SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Machine Learning Based Prediction Models of Air Quality Index for Mid-Sized Urban Cities

Vivek Mathur[1*], Divya Srivastava[2], Vinai Singh[3], Anish Kumar[4]

[1]Assistant Professor, Amity University, Uttar Pradesh, Lucknow Campus, Lucknow, India. E-mail: vmathur@lko.amity.edu
[2]Assistant Professor, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, India. E-mail: dsrivastava4@lko.amity.edu
[3]Associate Professor, Bansal Institute of Engineering and Technology, Lucknow, India. E-mail: vinaisingh17786@gmail.com
[4]Assistant Professor, Rajkiya Engineering College, Azamgarh, Uttar Pradesh, India. E-mail: anish@gecazamgarh.ac.in

Abstract
Air quality deterioration in mid-sized urban cities poses growing environmental and public-health challenges because pollutant concentrations can vary substantially with traffic activity, urban development, and atmospheric conditions. Reliable Air Quality Index (AQI) prediction is therefore important for timely air-quality management and early warning. This study develops and evaluates Machine-Learning (ML)-based prediction models for AQI using air-quality observations from a high-traffic monitoring station in Lucknow, India. A three-month dataset covering January–March 2024 and comprising 88 observations was analyzed using "$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O3" as predictor variables and AQI as the target. Four regression approaches—"Linear Regression, LASSO Regression, Random Forest (RF), and Extreme Gradient Boosting (XGBoost)"—were developed using an 80:20 division of the dataset into training and testing subsets. Their predictive performance was evaluated using "Mean Absolute Error (MAE), Mean Squared Error (MSE), R^2, Adjusted R^2, Akaike Information Criterion (AIC), and Schwarz's Bayesian Criterion (SBC)". RF achieved the best performance, with an MAE of 7.626, MSE of 108.288, R^2 of 0.774, and Adjusted R^2 of 0.650. It also produced the lowest AIC (96.326) and SBC (101.668), demonstrating its superior overall predictive performance. XGBoost showed moderate performance, while LASSO performed weakest. Findings support ensemble learning for reliable AQI prediction.

Keywords: Air Quality Index (AQI), Machine Learning, Air Quality Prediction, Random Forest, Urban Air Pollution.

I. Introduction

The atmosphere primarily consists of nitrogen (78%) and oxygen (21%), along with smaller proportions of gases such as argon and carbon dioxide [1]. The composition and quality of ambient air is however, very variable, depending on the geographical location, the altitude, meteorological conditions and the presence of natural or man-made pollutants. The particulate matter comprising dust, ash, pollen, aerosols and gaseous pollutants are highly contributing factors to environmental pollution, with negative impacts on the quality of the environment and human health [2–6]. AQI gives a simplified AQI which translates the concentrations of different pollutants into one index and one category, making it easy for the public and decision makers to comprehend the level of pollution. Under India's National Air Quality Index (NAQI), ambient air quality is assessed using eight key pollutants: PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, O_3, NH_3, and Pb. The resulting AQI value is interpreted through six health-based categories, ranging from "Good" (0–50) and "Satisfactory" (51–100) to "Moderately Polluted" (101–200), "Poor" (201–300), "Very Poor" (301–400), and "Severe" (401–500). As the AQI value increases, the associated pollution level and potential health risks also become progressively greater [7].

High AQI days can result in serious health impacts and necessitate preventive action, especially for vulnerable populations. Weather related variables, including temperature, wind velocity, humidity, rainfall, and solar irradiance, can substantially affect atmospheric pollution levels. These factors alter the movement, dilution, chemical conversion, and retention of airborne pollutants, thereby influencing the observed air-quality conditions alongside direct emission sources [8]. Air quality is a global issue impacting almost 99% of the world's population, with the World Health Organization (WHO) reporting that the monitoring and management of air quality was of vital importance [9]. There is a strong correlation between the tiny particle size ($PM_{2.5}$) and negative respiratory and cardiovascular consequences, making it one of the main pollutants and a major cause for worry. Asthma, other chronic respiratory disorders, heart disease, stroke, and lung cancer are only some of the illnesses linked to long-term exposure to polluted air [10].

Traditional air quality monitoring is based on well-known methods for measuring gaseous pollutants, like optical and beta attenuation, as well as particulate pollutants, such as gravimetric analysis and elemental analysis [11]. These methods are good for providing reliable measurements, but monitoring and reliable forecasts of air pollution can be difficult in urban areas, where air pollution is highly variable both temporally and spatially. The challenge is especially important for mid-size cities that experience rapid urbanization, more vehicles, building development, industrial and commercial growth, and shifting land use patterns which may lead to complex and variable pollution scenarios. In addition, meteorological factors may influence the dispersion of pollutants and atmospheric chemistry, making it difficult to predict future AQI based solely on monitoring data [12-15].

The ML approaches are also emerging as promising tools for prediction of air quality as they are suited to process multidimensional, nonlinear, and large-scale datasets and to discover complicated relationships between the concentrations of polluting elements, meteorological conditions and temporal patterns. While traditional statistical methods can involve preprogrammed relationships between variables, ML models can discover predictive patterns from historical observations. Artificial Neural Networks (ANNs) have been applied to a large extent due to their capabilities of representing the nonlinear relationship in air-quality data sets [16–19]. Support Vector Machines (SVMs) have also been shown to be effective in supervised prediction and classification problems [20–25], and RFs use ensembles of decision trees to model complex relationships and lessen the effect of any one predictor on the model [26,27]. Likewise, K-Nearest Neighbour (KNN) is a relatively simple data-driven approach that can also be used for the air-quality forecasting problem [28,29]. Such models can only forecast well if the data input is of good quality and quantity, has a sufficiently fine temporal resolution, and is representative of the study area, given the characteristics of the study area.

The use of data to predict air quality has been further extended with recent developments in computational intelligence. Quantum ML QML is a novel field of research that brings quantum computing and ML together, gaining momentum for its ability to tackle complex computational challenges and data in high dimensions [30-33]. In the field of classification and pattern recognition, Quantum SVM (QSVMs) have been explored and have been demonstrated to be potentially applicable to air quality related tasks [34-36]. Environmental and meteorological variables have also been explored with quantum-enhanced learning frameworks to create hybrid quantum-classical approaches [37,38]. These methods, however, are still relatively new and their practical benefits compared to traditional classical ML methods need to be assessed with real-world urban air-quality datasets.

While ML techniques are being increasingly used in air-quality forecasting, there are still several challenges. Differences in the emission sources, urban morphology, meteorological conditions, population density and local activities contribute to large variations in air-quality conditions between cities[39,40]. Therefore, the performance of prediction models built for large and metropolitan cities may not be the same as for the mid-sized urban cities. Moreover, numerous previous studies are mainly about prediction of individual pollutants, especially $PM_{2.5}$, rather than directly predicting the overall AQI that is a composite of information from multiple pollutants. Reliable AQI prediction models which include the pollutant concentrations and meteorological variables can thus help describe future air-quality conditions more completely.

In this context, a ML model for predicting air quality can be a relevant solution for prompt air-quality evaluation and public health decision-making in medium-sized cities. Accurate air-quality predictions can help authorities detect deteriorating conditions, issue warnings, develop mitigation plans, and manage urban air quality effectively. Thus, designing and validating robust ML forecasting models considering pollutants and meteorological factors is significant for reliable, location-specific forecasting. Here are the research objectives as shown below:
1. To develop ML-based models for predicting the AQI using major air pollutants, including $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3.
2. To evaluate the predictive performance of LR, LASSO Regression, RF, and XGBoost models for AQI prediction.
3. To compare the developed models using MAE, MSE, R^2, Adjusted R^2, AIC, and SBC to identify the most effective prediction approach.
4. To investigate the relative importance of individual air pollutants in AQI prediction and determine the major contributors identified by different ML models.

5. To identify the most suitable ML model for AQI prediction in the investigated urban environment, supporting improved air-quality assessment and management.

II. Data and Methods

A. Study Area and Air Quality Dataset

In the present study, air-qualitative condition in an important urban city is analyzed, where there is significant number of vehicles, urbanization and anthropogenic emission. Air quality measurements were taken from a monitoring station in a high traffic emission area, and during peak traffic period. The data is collected for 88 observations over a period of three months (January 2024 to March 2024). For the AQI prediction analysis, "$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3" were considered as the input features, while AQI served as the response variable. The selected pollutants were used to train the ML models and generate AQI predictions. Before model development, the observed pollutant concentrations and AQI values were examined statistically to understand their overall distribution and degree of variation. The dataset was summarized using minimum, maximum, mean, and standard deviation values. This data set can be used to assess the proficiency of various ML algorithms for predicting AQI with different levels of urban air quality.

TABLE I. DESCRIPTIVE STATISTICS OF THE DATASET USED IN THE CURRENT STUDY

Statistic	**$PM_{2.5}$**	**PM_{10}**	**NO_2**	**SO_2**	**CO**	**Ozone**	**AQI**
No. of observations	88	88	88	88	88	88	88
Minimum	1.000	17.000	4.250	2.250	0.240	33.400	41.000
Maximum	157.000	226.000	20.050	22.000	1.690	115.400	219.000
1st Quartile	28.375	71.750	5.513	6.150	0.740	40.450	70.500
Median	45.750	92.500	7.200	8.175	0.825	52.475	90.500
3rd Quartile	58.625	117.875	8.513	10.763	0.950	67.563	112.500
Mean	47.733	96.210	7.883	8.626	0.882	56.449	94.875
Variance (n-1)	735.551	1505.711	11.310	12.103	0.077	313.687	1116.731
Standard deviation (n-1)	27.121	38.803	3.363	3.479	0.278	17.711	33.418

B. Data Preprocessing and Exploratory Analysis

The data was preprocessed and analyzed using exploratory methods prior to the development of the AQI prediction models to ensure the suitability of the data in the statistical and ML analysis. These steps included data quality assessment, summarization of pollutant distribution, analysis of associations of variables and the final predictor matrix. The presence of the dataset's predictors $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 and the response variable AQI necessitated these steps to help identify variability, potential anomalies, and relationships that could affect model performance.

1) Data Quality Assessment

The first step was to examine the data set for missing observations, duplicate observations, inconsistencies or potential outliers. The basic descriptive statistics were calculated for each variable X as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathrm{SD} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

where n represents the number of observations, $\bar{X}$ is the sample mean, and SD is the standard deviation. These measures provide a general view of the variability and central tendency of pollutant concentrations and AQI.

2) Descriptive Statistical Analysis

Descriptive analysis was conducted for $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3, and AQI. The low, high, mean, and standard deviation values for each variable were used to describe their distribution. They also checked the coefficient of variation so we could compare the relative variability of the pollutants:

$$\mathrm{CV} = \frac{\mathrm{SD}}{\bar{X}} \times 100$$

A higher CV indicates greater relative variability, which may reflect stronger fluctuations in pollutant concentrations during the observation period.

3) Outlier and Distribution Analysis

Box-plot analysis was visually performed to understand the distribution and look for extreme observations. The Interquartile Range (IQR) for a variable X is the following:

$$\mathrm{IQR} = Q_3 - Q_1$$

where Q_1 and Q_3 represent the first and third quartiles, respectively. Observations below $Q_1 - 1.5 \times \mathrm{IQR}$ or above $Q_3 + 1.5 \times \mathrm{IQR}$ may be considered potential outliers. Extreme pollutant concentrations were not simply thrown out, however, because they could be a true event of high pollution rather than an error in measurement. Figure 1 shows box plots for the variables investigated.

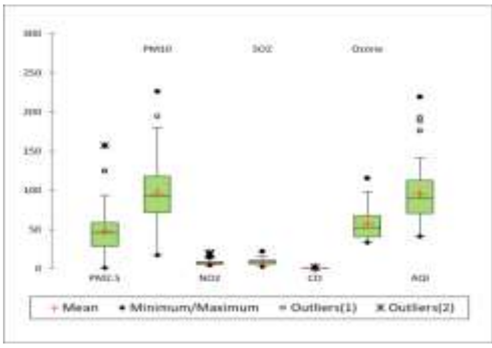

Figure 1. Box plots showing the distribution, central tendency, variability, and potential outliers of the air-quality variables.

4) Correlation Analysis

To examine the linear connections between AQI and pollution concentrations, Pearson correlation analysis was employed. When X and Y are two independent variables, the Pearson correlation coefficient is:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

A number close to +1 indicates a strong positive link, a value close to -1 indicates a strong negative association, and a value around zero indicates a weak linear correlation; the value of r can take on any value between -1 and +1. To find the relation between the six pollutant predictors and the AQI, the correlation matrix was made and displayed in a heatmap. The correlation map that is obtained is presented in Figure 2 and was also used to evaluate potential multicollinearity amongst predictors prior to model development.

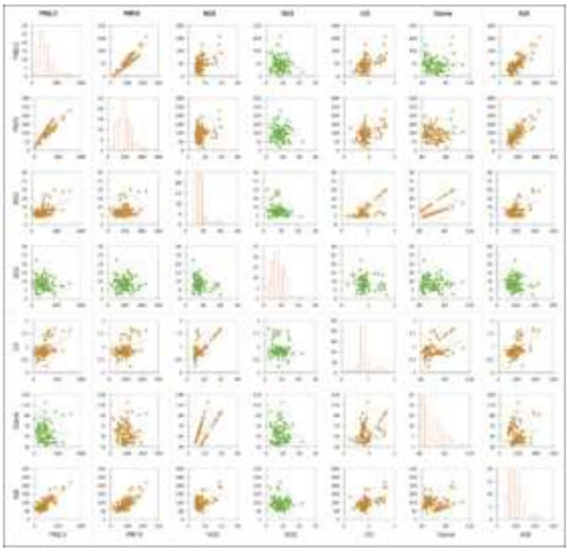

Figure 2. Correlation map showing relationships among the dataset variables.

C. AQI Prediction Framework

The proposed AQI prediction framework is based on the ML method and uses measured concentration of major air pollutants to determine the AQI. The independent variables in the present study were $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3, whereas AQI was regarded as the dependent variable. The prediction relationship can be represented as:

$$\mathrm{AQI} = f(\mathrm{PM}_{2.5}, \mathrm{PM}_{10}, \mathrm{NO}_2, \mathrm{SO}_2, \mathrm{CO}, \mathrm{O}_3)$$

where $f(\cdot)$ represents the nonlinear or linear mapping learned by the prediction model.

The predictor matrix can be expressed as:

$$\boldsymbol{X} = [\mathrm{PM}_{2.5}, \mathrm{PM}_{10}, \mathrm{NO}_2, \mathrm{SO}_2, \mathrm{CO}, \mathrm{O}_3]$$

and the corresponding target vector is:

$$\boldsymbol{Y} = [\mathrm{AQI}_1, \mathrm{AQI}_2, \dots, \mathrm{AQI}_n]^T$$

where n represents the total number of observations.

D. Training and Testing Strategy

The 88-observation data set was partitioned into a training set with 44 observations and a test set with 44 observations so that the ML models' prediction capabilities could be evaluated. In accordance with the study's protocol, the training of the model utilized 80% of the observations, or about 70 samples, while the testing phase made use of the remaining 20%, or around 18 samples. The six pollutant predictors and AQI were identified using the training set, and the unseen observations' predictions were tested using the testing set. This is the data partition:

$$D = D_{\text{train}} \cup D_{\text{test}}$$

where $D_{\text{train}} = 0.80D$ and $D_{\text{test}} = 0.20D$. The same training and testing subsets were applied to all four models to ensure a consistent and unbiased comparison of their predictive performance.

E. Machine Learning Prediction Models

Four ML models, which are LR, logistic regression, Support Vector Regression (SVR), and ANN were used to predict AQI based on the concentrations of $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3.The chosen models are both linear and nonlinear models of learning, which allows for a holistic comparison of their predictive performance. The baseline method was LR and LASSO introduced shrinkage and regularization. RF and XGBoost were chosen to model complex nonlinear relationships between pollutant levels and AQI.

1) Linear Regression (LR)

The baseline model was the LR due to its simplicity, interpretability, and low computational demand. It is based on the assumption that the dependent variable AQI is linearly dependent on the concentrations of pollutants. The model estimates the coefficients that minimize the difference between the observed AQI values and the estimated ones. The typical form of the LR equation is:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

where $\hat{y}$ represents predicted AQI, β_0 is the intercept, β_j represents the coefficient of the j^{th} predictor, and x_j denotes pollutant concentration. The coefficients are estimated by minimizing the residual sum of squares:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

LR provides a useful benchmark for determining whether more complex nonlinear models offer meaningful improvements in AQI prediction.

2) LASSO Regression

LASSO regression is an extension to LR which introduces an L_1 regularization. It was adopted to mitigate the effects of the less informative predictors and to attain stability in the model when predictors are related. The LASSO objective function is given by:

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

where λ controls the strength of regularization. Increasing λ increases coefficient shrinkage and may force some coefficients to exactly zero, effectively performing variable selection. The predicted AQI can be represented as:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

LASSO therefore provides both prediction and a means of identifying the relative contribution of pollutant predictors.

3) Random Forest

A multi-decision tree ensemble learning approach, RF improves resilience and captures nonlinear interactions by combining predictions. Every tree is trained on a bootstrap subset of the training data, and at each splitting stage, a randomly chosen subset of features is taken into account. To get the final prediction in regression, we average the results from all the trees:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(\boldsymbol{x})$$

The prediction of the b^{th} tree is denoted by $T_b(\boldsymbol{x})$, and B is the number of decision trees. Here is a conceptual expression of the ensemble's variance-reduction principle:

$$\widehat{\text{AQI}} = \text{Avg}[T_1(\boldsymbol{x}), T_2(\boldsymbol{x}), \ldots, T_B(\boldsymbol{x})]$$

RF is particularly suitable for environmental datasets because it can model nonlinear interactions among pollutant variables without requiring an explicitly defined functional relationship.

4) Extreme Gradient Boosting (XGBoost)

XGBoost is an ensemble learning approach that progressively generates decision trees, with each new tree focusing on reducing the prediction errors left by the preceding trees. Its optimization objective combines the model's prediction error with a regularization component, thereby limiting unnecessary model complexity and improving generalization:

$$\mathcal{L} = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega\,(f_k)$$

where l represents the loss function and $\Omega(f_k)$ controls the complexity of the k^{th} tree. The forecast is derived by combining the results of separate trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(\boldsymbol{x}_i)$$

XGBoost can model complex nonlinear relationships between pollutant concentrations and AQI via sequential error correction and regularization. It is capable of modelling interactions between several predictors and can be used with structured environmental data sets and comparative analysis of AQI prediction.

F. Model Training and AQI Prediction

The four selected ML models were trained using the training subset containing pollutant concentrations as input variables and AQI as the target variable. For each observation, the input vector was defined as $\boldsymbol{X}_i = [\mathrm{PM}_{2.5}, \mathrm{PM}_{10}, \mathrm{NO}_2, \mathrm{SO}_2, \mathrm{CO}, \mathrm{O}_3]^T$, while the corresponding measured AQI was represented by y_i .During training, each algorithm learned a mapping between pollutant concentrations and AQI:

$$\hat{y}_i = f(\boldsymbol{x}_i; \theta)$$

where f represents the selected ML model and θ denotes its learned parameters. Once models were trained, the fitted models were used to predict the AQI values for the independent testing subset. The prediction error for each observation was defined as:

$$e_i = y_i - \hat{y}_i$$

The observed and predicted AQI values were then correlated against each other to evaluate the performance of the models in predicting the AQI, namely LR, LASSO Regression, RF and XGBoost. The same predictor variables, partition of the data into training and test sets, and evaluation method were used for all of the models for the purposes of making the models comparable.

G. Model Performance Evaluation

The four ML models were evaluated for their performance by comparing their predictions with the measured AQI values in the testing data set. The model's complexity, accuracy of predictions, and goodness of fit were assessed using several statistical methods. The lower the error value, the more accurate the prediction, and the higher the R^2 and Adjusted R^2 values, the closer the model's fit to reality. The model selection was also compared with AIC and SBC/BIC.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\,y_i - \hat{y}_i|$$

where y_i and $\hat{y}_i$ represent observed and predicted AQI values, respectively. Lower MAE indicates better accuracy.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

MSE gives greater weight to larger prediction errors, with lower values indicating improved performance.

Coefficient of Determination (R^2)

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

where $\bar{y}$ is the mean observed AQI. Higher R^2 indicates better explanatory capability.

Adjusted R^2

$$R^2_{\mathrm{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p is the number of predictors. Higher Adjusted R^2 indicates better fit after accounting for model complexity.

Akaike Information Criterion (AIC)

$$\mathrm{AIC} = 2k - 2\,ln(L)$$

where k is the number of estimated parameters and L is the maximum likelihood. Lower AIC indicates a preferable model.

$$\mathrm{SBC} = k\,ln(n) - 2\,ln(L)$$

where n is the sample size. Lower SBC/BIC indicates a better balance between model fit and complexity.

H. Comparative Analysis of Prediction Models

The predictive capabilities of LR, LASSO Regression, RF, and XGBoost for AQI prediction were evaluated in a comparative comparison. For the sake of fairness, we compared the four models using the same testing dataset and the same predictor variables. The model's performance was evaluated with MAE, MSE, R^2, Adjusted R^2, AIC, and SBC/BIC. Lower values of MAE and MSE show less prediction error, whereas higher values of R^2 and Adjusted R^2 show a better fit to the model. The evaluation of the balance between model fit and complexity was done using AIC and SBC/BIC. The comparative numerical results are presented in the Results section.

I. Prediction–Observation and Residual Analysis

Prediction–observation and residual analyses were performed to further assess the reliability and consistency of the developed AQI prediction models. The predicted AQI values generated from the testing dataset were compared with the corresponding observed AQI values to determine how closely each model reproduced the actual air-quality conditions. Residuals were examined to identify systematic prediction errors, deviations, or potential model bias. The residual for each observation was calculated as:

$$e_i = y_i - \hat{y}_i$$

where e_i represents the residual, y_i is the observed AQI, and $\hat{y}_i$ is the predicted AQI. Residuals distributed closely around zero indicate better prediction consistency.

III. Results And Discussion

A. Random Forest (RF) Model

The AQI model achieved a good prediction accuracy with the predicted values mostly located around the 1:1 reference line as seen in Figure 3 for the RF model. The model had a relatively low prediction error with an MAE of 7.626, MSE of 108.288, and (R^2) of 0.774, demonstrating that it had good explanatory power. $PM_{2.5}$ was the most important variable with mean error increase of about 12.7, followed by CO (6.7) and PM_{10} (4.2). Importance of NO_2 was around 3.9 and that for Ozone was around 3.2. The contribution was the lowest for SO_2, with about 0.2.

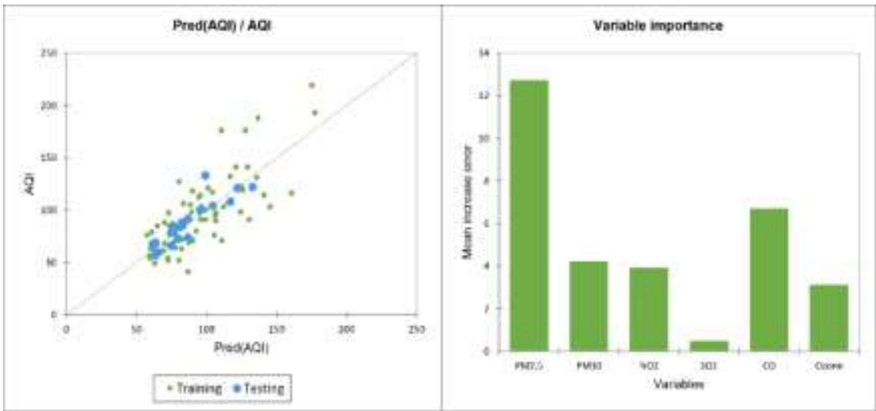

Figure 3. Statistical plots for RF model

B. XGBoost Model

As illustrated in Figure 4, there was good agreement between the predicted and observed AQI values with the most points lying near the diagonal reference line for the XGBoost model both during training and testing. The model had an MAE of 12.963, MSE of 241.748, and R^2 of 0.495, which were moderate for the model. Based on Gain, the variable importance analysis reveals that the contribution of $PM_{2.5}$ is the highest among the variables followed by NO_2 and PM_{10}. Moderate contributions came from CO and Ozone while comparatively lower contribution came from SO_2 in Gain and Cover. When it comes to Frequency, $PM_{2.5}$ and SO_2 were more common amongst the trees. From the results of the XGBoost model, it is seen that $PM_{2.5}$ is the most important predictor.

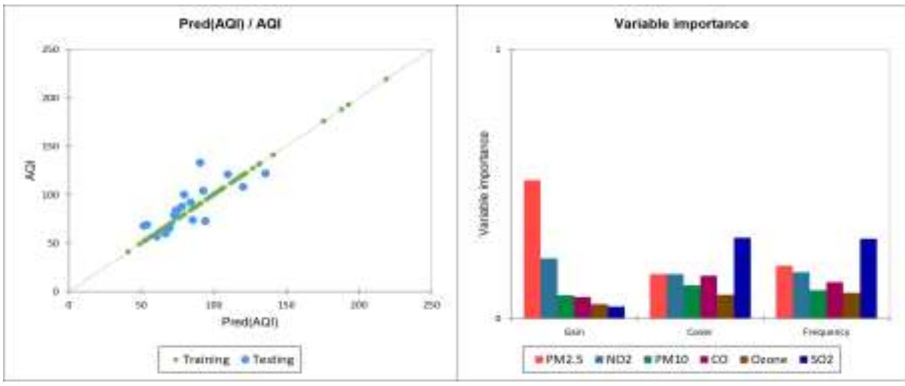

Figure 4. Statistical plots for XGBoost model

C. LASSO Regression Model

LASSO regression showed relatively poor performance in predicting AQI, with considerable differences between predicted and observed values, particularly at higher AQI levels. As shown in Figure 5, the model produced an MAE of 17.994, an MSE of 637.223, and an R^2 of -0.332, indicating limited predictive capability. The variable-importance analysis identified CO as the most influential predictor, with an absolute regression coefficient of approximately 23.5, whereas $PM_{2.5}$ had a substantially smaller coefficient of about 1.1. This indicates that CO contributed considerably more to the model than $PM_{2.5}$, while the remaining predictors were reduced through L_1 regularization. Overall, LASSO demonstrated the weakest AQI prediction performance among the evaluated models.

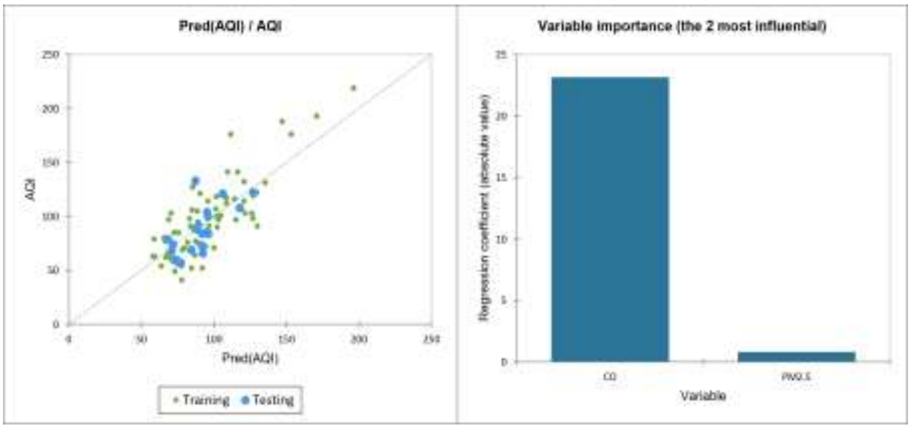

Figure 4. Statistical plots for LASSO Regression model

D. Linear Regression Model

The results of the LR model exhibited moderate agreement between the predicted and observed AQI values, exhibiting some scatter around the 1:1 reference line. The model achieved an MAE of 13.772, MSE of 286.416, and (R^2) of 0.401. The standardized coefficient plot shows that PM_{10} had the highest positive coefficient (around 0.35), followed by $PM_{2.5}$ (0.21) and NO_2 (0.26). CO also had a positive contribution of around 0.20, but SO_2 had a small positive coefficient. On the contrary, the coefficient in the case of O_3 was negative, around –0.18. The overall performance of the model was satisfactory with larger differences at higher AQI levels.

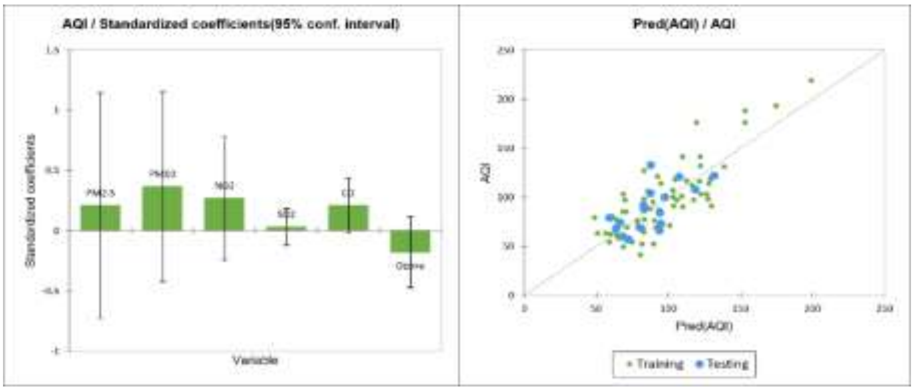

Figure 4. Statistical plots for LR model

E. Comparative Performance of ML Models

The comparative results in Table 2 indicate that RF provided the strongest AQI prediction, recording the minimum MAE of 7.626 and MSE of 108.288, along with the maximum R^2 of 0.774 and Adjusted R^2 of 0.650. XGBoost demonstrated the next-best performance, yielding an MAE of 12.963, an MSE of 241.748, and an R^2 value of 0.495. The Moderate performance for LR was obtained with the MAE of 13.772 and (R^2) of 0.401. LASSO performed weakest, with an MAE of 17.994, MSE of 637.223, and negative (R^2) of −0.332. RF also had the minimum value of AIC (96.326) and SBC (101.668), which was the best model-selection criteria among the models that were tested.

TABLE I. PERFORMANCE OF VARIOUS MODELS

Performance metrics	**Linear regression**	**Random forests**	**Extreme Gradient Boosting**	**LASSO regression**
MAE	13.772	7.626	12.963	17.994
MSE	286.416	108.288	241.748	637.223
R^2	0.401	0.774	0.495	-0.332
Adjusted R^2	0.075	0.650	0.219	-1.059
Akaike's AIC	113.834	96.326	110.782	128.228
Schwarz's SBC	119.176	101.668	116.124	133.570

IV. Conclusion and Future Scope

This study developed and evaluated ML-based models for predicting the AQI using air-quality observations from a high-traffic monitoring station in Lucknow, India. This analysis was based on 88 observations obtained during a three-month period from January to March 2024 and consisted of $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO and O_3 as predictor variables. The four regression models namely “LR, LASSO Regression, RF, and XGBoost” were assessed based on “MAE, MSE, R^2, Adjusted R^2, AIC and SBC”. Of all the methods evaluated, RF had the best predictive performance, with the lowest MAE of 7.626 and MSE of 108.288, and the highest R^2 of 0.774 and Adjusted R^2 of 0.650. It also had the smallest AIC (96.326) and SBC (101.668) values, which means it had good model selection performance. The model XGBoost demonstrated a moderate predictive ability, with an MAE of 12.963 and an R^2 of 0.495, while LR performance was less impressive, with an MAE of 13.772 and an R^2 of 0.401. LASSO showed the poorest result with an MAE of 17.994 units and R^2 of −0.332. Overall, the results show the capability of ensemble-based learning to predict AQI in the analysed urban environment and in general the efficiency of RF. Long term, multi-station, meteorological and temporal data should be included in future studies to build more robust and generalizable AQI forecasting models.

References

[1] Iqbal, M. Chapter 5-a Cloudless-Sky Atmosphere and its Optics (Academic Press, UK, 1983).

[2] Chen, Y, Chen, R, Chen, Y, Don, X, Zhu, J, Liu, C, Donkelaar, A, Martin, R.V., Li, H, Kan, H, Jiang, Q, Fu, C, . The prospective effects of long-term exposure to ambient pm2.5 and constituents on mortality in rural east china. Chemosphere 280, 130740 (2021).

[3] Darrow, L. A., Hess, J, Rogers, C.A., Tolbert, P. E., Klein, M, Sarnat, S.E., Ambient pollen concentrations and emergency department visits for asthma and wheeze. J. Allergy Clin. Immunol. 130(3), 630–638 (2012).

[4] Dennekamp, M. & Abramson, M. J. The effects of bushfire smoke on respiratory health. Respirology 16(2), 198–209 (2011).

[5] Ram, K., Singh, S., Sarin, M., Srivastava, A. & Tripathi, S. Variability in aerosol optical properties over an urban site, kanpur, in the indo-gangetic plain: A case study of haze and dust events. Atmos. Res. 174, 52–61 (2016).

[6] Mathur, V., Srivastava, D., Pandey J. and Mishra,V., “Soft computing techniques for atmospheric pollution and traffic emission prediction” Environmental Modelling & Software, Vol.197, 106828. (SCI, Scopus Indexed) (2025)

[7] WAQ, W. I. World’s air pollution: Real-time air quality index. https:// waqi. info/ (2020). Kinney, P. L. Climate change, air quality, and human health. Am. J. Prev. Med. 35(5), 459–467 (2008).

[8] Kinney, P. L. Climate change, air quality, and human health. Am. J. Prev. Med. 35(5), 459–467 (2008).

[9] Gupta, N., Yadav, V.K., Gacem, A, Dossari, M.A., Yadav, K.K., El-Gawaad, S.A., Khedher, N.B., Chaudhary, N, Kumar, P, Cavalu, S, Deleterious effect of air pollution on human microbial community and bacterial flora: A short review. Int. J. Environ. Res. Public Health 19(23), 15494 (2022).

[10] Carrero, J., Goienaga, N., De Vallejuelo, S.F.-O., Arana, G. & Madariaga, J. Classification of archaeological pieces into their respective stratum by a chemometric model based on the soil concentration of 25 selected elements. Spectrochim. Acta, Part B 65(4), 279–286 (2010).
[11] Kumar, A., Dhawan, S, Kumar, M.V., Khare, M, Nagendra, S,M.S., Dubey, S.K., Mehta, D.S., Detection and identification of shape, size, and concentration of particulate matter in ambient air using bright field microscopy-based system. Atmos. Pollut. Res. 2023, 101913 (2023).
[12] Triantafyllou, E., Diapouli, E., Tsilibari, E.M., Adamapoulas, A.D., Biskas, G., Eleftheriadis, K., Assessment of factors influencing pm mass concentration measured by gravimetric & beta attenuation techniques at a suburban site. Atmos. Environ. 131, 409–417 (2016).
[13] Wang, Z., Calderon, L., Pattan, A. P., Allacci, M.S., Senick, J., Wener, R., Andrews, C.J., Mainelis, G., Comparison of real-time instruments and gravimetric method when measuring particulate matter in a residential building. J. Air Waste Manage. Assoc. 66(11), 1109–1120 (2016).
[14] Watson, J. G., Tropp, R. J., Kohl, S. D., Wang, X. & Chow, J. C. Filter processing and gravimetric analysis for suspended particulate matter samples. Aerosol Sci. Eng. 1, 93–105 (2017).
[15] Yadav, A., Srivastava, D. and Mathur, V. (2024), "Expository Assessment of Air Quality Scenario with Sentinel-5 Precursor TROPOMI Explorer Sensor" Nature Environment and Pollution Technology Journal . Vol 24(2), pp1-12. ISSN: 0972-6268 (Scopus Indexed)
[16] Agarwal, S., Sharma, S., Suresh, R., Rahman, M.H., Vranckx, S., Maiheu, B., Blyth, L., Janssen, S., Gargava, P., Shukla, V.K., Batra, S., Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. Sci. Total Env. 735, 139454 (2020).
[17] Maleki, H., Saraashian, A., Gaudarzi, G., Babali, Z., Birgani, Y.T., Rahmati, M., Air pollution prediction by using an artificial neural network model. Clean Technol. Environ. Policy 21, 1341–1352 (2019).
[18] Patni, J. C. & Sharma, H. K. Air quality prediction using artificial neural networks. In 2019 international Conference on Automation, Computational and Technology Management (icactm) 568–572 (2019).
[19] C. R. Kataria, A. Vaish, M. Singh, A. D. Mishra, D. Srivastava and S. W. A. Rizvi, "Predicting Air Quality Index: Deep Learning Perspective," 2026 5th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0, Raigarh, India, 2026, pp. 1-5, doi: 10.1109/OTCON69332.2026.11629795 (Scopus Indexed)
[20] Cristianini, N. & Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge University Press, Cambridge, 2000).
[21] Hossain, M. T., Hassain, A., Meem, M., Monir, M.F., Miah, M.S.U., Sarwar, T.B., Impact of covid-19 lockdowns on air quality in bangladesh: Analysis and aqi forecasting with support vector regression. In 2023 4th International Conference for Emerging Technology (incet) 1–6 (2023).
[22] Leong, W., Kelani, R. & Ahmad, Z. Prediction of air pollution index (api) using support vector machine (svm). J. Environ. Chem. Eng. 8(3), 103208 (2020).
[23] Liu, C.-C., Lin, T.-C., Yuan, K.-Y. & Chiueh, P.-T. Spatio-temporal prediction and factor identification of urban air quality using support vector machine. Urban Climate 41, 101055 (2022).
[24] Lu, W., Wang, W., Leung, A.Y.T., Lo, S.M., Yuen, R.K.K., Xu, Z., Air pollutant parameter forecasting using support vector machines. In Proceedings of the 2002 International Joint Conference on Neural Networks, ijcnn'02 (cat. no. 02ch37290), vol. 1 630–635 (2002).
[25] Mathur, V., Srivastava, D., Singh, V., and Kumar, A. (2026). Support Vector Machine Based Air Quality Prediction of a Mid-Sized Urban City. In: Reddy, K.R., Nagabhatla, N., Singh, S.K., Soundharajan, B., Kumari, M., Jain, S.K. (eds) Recent Advances in Geomatics,
Water Resources and Environmental Engineering. TRACE 2024 2024. Lecture Notes in Civil Engineering, vol 727. Springer, Singapore. https://doi.org/10.1007/978-981-95-0736-8_19 (Scopus Indexed)
[26] Hasnain, A., Sheng, Y., Hashmi, M.Z., Bhatti, U.A., Ahmed, Z., Zha, Y., Assessing the ambient air quality patterns associated to the covid-19 outbreak in the yangtze river delta: A random forest approach. Chemosphere 314, 137638 (2023).
[27] Zimmerman, N., Presto, A.A., Kumar, S.P.N., Gu, J., Hauryliuk, A., Robinson, A.L., Subraminium, R., A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos. Meas. Tech. 11(1), 291–313 (2018).
[28] Qin, Z., Cen, C. & Guo, X. Prediction of air quality based on knn-lstm. J. Phys: Conf. Ser. 1237, 042030 (2019).
[29] Utku, A. & Can, U. Machine learning-based a comparative analysis for air quality prediction. In 2022 30th Signal Processing and Communications Applications Conference (siu) 1–4 (2022).
[30] Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J.C.,Barends, R., Biswas, R., Boixa, S., Brandao, F.G.S.L., Buell, D.A., Burkett, B., Chen, Y., Chen, Z., Chiaro, B., Collins, R., Courtney, W., Dunsworth, A.,

Farhi, E., Faxen, B., Fowler, A., Gidney, C., Giustina, M., Graff, B., Guerin, K., Martinis, J.M., Quantum supremacy using a programmable superconducting processor. Nature 574(7779), 505–510 (2019).
[31] Schuld, M. & Petruccione, F. Supervised Learning with Quantum Computers 17th edn. (Springer, Uk, 2018).
[32] Kopczyk, D. Quantum machine learning for data scientists. arXiv preprint arXiv: 1804. 10068 (2018).
[33] Outeiral, C., Strahm, M., Shi, J., Morris, G.M., Benjamin, S.C., Deanne, C.M., The prospects of quantum computing in computational molecular biology. Wiley Interdiscipl. Rev. Comput. Mol. Sci. 11(1), e1481 (2021).
[34] Aithal, P. Advances and new research opportunities in quantum computing technology by integrating it with other icct underlying technologies. Int. J. Case Stud. Business, IT Educ. (IJCSBE) 7(3), 314–358 (2023).
[35] Bindhu, V. Cyber security analysis for quantum computing. J. IoT Soc. Mobile Anal. Cloud 4(2), 133–142 (2022).
[36] Kapoor, N., Kumar, A., Kumar, A., Kumar, A., Arora, H.C., Jahangir, H., Quantum computing for indoor environmental quality: A leapfrogging technology. In Handbook of Research on Quantum Computing for Smart Environments 191–216 (IGI Global, 2023).
[37] Zheng, K., Van Griensven, J. & Fraser, R. A quantum machine learning approach to spatiotemporal emission modelling. Atmosphere 14(6), 944 (2023).
[38] Wang, J., Li, H., Yang, H. & Wang, Y. Intelligent multivariable air-quality forecasting system based on feature selection and modified evolving interval type- 2 quantum fuzzy neural network. Environ. Pollut. 274, 116429 (2021).
[39]Hashi, Abdinasir Ismael, and Mr Osman Abdullahi Jama. "Evaluating the Performance of AI-Based Software Tools in Intelligent Decision-Making Systems." International Journal of Computing and Engineering 7, no. 22 (2025): 1-20.
[40]Shanker, Ravi. "Edge-Accelerated Analytics for Encrypted Network Traffic using Optimized Machine Learning." In 2025 International Conference on Computing Technologies & Data Communication (ICCTDC), pp. 1-7. IEEE, 2025