hour(T), day(T), month(T), year(T), \ldots.\} \quad (2)$$

Where:

hour(T)	represents the hour of the day.
day(T)	is the day of the month.
month(T)	is the month.
year(T)	is the year.

Lag Features: Lag features for 1 year and 2 years are generated to capture the temporal dependencies:

$$pm\_lag\_1Y(i) = y_i - 365 * 24, pm\_lag\_2Y(i) = y_i - 730 * 24 \quad (3)$$

Where $pm\_lag\_1Y(i)$ and $pm\_lag\_2Y(i)$ represent the lagged values of the target pollutant for 1 year and 2 years, respectively.

Spatiotemporal Grouping and Aggregation: For spatiotemporal analysis, the data is grouped at different time intervals to study trends:

Let G represent the spatiotemporal groups:

$$G = \{daily, monthly, yearly\} \quad (4)$$

The data is aggregated using the mean:

$$MeanPollu\,tan\,t\,(G) = \frac{1}{|G|}\sum_{i \in G} y_i \quad (5)$$

Correlation matrix and feature selection

By correlating all the input variables in a matrix, the most critical features for prediction are determined. To select relevant features, the correlation of the target pollutant with other features is calculated:

$$Corr(X_j, y) = \frac{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (6)$$

Where $Corr(X_j,y)$ is the Pearson correlation coefficient between feature X_j and the target y, and X_j and $\bar{y}$ are the mean values of X_j and y, respectively.

Prediction model

Thezlevels of various pollutants are predicted by building Random Forest, Gradient Boosting, and XGBoost models. Each of these has its own unique strengths. A Random Forest Regressor is an ensemble method that creates many decision trees and averages their prediction for more accuracy in reducing overfitting. They work well with large datasets and high-dimensional data, making them reliable and straightforward to interpret. It may become computationally expensive when the data is sparse and less effective. Random Forests reduces the effect of overfitting of the model to specific data instances and noise through the creation of many subsets of the training dataset using a method called bagging or bootstrap aggregation. Only a random subset of features was considered in every split of the decision trees. This leads to diversity and reduces the model's reliance on individual predictors. This average of multiple decision tree final predictions helps smoothen out the errors of individual trees and hence reduces the variance. In real-world applications, Random Forests handles overfitting with out-of-bag error estimation. Such an estimation is unbiased with respect to the performance of the model without requiring a separate validation set, which leads to a more suitable tuning of hyperparameters. The Gradient Boosting Regressor uses an approach to build a sequence of decision trees. Each tree is used to rectify the past ones in terms of errors. It iterates well enough to capture the intricate data patterns and obtain highly accurate solutions. Although it performs pretty well in modeling non-linear relationships, Gradient Boosting, however, is more prone to overfitting, and it is generally slower to train than the Random Forest. It requires careful tuning of its parameters to avoid any performance issues. XGBoost is a regularized version of Gradient Boosting, which overcomes many inefficiencies since techniques for regularization are used to avoid overfitting as it would speed up the process. XGBoost provides fantastic results and is the key parameter of use in structured data problems and competition within machine learning, but is rather complex when compared to both Random Forest or simple Gradient Boosting, yet it can be quite versatile as it runs fast and also works well even on missing data or imbalanced datasets. Although XGBoost has a good number of strengths, the approach itself still requires careful tuning to prevent overfitting, and it becomes fairly complex to implement. Various machine learning models $f(X;\theta)$ have used to predict the air quality, where θ represents the model parameters:

$$\hat{y}_i = f(X_i; \theta) \quad (7)$$

The models considered include:

Random Forest Regressor:

$$f_{RF}(X; \theta) = \frac{1}{T}\sum_{t=1}^{T} h_t(X; \theta_t) \quad (8)$$

Where each h_t is a decision tree.

Gradient Boosting Regressor:

$$f_{GB}(X; \theta) = \sum_{m=1}^{M} \gamma_m h_m(X; \theta_m) \quad (9)$$

Where h_m are weak learners, and γ_m are the weights.

XGBoost (Extreme Gradient Boosting):

$$f_{XGB}(X; \theta) = \sum_{k=1}^{K} h_k(X; \theta_k) + \lambda \parallel \theta_k \parallel^2 \quad (10)$$

Where $\lambda \parallel \theta_k \parallel^2$ is the regularization term.

Model training and evaluation

All the models that were trained on the data and analyzed for the use of RMSE, MAE, and R2. The models are trained using a training set $\{X_{train}, y_{train}\}$, and the predictions are evaluated on a test set $\{X_{test}, y_{test}\}$.

The performance is measured using several metrics:

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (11)$$

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (12)$$

R-squared (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (13)$$

Cross-validation and hyperparameter tuning

The process involves tuning hyperparameters for models using randomized search together with time series cross-validation. This targets optimization of model performance. Randomized search is used with cross-validation. The objective is to minimize the RMSE by using k-fold time series cross-validation. This study used k-fold time-series cross-validation, which is crafted with consideration for the temporal nature of time-series data. As such, this means the training set will always appear before the validation set so that the data remains intact. This is contrary to traditional k-fold cross-validation, which is designed around independent and identically distributed data points. The split here preserves the temporal nature of the data so that no future results can predict earlier results.

$$\theta^* = arg\,\min_{\theta} \frac{1}{k}\sum_{j=1}^{k} RMSE(y_{val}^{(j)}, f(X_{val}^{(j)}; \theta)) \quad (14)$$

Where $X_{val}^{(j)}$, $y_{val}^{(j)}$ are the validation splits.

Prediction on future data

These models are then used to forecast air quality over time within datasets across future time points. The effectiveness of the models is validated by comparing the predicted values against the actual pollutant concentrations. For future air quality predictions, a dataset X_future is created by extrapolating datetime features. The future values are predicted as:

$$\hat{y}_{future} = f(X_{future}; \theta^*) \quad (15)$$

The results of the model predictions are compared with actual data to evaluate the model's effectiveness in predicting air quality in future periods.

This approach results in a strong basis for analyzing and predicting air quality within urban cities based on spatiotemporal data and machine learning methods.

Data Collection

It is sourced from different platforms, like the air quality monitoring stations operated by the Indian government across the country. The dataset contains both hourly and daily measurements of key pollutants, including PM2.5, PM10, NOx, CO, SO2, and O3, as well as meteorological parameters like temperature, humidity, wind speed, and wind direction. The data ranges between 2010 and 2023 [20]. This has been aggregated from multiple CSV files whereby each file represents data from a different monitoring station. Details of the dataset are outlined in Table 1. To capture well the spatial features, metadata have been included on the location of the monitoring stations. The dataset was supplemented with meteorological data sourced from other locations since air quality is highly exposed to the weather.

Figure 2 displays plots that indicate the time series of four major indicators of air quality, which include PM2.5, Carbon Monoxide (CO), O3 Concentration, and Nitrogen Compounds (NOx), at periodic intervals across several years over Indian urban cities. Each plot compares daily, monthly, and yearly groupings to reveal short-term fluctuations and long-term trends, which form the foundations for spatiotemporal analysis as well as for machine learning-based prediction of air quality.

In Figure 2(a) that presents curve of the time series of PM2.5 and PM10 concentrations which show that the variations are quite large along with high peaks corresponding to strong air pollution periods. Such periods may coincide with the peak seasons for celebrations, agricultural burning, and industrial effluents. Average yearly values have a tendency to decline

marginally with time, suggesting a probable enhancement in the quality of air due to regulations or shift in the source of emissions.

Figure 2(b) Carbon Monoxide The carbon monoxide is a pollutant much more directly associated with vehicle exhausts and incomplete combustion. Though the concentrations are not as high as the particulate matter, there are marked spikes perhaps during periods of heightened traffic or industrial operations. Annual averages seem quite stable indicating short term variability exists but the long term concentration of carbon monoxide has not shifted significantly over time.

Figure 2(c) plots the concentration of ozone, a secondary pollutant formed through photochemical reactions in the atmosphere. The pattern of ozone concentrations is strongly seasonal, with periodic peaks that may be associated with variability in temperature and strength of sunlight, which are the primary catalysts for ozone formation. The general downward trend in yearly averages suggests that conditions or precursors for the formation of ozone may be decreasing, perhaps because of control measures on emissions of nitrogen oxides and volatile organic compounds.

Finally, Figure 2(d) displays the nitrogen oxides and nitrogen dioxide concentrations, both of which are pollutants closely associated with traffic and industrial activities. The nitrogen compounds have spiky periodic behaviour that may reflect periods when vehicles or industrial production rates are higher. However, the average yearly trend is pretty stable, indicating that despite episodes where emissions are elevated, long-term levels for the nitrogen compounds have changed little.These plots collectively shed insight into the time dependency of air pollutants in urban Indian environments. The periodic patterns-including seasonal spikes and event-based pollution peaks-provide valuable input into developing machine learning models that predict air quality from historical data.

The Figure 3 is pairplot matrix that shows the correlation of many air quality variables such as PM2.5, Nitric Oxide (NO), Nitrogen Dioxide (NO2), CO, O_3, and NOx. A histogram of the individual variables is shown in each diagonal element, and the pairwise relationships between variables are represented by scatter plots in the off-diagonal elements.

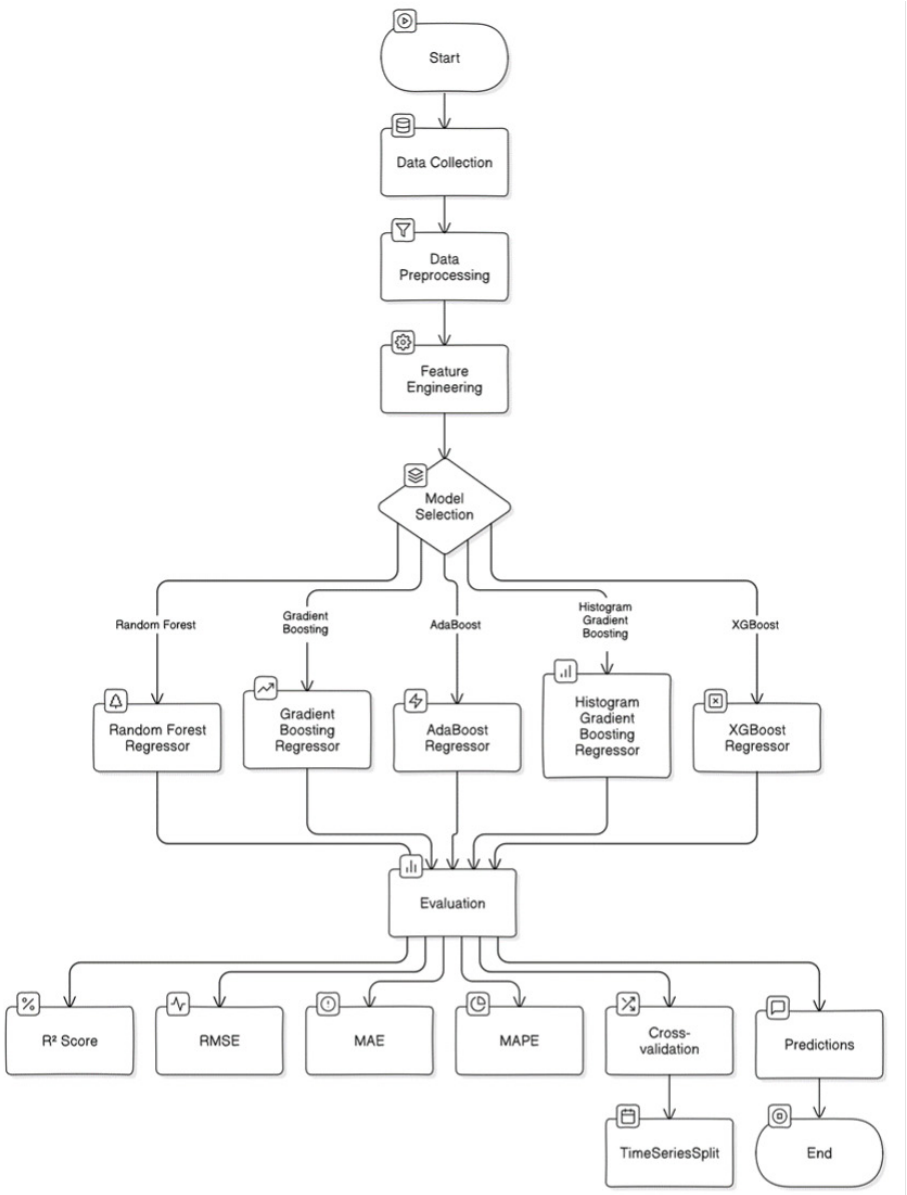

Figure 1. Flowchart of the machine learning pipeline for air quality prediction

Table 1. Description of the dataset used for air quality prediction

Indian Cities	No. of Instances	Parameters	Duration
North India: Delhi, Haryana, Himachal Pradesh, Jammu and Kashmir, Punjab, Uttarakhand, Uttar Pradesh, Chandigarh (Union Territory) South India: Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, Telangana, Puducherry (Union Territory), East India: Arunachal Pradesh, Assam, Bihar, Jharkhand, Odisha, Sikkim, West Bengal, Manipur, Meghalaya, Mizoram, Nagaland, Tripura, West India: Goa, Gujarat, Maharashtra, Rajasthan, Madhya Pradesh, Chhattisgarh	2796171	57	2010 to 2023

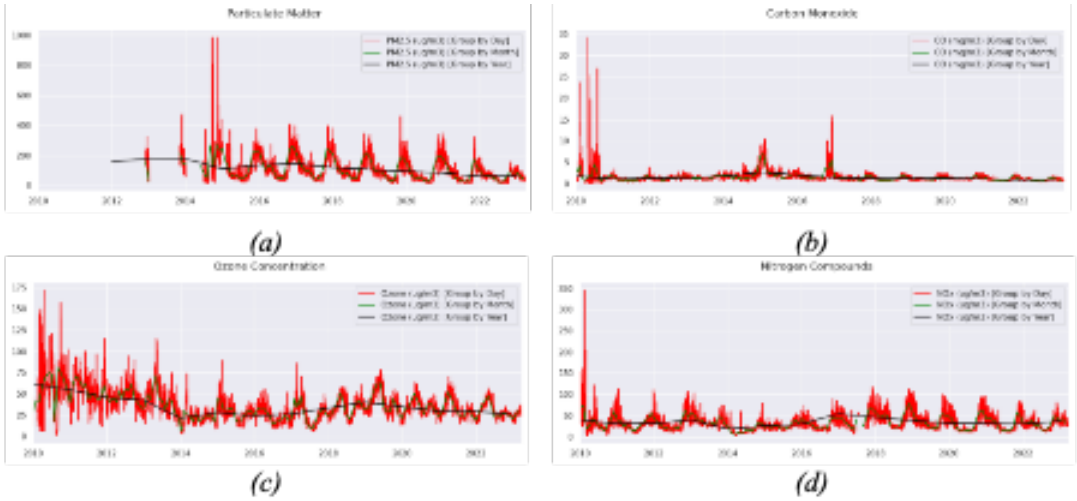

Figure 2. Time series plots of four major air quality parameters—(a) Particulate Matter, (b) Carbon Monoxide, (c) Ozone Concentration, and (d) Nitrogen Compounds—for years in Indian cities

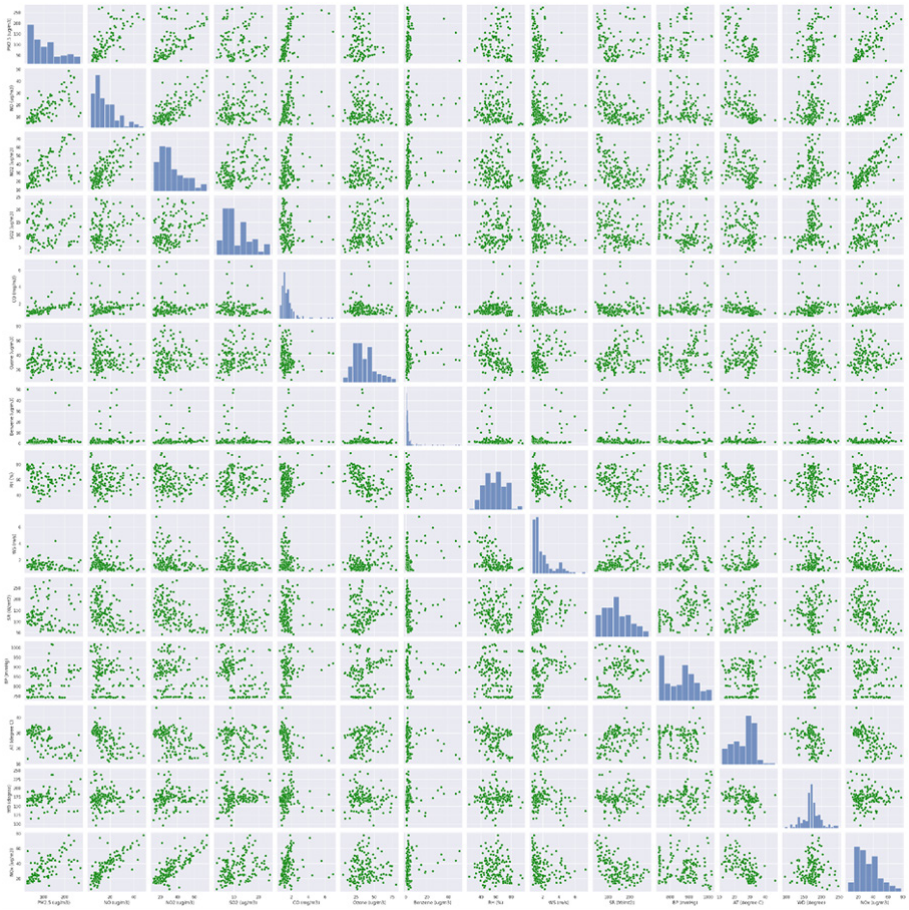

Figure 3. Pairplot matrix that visualizes the relationships between several air quality parameters, including PM2.5, NO, NO2, CO, O_3, and NOx

The scatter plots indicate potential correlations between other pollutants. For example, the upward trend in the scatter plots suggests positive correlation between NOx and NO2, and also between PM2.5 and NOx. The diagonal histograms give an idea of the distribution of each pollutant, some of which are skewed, such as CO and O_3, so that there might be more extreme values in their distributions for some pollutants. Other relationships seem more scattered, thus more weakly correlated, such as O_3 and PM2.5 or O_3 and NOx. This matrix is helpful in identifying a number of key pollutant interactions that could be informative as part of any sort of multivariate analysis or feature selection aimed at machine learning models targeted toward the prediction of air quality.

Data Preprocessing

After collecting the raw data, several preprocessing steps have been performed to prepare it for analysis. Firstly, the "From Date" column has been converted into a datetime index to conduct effective time-series analysis. In addition, the redundant column "To Date" has been dropped. Finally, merging of duplicate measurements of the same pollutant. Under different slightly different names, some of the pollutants were listed as "Xylene" and "MP-Xylene." For consistency, a specialized function was created that aggregated these columns. Handling missing data was another pre-processing task. The two strategies put in place were: in the interpolation and fill-ins, the "pad" method was used for sparse features with missing data, and for columns where a feature had some reasonable percentage of missing values, columns were dropped where the missingness went above a predefined threshold. From this point, the dataset was ready to use for further analysis. From here, cleaned and merged data for feature extraction and analysis was presented.

Feature Engineering

To enhance the performance of MACHINE learning models, several new features were engineered out of the existing dataset. The hour of the day, day of the week, day of the month, month, and year temporal features were extracted from the data in order to capture seasonality and daily/weekly pollution patterns. Figure 4 illustrates detailed analysis of PM2.5 concentrations at different temporal scales integral to feature engineering process in our study.

Daily trends are evident in "PM2.5 (μg/m^3) by day/month," which captures the daily periodic feature fluctuations in pollution. More seasonal trends are present in the "PM2.5 (μg/m^3) by month" box plot, clearly making evident the role of monthly features in this system. The "PM2.5 (μg/m^3) by weekday" plot illustrates the weekly trends and the good correlation with the day of the week feature to capture pollution cycles happening based on days of the week. PM2.5 (μg/m^3) by week/year" produces a heatmap-style plot, so that both weekly and yearly data may be available and thus potential long-term trends are available to justify lag features from previous years. Finally, "PM2.5 (μg/m^3) by year" shows some historical levels of pollution, which becomes important as lag variables in the model for enhancing accuracy. The visualizations together depict how the engineered features can capture temporal dependencies, seasonality, and historical trends and improve the model's ability to predict air quality in urban cities. Although the inclusion of meteorological data is not taken as visualization in itself, it is represented through the seasonal and temporal variations-that reflect the role of weather in the pollution levels-increased levels. Furthermore, the use of lag features was necessary to capture temporal dependencies in the data. For instance, the past year and two years ago were considered as lag variables for the concentrations of PM2.5 in its trend of historical pollution in order to improve the accuracy of its prediction. Meteorological factors such as temperature, wind speed, wind direction, and humidity were added as major predictors because their effects are critical in the dispersion and concentration of pollutants. These feature engineering steps were to enhance the ability of the model so that it can work better on predictive abilities of the pollution levels.

Model Selection

Several ensemble-based algorithms for air quality forecasting, focusing on PM2.5 concentrations in particular because these have the largest health impacts.

Random Forest Regressor [21, 22]: This is a strong, non-linear model that uses multiple decision trees to produce a prediction by reducing overfitting from averaging the outputs generated by multiple trees.

Gradient Boosting Regressor [23]: Ensemble learning technique arranged sequentially where each successive tree corrects the errors made by the previous one.

AdaBoost Regressor [22]: It is an adaptive boosting model that combines weak learners into a strong predictor, mainly where lots of noise are found in the data.

Histogram-Based Gradient Boosting Regressor [24]: The optimized version of the gradient boosting mechanism, which uses histograms. The above three models are very efficient, mainly in computation and utilization of memory instead of the regular linear regressor.

XGBoost Regressor [22]: This is a high-performance, efficient gradient boosting algorithm. XGBoost is an abbreviation of "Extreme Gradient Boosting." It creates decision trees sequentially. New trees correct the residual errors from the previous trees. Regularization prevents overfitting, parallel processing makes the computation faster, and the tree-pruning mechanism prevents over-complexity.

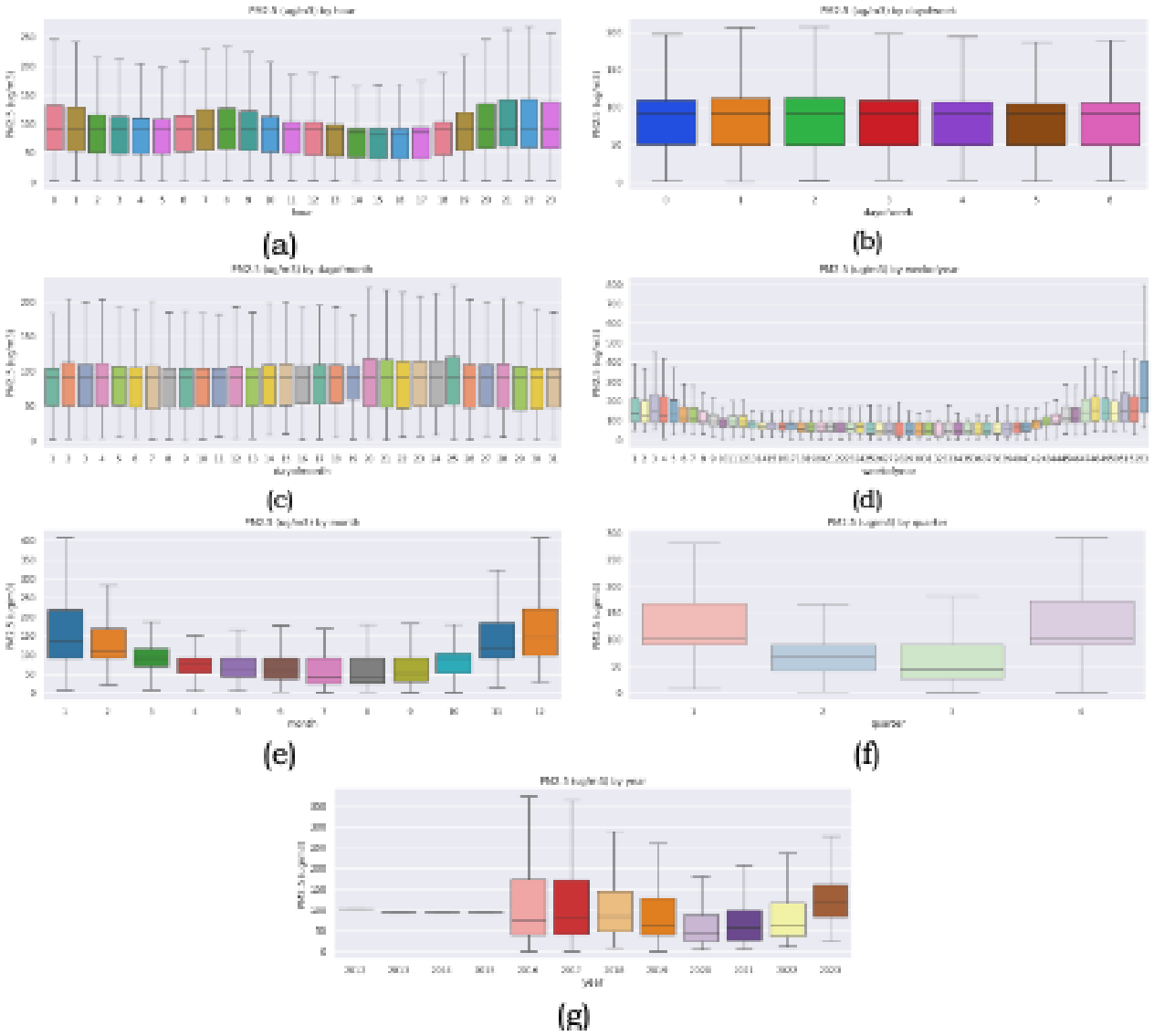

Figure 4. PM2.5 concentrations by hour, day, month, weekday, and year, illustrating the temporal behavior and highlighting seasonal patterns in pollution levels

RESULTS AND DISCUSSIONS

The dataset splitted into training and test datasets, with 80% to training and 20% to testing. The following evaluation metrics would determine the performance of the different models.

•R^2 Score: It means the percentage of the variance in the dependent variable that can be explained by the independent variables. In simple words, it measures the variance in the dependent variable which this model explains.

•Root Mean Squared Error (RMSE): It is a measure for the average magnitude of prediction errors. The better the performance, the smaller the value is.

•Mean Absolute Error (MAE): Average absolute difference between true values and predicted values.

•Mean Absolute Percentage Error (MAPE): The errors are usually represented in percent form, which provides understanding of relative accuracy due to the model.

Table 2 and Figure 5 will give an overview summary of performance metrics for some of the machine learning models applied in the prediction of air quality, particularly PM2.5 concentrations.

The Random Forest model gives a very high R^2 value of 0.99 on the training set, which means an excellent fit, but at the same time, the other R^2 value is low at 0.35 for the test, which would indicate overfitting. It also has a relatively high RMSE of 63.60, MAE of 44.98, and MAPE of 0.57. The Gradient Boosting model shows more balanced performance across folds, with an R^2 of 0.54 for the training set and 0.48 for the test set and lower error metrics generally (RMSE: 56.46, MAE: 39.60, MAPE: 0.50), suggesting better generalization. Lowest R^2 was with AdaBoost at 0.27 on the training set and 0.28 on the test set, while having the highest RMSE of 66.86, MAE of 52.34, and MAPE of 0.94. Histogram GB and XGBoost are at a decent point. Histogram GB is at R^2 = 0.72 training and 0.43 test, and XGBoost is at R^2 = 0.82 for the training set and 0.38 on the test set; both of the models have

related error metrics with a slight edge going to Histogram GB in RMSE and MAE. The table overall shows a trade-off between model complexity and generalization, and Gradient Boosting is relatively robust in this regard.

All models were tuned with the RandomizedSearchCV algorithm. It had found the best combination of hyperparameters that maximally allowed the optimal model configuration. Some of the critical features, namely number of trees, learning rate and maximum depth, were optimized. Cross-validation was also carried out using a time series-split technique. This ensured robust testing of every model on more than one-fold. The predictions were evaluated not only on the test data, but also by visualization methods in order to estimate the generalization capabilities and the potential for forecasting into the future. All the ensemble models were tested for hyperparameter optimization through Randomized Search Cross-Validation to fine-tune the models for the best-performing parameter configurations. The time it took to fit each model, as well as subsequent prediction times, was recorded with the view of evaluating efficiency based on computations.

Table 2. Performance metrics (RMSE, MAE, MAPE, and R^2) for machine learning models used in predicting PM2.5 levels

Model	RMSE	MAE	MAPE	R^2 (Train)	R^2 (Test)
AdaBoost	66.86	52.34	0.94	0.27	0.28
Gradient Boosting	56.46	39.6	0.5	0.54	0.48
Random Forest	63.6	44.98	0.57	0.99	0.35
XGBoost	61.72	42.4	0.53	0.82	0.38
Histogram GB	59.17	41.56	0.52	0.72	0.43

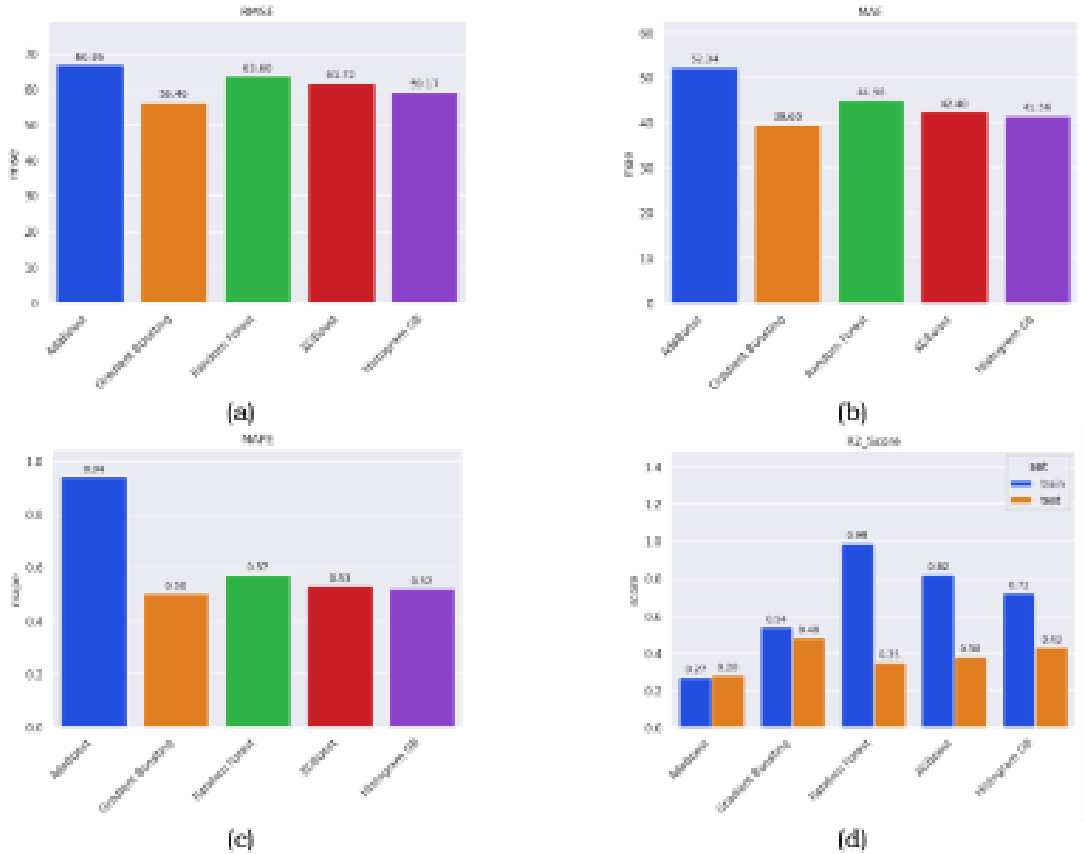

Figure 5. Performance comparison of AdaBoost, Gradient Boosting, Random Forest, XGBoost, and Histogram GB models with the help of RMSE, MAE, MAPE, and R^2 for the prediction of the air quality

Prediction and Visualization

Following model training, the best-performing models applied towards forecasting future PM2.5 levels from 2023 to 2024. Lag features added and also engineered temporal variables within the models for a better forecast. Below are graphical representations of the actual vs. predicted values as shown in Fig 6, focusing on different time horizons, including daily, monthly, and yearly-all concerning assessing model performance over the horizon of times. Heatmaps alse produced to denote the interlinkages between diverse pollutants and meteorological factors.

The Figure 6 compares the prediction of the different concentrations of PM2.5 by various machine learning models. The computed values presented above are plotted on the same graph as actual values for comparison of which machine learning model is good at predicting the predicted PM2.5 concentration levels.

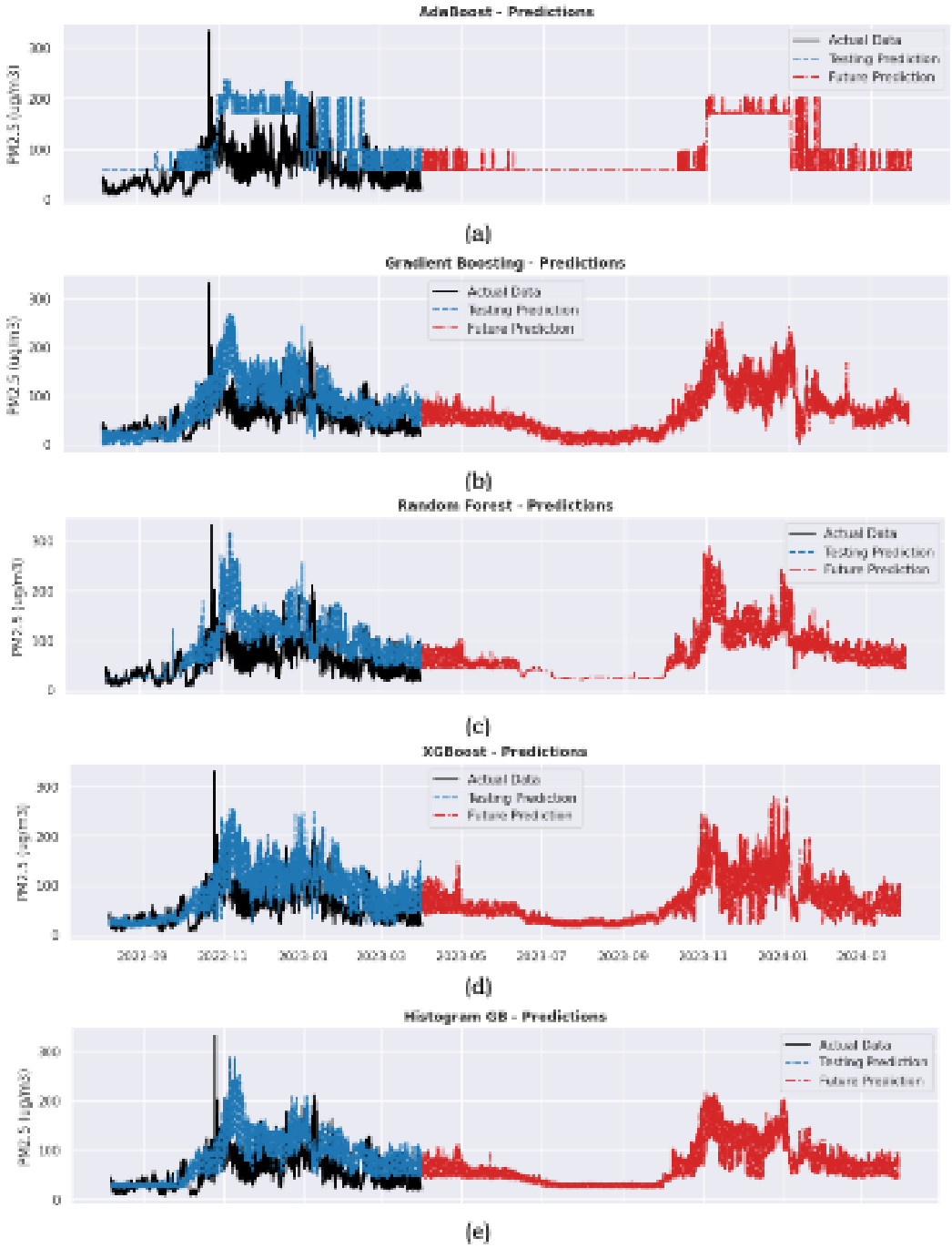

Figure 6. Comparison of predicted PM2.5 levels for 2023-2024 using machine learning models, highlighting the strengths and weaknesses of each model in forecasting future air quality trends

Here, the models under consideration are AdaBoost, Gradient Boosting, Random Forest, XGBoost, and Histogram Gradient Boosting. The actual and predicted values are differentiated in the plot by color representation: actual historical concentration of PM2.5 is blue, predictions during the testing phase overlapping the actual data orange, and the future predictions, where the models predict the concentration of PM2.5 beyond the data available - green.

Now let's examine performance of every model below:

AdaBoost: Test predictions for AdaBoost are very much like real data but noisier than that of Gradient Boosting. It shows a lot of variability in its future predictions, hence with sharp jumps, this must perhaps be a model that will have trouble making long-term predictions.

Gradient Boosting: Gradient Boosting exhibits considerably more regular patterns when tested and fits the actual pretty well. The future predictions are smoother than those from Random Forest but still present sudden growths at certain periods.

Random Forest: Generally, this model provides a reasonable fit to actual data during the forecasted period (blue and orange regions). At the same time its prospects for the future (green) are chaotic with deep oscillations and have no trend, meaning that it fails to generalize to future PM2.5 concentration.

XGBoost: XGBoost is the best fit for the testing phase, where its predictions of the orange curve closely follow actual data of the blue curve. The predictive curve of the future tends to be relatively smooth and follows logically, which indicates good generalization of XGBoost for the concentration of PM2.5 in the future.

Histogram Gradient Boosting: Histogram GB predicts even better compared to AdaBoost while being relatively less stable compared to Gradient Boosting. Their subsequent predictions are smoother and have an upward trend, though there still exist some degree of fluctuations.

In a summary, XGBoost is the best model for both short-term and longer-term predictions of PM2.5 concentrations. It has the smoothest prediction and least erratic jump compared to the other models in the future prediction. In terms of longer-term predictions, Random Forest and AdaBoost are the worst models as they exhibit high variability in future predictions.

CONCLUSION

The study proposed applications of machine learning techniques such as Random Forest, Gradient Boosting, AdaBoost, Histogram-Based Gradient Boosting, and XGBoost in predicting PM2.5 concentrations in Indian urban cities. Spatiotemporal analysis integrated with advanced predictive modeling to prove that machine learning does bear high potential in accurately forecasting air-quality trends which can lead toward more effective decision-making in environmental management. Gradient Boosting had nearly balanced performance over folds, with R^2 values of 0.54 for training and 0.48 for test, and lower error metrics, such as RMSE: 56.46, MAE: 39.60, MAPE: 0.50. Nevertheless, it has its limitations, such as sensitivity to overfitting, the need for hyperparameter tuning, and slower training. XGBoost had better fit during testing and more smooth, logically consistent predictions for future PM2.5 concentrations, which means better generalization for long-term forecasting. While Gradient Boosting fairly balanced R^2 scores pointed to reduced overfitting, the XGBoost showed advantages in visualization results and thus long-term forecasting. The strengths of the model are in handling non-linear associations, resistance to anomalies and missing data, and its ability to pick out complex interactions among features. However, despite such advantages, the model is demanding in terms of adjustment in avoiding overfitting and consumes more resources than simpler models, which poses a problem in resource-constrained environments. It was noticed that the Random Forest model did well for the training data but failed to generalize properly for the test data, effectively communicating the issues with overfitting in complex environmental datasets. AdaBoost was very weak while Histogram-Based Gradient Boosting and XGBoost resulted in mediocre accuracy with good predictive consistency shown by the latter.

Its inclusion of meteorological data and the incorporation of temporal feature engineering-by increasing the models' capability to capture short-term fluctuations together with long-term pollution trends-allowed it to understand better the factors degrading air quality in Indian cities. During study, seasonal peaks in pollution were also emphasized, such as resulting from crop burning along with adverse meteorological conditions during winter months.

From a policy point of view, this work shows the potential for such machine learning to be an effective tool in managing air quality, providing reliable near-real-time forecasts, enabling policymakers to intervene proactively and reduce public health risks and alleviate the impact of extreme events. Furthermore, it opens up possibilities where real-time air quality monitoring is to be integrated with machine learning models, thereby providing a dynamic and adaptive solution for dealing with air pollution in rapidly urbanizing regions.

Future research would focus on widening the scope of this study to more cities, with even finer granularities of data. The increasing use of more sophisticated machine learning techniques, such as deep learning, or perhaps even hybrid models that combine the best features of several, could further improve predictive accuracy and robustness. These challenges will be significant in the development of scalable, reliable, and actionable air quality prediction systems that can function across diverse urban environments in India and other geographies. In order to extend the applicability of the model beyond the Indian scenario, it is important to introduce elements that are specific to local sources of pollution, local environmental conditions, and regulatory structure. Retraining of the model with data from diverse urban locations across the world may further improve the generality of the model. More importantly, accurate feature engineering needs to be

ensured, especially when incorporating city-specific variables like industrial activities, transportation dynamics, and geographical characteristics.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

USE OF AI FOR WRITING ASSISTANCE

Not declared.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

1. H. Liu, Q. Han, H. Sun, J. Sheng, and Z. Yang, "Spatiotemporal adaptive attention graph convolution network for city-level air quality prediction," Scientific Reports, vol. 13(1), pp. 13335, 2023, doi: 10.1038/s41598-023-39286-0.
2. J. Duan, Y. Gong, J. Luo, and Z. Zhao, "Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer," Scientific Reports, vol. 13(1), pp. 12127, 2023, doi: 10.1038/s41598-023-36620-4.
3. D. M and R. V, "Novel Regression and Least Square Support Vector Machine Learning Technique for Air Pollution Forecasting," International Journal of Engineering Trends and Technology, vol. 71(4), pp. 147–158, 2023, doi: 10.14445/22315381/IJETT-V71I4P214.
4. X. Zhang, X. Jiang, and Y. Li, "Prediction of air quality index based on the SSA-BiLSTM-LightGBM model," Scientific Reports, vol. 13(1), pp. 5550, 2023, doi: 10.1038/s41598-023-32775-2.
5. M. Bonas and S. Castruccio, "Calibration of SpatioTemporal forecasts from citizen science urban air pollution data with sparse recurrent neural networks," The Annals of Applied Statistics, vol. 17(3), 2023, doi: 10.1214/22-AOAS1683.
6. R. López-Blanco, M. Chaveinte García, R. S. Alonso, J. Prieto, and J. M. Corchado, "Pollutant Time Series Analysis for Improving Air-Quality in Smart Cities," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8(3), pp. 98, 2023, doi: 10.9781/ijimai.2023.08.005.
7. R. Guo, Y. Qi, B. Zhao, Z. Pei, F. Wen, S. Wu, and Q. Zhang, "High-Resolution Urban Air Quality Mapping for Multiple Pollutants Based on Dense Monitoring Data and Machine Learning," International Journal of Environmental Research and Public Health, vol. 19(13), pp. 8005, 2022, doi: 10.3390/ijerph19138005.
8. M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey," Artificial Intelligence Review, vol. 56(9), pp. 10031–10066, 2023, doi: 10.1007/s10462-023-10424-4.
9. S. Chowdhury, A. Pillarisetti, A. Oberholzer, J. Jetter, J. Mitchell, E. Cappuccilli, B. Aamaas, K. Aunan, A. Pozzer, and D. Alexander, "A global review of the state of the evidence of household air pollution's contribution to ambient fine particulate matter and their related health impacts," Environment International, vol. 173, pp. 107835, 2023, doi: 10.1016/j.envint.2023.107835.
10. J. Cheng, F. Li, L. Liu, H. Jiao, and L. Cui, "Spatiotemporal Variation Air Quality Index Characteristics in China's Major Cities During 2014–2020," Water Air & Soil Pollution, vol. 234(5), pp. 292, 2023, doi: 10.1007/s11270-023-06304-w.
11. R. R. Behera, D. R. Satapathy, A. Majhi, and C. R. Panda, "Spatiotemporal variation of atmospheric pollution and its plausible sources in an industrial populated city, Bay of Bengal, Paradip, India," Urban Climate, vol. 37, pp. 100860, 2021, doi: 10.1016/j.uclim.2021.100860.
12. A. A. Khan, K. Garsa, P. Jindal, P. C. S. Devara, S. Tiwari, and P. B. Sharma, "Demographic Evaluation and Parametric Assessment of Air Pollutants over Delhi NCR," Atmosphere (Basel), vol. 14(9), pp. 1390, 2023, doi: 10.3390/atmos14091390.
13. Vaishali, G. Verma, and R. M. Das, "Influence of Temperature and Relative Humidity on PM2.5 Concentration over Delhi," MAPAN, vol. 38(3), pp. 759–769, 2023, doi: 10.1007/s12647-023-00656-8.
14. K. K. Rani Samal, K. Sathya Babu, A. Acharya, and S. K. Das, "Long Term Forecasting of Ambient Air Quality Using Deep Learning Approach," in IEEE 17th India Council International Conference (INDICON), IEEE, Dec. 2020, pp. 1-6. doi: 10.1109/INDICON49873.2020.9342529.
15. M. Ansari and M. Alam, "An Intelligent IoT-Cloud-Based Air Pollution Forecasting Model Using Univariate Time-Series Analysis," Arabian Journal for Science and Engineering, vol. 49(3), pp. 3135–3162, 2024, doi: 10.1007/s13369-023-07876-9.
16. K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," International Journal of Environmental Science and Technology, vol. 20(5), pp. 5333–5348, 2023, doi: 10.1007/s13762-022-04241-5.
17. S. M. Selvi, K. Ravikumar, A. D. Rajendran, A. B.

Bagavathi, N. Narayanan, and V. Mangottiri, "Assessment of Air Quality Index in major cities of India - Lessons from Lockdown," IOP Conference Series Materials Science and Engineering, vol. 955(1), pp. 012079, 2020, doi: 10.1088/1757-899X/955/1/012079.

18. V. Sharma, S. Ghosh, S. Dey, and S. Singh, "Modelling PM2.5 for Data-Scarce Zone of Northwestern India using Multi Linear Regression and Random Forest Approaches," Annals of GIS, vol. 29(3), pp. 415–427, 2023, doi: 10.1080/19475683.2023.2183523.
19. A. Masood and K. Ahmad, "Prediction of PM2.5 concentrations using soft computing techniques for the megacity Delhi, India," Stochastic Environmental Research and Risk Assessment, vol. 37(2), pp. 625–638, 2023, doi: 10.1007/s00477-022-02291-2.
20. Central Pollution Control Board, "CPCB|Central Pollution Control Board," cpcb.nic.in, 2019. https://cpcb.nic.in/
21. D. Kothandaraman, N. Praveena, K. Varadarajkumar, B.M. Rao, D. Dhabliya, S. Satla, and W. Abera, "Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning," Adsorption Science & Technology, vol. 2022, pp. 1–15, 2022, doi: 10.1155/2022/5086622.
22. T. Toharudin, R.E. Caraka,I.R. Pratiwi,Y. Kim, P.U. Gio, A.D. Sakti, M. Noh, F.A.L. Nugraha, R.S. Pontoh, T.H. Putri, T.S. Azzahra, J.J. Cerelia, G. Darmawan, and B. Pardamean, "Boosting Algorithm to Handle Unbalanced Classification of PM2.5 Concentration Levels by Observing Meteorological Parameters in Jakarta-Indonesia Using AdaBoost, XGBoost, CatBoost, and LightGBM," IEEE Access, vol. 11, pp. 35680-35696, 2023, doi: 10.1109/ACCESS.2023.3265019
23. N. Doreswamy, H. K. S, Y. Km, and I. Gad, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," Procedia Computer Science, vol. 171, pp. 2057–2066, 2020, doi: 10.1016/j.procs.2020.04.221.
24. A. Sarkar, S. S. Ray, A. Prasad and C. Pradhan, "A Novel Detection Approach of Ground Level Ozone using Machine Learning Classifiers," Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 428-432, doi: 10.1109/I-SMAC52330.2021.9640852