Transforming Air Quality Index Prediction Using Machine Learning: Key Insights from the Taj Trapezium Zone (TTZ)

Swati Varshney
Department of Computer Application, Invertis University, Bareilly, Uttar Pradesh, India
swativarshney2403@gmail.com (corresponding author)

Jitendra Nath Shrivastava
Department of Computer Science and Engineering, Invertis University, Bareilly, Uttar Pradesh, India
jitendranathshrivastava@gmail.com

Neha Gupta
School of Computer Science and Information Technology, Symbiosis University of Applied Science, Indore, Madhya Pradesh, India
neha.gupta@suas.ac.in

Received: 16 May 2025 | Revised: 23 June 2025 | Accepted: 9 July 2025

 DOI: https://doi.org/10.48084/etasr.12168

ABSTRACT

The exponential growth of air pollution is alarming for today's modern society. The Air Quality Index (AQI) is the main index used to gauge the severity of air pollution. In light of the recent Commission for Air Quality Management (CAQM) mandate for three days' advance upgrading of the Graded Response Action Plan (GRAP) stage, this research focuses on predicting the AQI to minimize negative economic losses through the correct adoption of the GRAP stage. The current research paper aims to predict the AQI in the Taj Trapezium Zone (TTZ) using machine learning algorithms. In this study, the main air contaminants that affect air quality—PM2.5, PM10, CO, NO_2, NH_3, NO_x, O_3, SO_2, benzene, and toluene—along with meteorological factors such as temperature, humidity, wind speed, wind direction, and pressure, are used for data analysis and AQI prediction using machine learning algorithms. The study also aims to identify the most dominant pollutants in the TTZ area. The data source is a real-time dataset from air quality monitoring stations operated by the Central Pollution Control Board (CPCB) in Agra, one of the key cities within the TTZ area. This study uses four machine learning algorithms: AdaBoost, XGBoost, CatBoost, and LightGBM to calculate and compare AQI prediction accuracy. Four statistical performance metrics are used, namely: R^2 score, Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE). The findings show that the XGBoost algorithm forecasts the AQI with the highest accuracy and best performance in the TTZ region. The study also found that PM10 is the most dominant pollutant in the TTZ area. This indicates that more stringent control measures are needed to curb PM10 pollution and improve the AQI. The use of these predicted AQI values may help CAQM in real-time GRAP stage decision making.

Keywords-Air Quality Index (AQI); Graded Response Action Plan (GRAP); machine learning; Commission for Air Quality Management (CAQM); ensemble techniques; XGBoost; AdaBoost; LightGBM; CatBoost

I. INTRODUCTION

Air pollution has become a significant environmental issue in numerous parts of the world, especially in developing nations. This issue becomes more critical in urban areas because of high population density, industrialization, and transportation. The air pollution menace is alarmingly high due to ever-increasing industrialization and population growth. There are both immediate and long-term health implications of air pollution. This amounts to an economic loss in productivity. To measure air quality in a homogenous parameter, the Air Quality Index (AQI) is a key measure of air pollution. It is measured against eight pollutants, namely PM2.5 (particulate matter having a diameter of 2.5 microns or less), PM10 (particulate matter having a diameter of 10 microns or less), CO (carbon monoxide), NO_2 (nitrogen dioxide), NH_3 (ammonia), NO_x (nitrogen oxides), O_3 (ozone), and SO_2 (sulphur dioxide). The air quality category, which is taken from

the Central Pollution Control Board (CPCB) (national AQI) [1, 2], and its health impact is shown in Table I.

TABLE I. INDIA AQI CATEGORIES AND THEIR HEALTH IMPACTS

AQI	Remark	Color Code	Possible health impacts	GRAP stages
0-50	Good		Minimal impact	–
51-100	Satisfactory		Minor breathing discomfort to sensitive people	–
101-200	Moderate		Breathing discomfort to the people with lungs, asthma, and heart diseases	–
201-300	Poor		Breathing discomfort to most people on prolonged exposure	GRAP 1
301-400	Very poor		Respiratory illness on prolonged exposure	GRAP 2
401-450	Severe		Affect healthy people and seriously impacts those with existing diseases	GRAP 3
451 and above	Severe plus		Extreme health impact	GRAP 4

Governments are taking significant steps to minimize AQI deterioration. In India, the Commission for Air Quality Management (CAQM) is the apex body responsible for AQI improvement initiatives. It uses the tool known as Graded Response Action Plan (GRAP), where the GRAP stages are triggered based on the AQI movement upward or downward. As the AQI worsens, the measures intensify, involving seizure or closure of economic activities in a phased manner or vice versa. Thus, each GRAP stage has an increasing economic cost. Recently, CAQM has decided to upgrade/downgrade GRAP stage three days in advance of AQI changes. This necessitates the ability to forecast ahead of time, minimizing economic losses. Historical monuments are profoundly affected by deteriorating air quality. Among the most iconic structures grappling with this environmental challenge is the Taj Mahal, a breathtaking masterpiece renowned for its pristine white marble façade. Unfortunately, this exquisite marble is succumbing to a yellowing discoloration, a direct consequence of the relentless decline in air quality. CAQM has included the Taj Mahal in its GRAP response program.

This research considers the Taj Trapezium Zone (TTZ), a protective area spanning approximately 10,400 km² around the Taj Mahal, designed to shield this symbol of love from the encroaching threats of environmental pollution. The zone serves not only as a sanctuary for the monument but also as a crucial buffer against the pollutants that endanger its timeless beauty.

The first steps in developing effective air pollution mitigation strategies are monitoring and forecasting. Traditional methods employ statistical and mathematical models to achieve this. However, these models often produce inaccurate and inefficient results when predicting the AQI. These models treat fresh and old data equally and use basic mathematics.

Advanced technology has enabled the development of various methods for forecasting air pollution, including deterministic, statistical, and machine learning strategies. Machine learning techniques are being adopted at an increasing rate, particularly for predicting time-series data. This research paper employs machine learning techniques to accurately predict the AQI and prominent pollutants in the TTZ area. The study uses real data from three CPCB monitoring stations in Agra from January 2022 to December 2024 (36 months) [3]. The paper also uses correlation matrices to identify the most dominant pollutant.

II. RELATED WORK

Authors in [4] effectively used eight robust machine learning models for AQI prediction. This paper provides precise hourly forecasts for Azamgarh, India. The authors concluded that XGBoost is the best model for predicting AQI. Authors in [5] assessed the AQI for the smart cities of Ahmedabad, Delhi, Lucknow, Gurugram, and Mumbai in India. This research compared several machine learning methods, such as Random Forest Regression (RFR), decision tree regression, linear regression, XGBoost, and a proposed hybrid model that integrates random forest and XGBoost techniques. Determining which pollutants are most impacted in smart cities is facilitated by this study. Authors in [6] utilized data from Shijiazhuang, Hebei Province, China to forecast the AQI. The XGBoost model achieved the best performance on the hourly scale. Authors in [7] showed that a machine learning approach using six years of data collected from 23 cities in India was effective. The authors reported that the Gaussian Naive Bayes model achieved the highest accuracy, whereas the XGBoost model outperformed the other machine learning algorithms.

Authors in [8] selected ten meteorological data points for AQI prediction in Jinan, China. The paper assessed the effectiveness of various machine learning models using a classification approach. Among these models, LightGBM emerged as the most effective. Authors in [9] proposed the integration of statistical techniques with machine learning approaches to predict the AQI. To forecast the annual PM2.5 concentration in Hyderabad, India, the paper employed time series analysis, regression, and Adaboost. Authors in [10] showed that applying artificial intelligence methods provides promising results for AQI forecasting in Taiwan. This study demonstrated that the machine learning methods AdaBoost and stacking ensemble can outperform popular methods in the literature, such as Support Vector Machine (SVM), random forest, and Artificial Neural Networks (ANNs). AdaBoost and stacking ensemble can be considered new and superior alternatives for AQI forecasting. Authors in [11] proposed a new stacking ensemble model that can forecast PM2.5 concentrations in Beijing and Istanbul. Several machine learning models where integrated to create the suggested model. The authors concluded that the stacking ensemble model best predicts the AQI.

Authors in [12] used a deep learning solution to forecast PM2.5 levels in Beijing and other cities in China. This research utilizes various deep learning techniques to effectively predict outcomes: Convolutional Neural Network (CNN), Gated

Recurrent Unit (GRU), Bidirectional GRU (Bi-GRU), Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and a hybrid model that combines a CNN and an LSTM. Authors in [13] selected Keelung City and Xizhi District in New Taipei City as the locations for their research. The authors employed five different machine learning models to forecast the results: Deep Neural Network (DNN), M5 Decision Tree (M5P) algorithm, M5 Rules Decision Tree (M5Rules) algorithm, Alternating Model Tree (AMT), and Multiple Linear Regression (MLR). The authors concluded that hybrid prediction systems provide greater accuracy than individual models when forecasting future PM2.5 concentrations. Authors in [14] proposed a machine learning model that combines the decision tree method for prediction with Grey Wolf Optimization (GWO). The air quality information of major cities in India was obtained from the Kaggle repository. The effectiveness of the model was evaluated and validated through various metrics and compared with other competing models. Authors in [15] applied machine learning techniques to forecast hourly air pollution levels on the basis of meteorological data from the preceding days. Emphasis was placed on various regularization methods to determine the most effective model. The authors introduced parameter-reducing formulations and consecutive-hour-related regularizations, which significantly enhance performance.

Authors in [16] compared three machine learning models—Support Vector Regression (SVR), RFR, and CatBoost Regression (CR)—across four different Indian cities: Hyderabad, Bangalore, Kolkata, and New Delhi using the Synthetic Minority Oversampling Technique (SMOTE). Authors in [17] analyzed various machine learning models such as the multilayer linear perceptron algorithm and the Gaussian air dispersion model. The paper analyzed the distribution of pollutants, their levels, and their proximity to the source using dummy data. Authors in [18] identified the most polluted areas in Jordan during the period 2017-2019 and evaluated the levels of various pollutants. By employing machine learning algorithms, this study identified the pollutants that influence the AQI. Authors in [19] utilized machine learning techniques to assess the levels of SO_2 in the air across Maharashtra, India. The paper examined different meteorological elements, including the combustion of fossil fuels and industrial operations. The paper concluded that the prediction for the entire state is not helpful as aggregate units; it should be carried out at smaller units or on a city basis. Authors in [20] analyzed the combination of a neural network and a boosting model using a Kaggle dataset. They concluded that machine learning models are effective for AQI prediction. Authors in [21] employed machine learning techniques in the desulfurization process in the oil industry. Authors in [22] introduced a new technique to forecast California's air quality. The authors used SVR with a radial basis kernel to forecast pollution and particle matter. Authors in [23] utilized two machine learning approaches: neural network and SVM. The data were obtained from the CPCB, Ministry of Environment, Forest and Climate Change, Government of India, for Delhi. The paper concluded that to forecast the AQI for Delhi, SVM has the highest accuracy. In [24], the authors evaluated three machine learning algorithms: random forest, ANN, and SVM. The result showed that these methods are reliable for AQI prediction. It was found that the most effective algorithm is random forest. Authors in [25] showed that a predictive model using SVM regression, geographically weighted regression, ANN and auto-regressive nonlinear neural network is effective at evaluating air pollutants on the basis of PM2.5 and PM10 concentrations in Tehran, Iran. Authors in [26] researched supervised learning algorithms in machine language among three types of models (unsupervised learning, supervised learning, and reinforcement learning) for air pollution prediction. They used sensor-based real-time data for their analysis.

III. METHODOLOGY

The methodology of this study involves data collection, dataset description, data preprocessing/cleaning, data balancing, dataset splitting, data training/testing, forecasting with four machine-learning algorithms, and evaluating the performance. Figure 1 depicts the flowchart of this methodology.

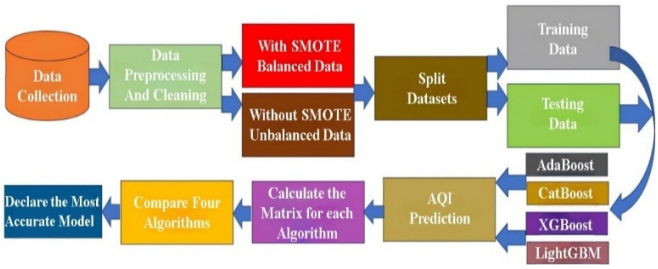

Fig. 1. Overall architecture of AQI prediction for TTZ.

In this section, we will delve into the detailed steps involved in executing this work, providing a comprehensive overview of the processes and methodologies utilized to achieve our objectives.

A. Data Collection

Poor air quality is a significant barrier to social growth. In the TTZ area, the Petha manufacturing sector, glass production facilities, oil refineries, and furnaces are the main contributors to air pollution. This study employs real-time data obtained from the Uttar Pradesh Pollution Control Board (UPPCB) from January 2022 to December 2024. There are six air monitoring stations in Agra. For this study, data from three of these stations—Awas Vikas (Upeida industrial area), Shastripuram (Rohta industrial area), and Manoharpur (dense population with a large number of small Industries)—were used to predict the AQI. This provided more varied and robust analytical data. The air monitoring stations were chosen due to the longest available dataset. The total dataset includes approximately 52,500 data points arranged in 3,289 rows and 16 columns. As presented in Table II, the dataset includes various features such as PM2.5, PM10, O_3, NH_3, SO_2, CO, NO_2, NO_x, temperature, humidity, wind speed, wind direction, pressure, benzene, toluene, and o-xylene.

TABLE II. SUMMARY OF STUDY AREA AND OBSERVED VARIABLES

Study area	Type	Variables
TTZ	Meteorological conditions	Temperature, humidity, wind speed, wind direction, pressure
	Criteria gases	O_3, NH_3, SO_2, CO, NO_2, NO_x, benzene, toluene, o-xylene
	Particulate	PM2.5, PM10

B. Data Description

The AQI unequivocally serves as a vital indicator of air quality, transforming pollutant concentrations into a single, authoritative number. AQI calculation employs a standardized formula to create sub-indices for key pollutants, including PM2.5, PM10, O_3, NO_2, SO_2, CO, and NH_3. The overall AQI value is determined by the highest sub-index among these pollutants, providing a definitive assessment of air quality.

For PM2.5, PM10, SO_2, NO_2, and NH_3, the average concentration is measured over a 24-hour period, whereas O_3 and CO are assessed based on the maximum concentration over an 8-hour span. To guarantee a reliable AQI calculation, it is imperative that at least one of the fine particulate pollutants (PM2.5 or PM10) is included, along with data from a minimum of three pollutants. This comprehensive methodology delivers a clear and impactful understanding of the air we breathe and its significant implications for health and well-being.

C. Data Preprocessing

The effectiveness of data visualization and the precision of machine learning algorithms heavily rely on the quality of the data. The data preprocessing steps are fundamental in enhancing the performance and overall effectiveness of machine learning algorithms. By carefully optimizing these steps, the study can significantly improve the quality of models and their outcomes. Preprocessing steps perform various operations such as handling missing values, reducing outliers, and feature engineering.

1) Handling Missing Values

Eliminating invalid and missing rows often leads to inaccuracies in results. It is beneficial to employ methods like multiple imputation and maximum likelihood, which effectively replace missing values with informed estimates. In this research, the researchers have chosen mean imputation to address missing data, which helps improve analysis and provides clearer insights. This approach contributes positively to the overall quality of the results. Upon examining Figure 2, it becomes evident that toluene exhibits the highest quantity of missing values among all the variables presented.

To address this significant data gap, the missing values for Toluene are primarily filled using the mean values. This approach ensures that the dataset maintains its integrity while minimizing the impact of absent data on subsequent analyses.

2) Outlier Detection and Removal

Outliers are critical data points that deviate significantly from the rest of a dataset, manifesting as either unusually high or low values. Their sources include measurement errors, data entry mistakes, experimental errors, and genuine natural variations. In this research paper, the authors have adeptly applied z-scores, interquartile range, and isolation forest techniques to effectively manage outliers. The code combines each technique's advantages and disadvantages to provide more reliable outlier detection and removal.

Station Name	0
Date	0
PM10	8
PM2.5	12
CO	0
SO2	1
NO2	2
NH3	60
Nox	0
O3	21
Temprature	6
Humidity	6
Wind Speed	6
Wind Direction	6
Pressure	12
Benzene	106
Toulene	1215
Oxlene	1204

Fig. 2. Missing values of the different air pollutants.

Time series analysis provides a deeper understanding of historical patterns and helps forecasting models become more accurate. The time series patterns for PM2.5, PM10, O_3, CO, NO_2, and benzene are displayed in Figures 3, 4, 5, 6, 7, and 8, respectively. They demonstrate how O_3 deviates greatly from the overall trend throughout the time period with multiple clear and sharp spikes. When compared to other pollutants, these spikes are more noticeable and regular. Compared to pollutants like PM2.5 or PM10, however, the observed patterns seem less regular, even though seasonal fluctuations are expected for O_3. Spikes are also considered outliers.

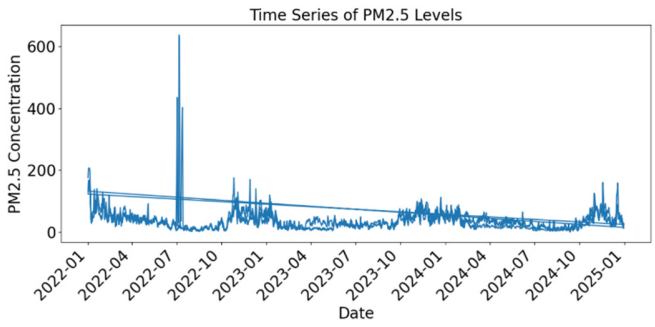

Fig. 3. Time series pattern of PM2.5 concentration levels.

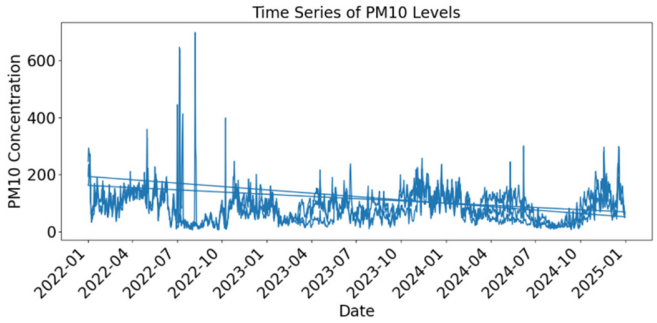

Fig. 4. Time series pattern of PM10 concentration levels.

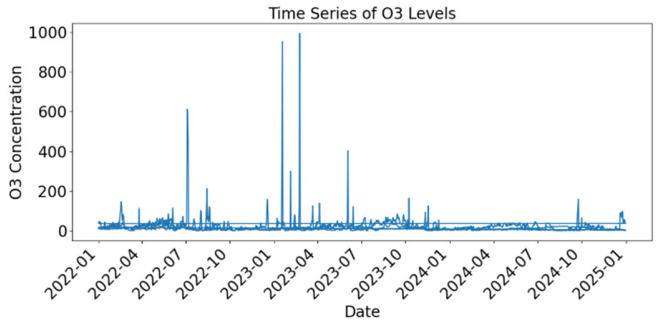

Fig. 5. Time series pattern of O_3 concentration levels.

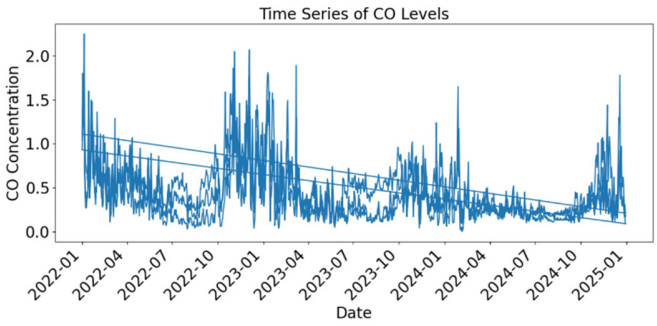

Fig. 6. Time series pattern of CO concentration levels.

3) Feature Engineering and Scaling

Feature engineering involves creating new features from existing ones or adding relevant variables not present in the original dataset. This can enhance model performance by capturing complex relationships. It includes transforming existing features for better representation, such as using mathematical functions, creating interaction terms, or generating polynomial features. In this research, log transformation and square root transformation have been used. It is also important to select the most relevant features through techniques like feature importance scores or correlation analysis, which helps reduce complexity and improve interpretability. Subsequent to feature engineering, the data are transformed into numerical features within a defined range, usually between 0 and 1, during the normalization step of feature scaling. In this study, the data were normalized using the MinMaxScaler. After this step, the data are prepared for splitting into training and testing sets and the scaled training data can then be used to train the model.

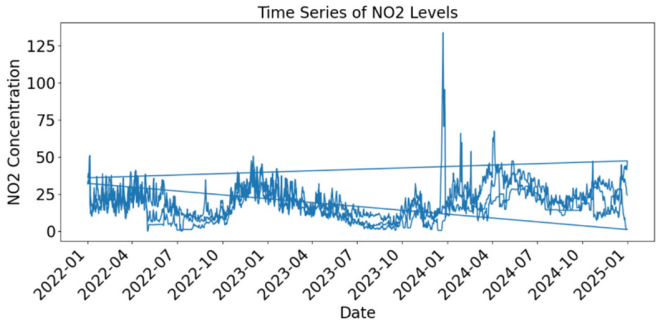

Fig. 7. Time series pattern of NO_2 concentration levels.

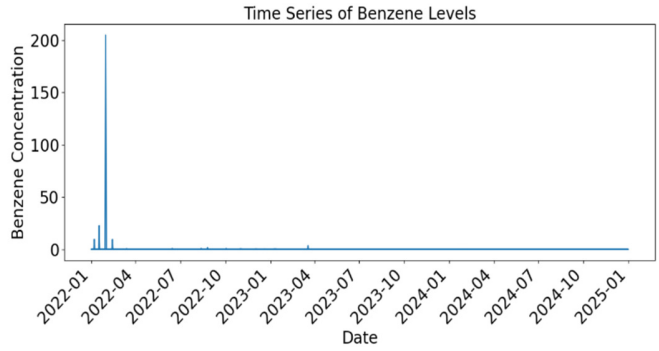

Fig. 8. Time series pattern of benzene concentration levels.

D. Applying the Synthetic Minority Oversampling Technique

SMOTE is a powerful algorithm that creates synthetic samples for the minority class, effectively balancing imbalanced datasets. This technique is essential for overcoming the problem of overfitting that can arise from random oversampling. In this research study, we utilize data both with and without SMOTE to predict outcomes, showcasing the significant advantages of this method.

E. Splitting the Dataset

In the current study, a strategic partitioning of the datasets into training and test sets, adhering to an effective 80:20 ratio is performed. This thoughtful division is based on empirical studies. It ensures that a robust 80% of the data is utilized to train the model, empowering it to learn from a rich and diverse set of examples. Conversely, the remaining 20% is reserved for rigorously testing the model's predictive accuracy against the original data. By systematically comparing the values predicted by sophisticated machine learning algorithms with the actual outcome, the process gives crucial insights into the model's effectiveness.

F. Applying Machine Learning Techniques

In this phase, machine learning is implemented on the cleansed and partitioned datasets. This research paper employs four algorithms: XGBoost, CatBoost, AdaBoost, and LightGBM. These algorithms are designed and programmed as models for evaluation purposes. The software used for this was developed in Python (version 3.10), mainly using the Pandas, Numpy, Seaborn, Matplotlib, and Scikit-learn packages.

G. Air Quality Index Prediction

Machine learning algorithms are employed to determine the AQI value. In this study, both tabular and scatter plot representations are utilized to depict the most accurate algorithms for AQI prediction in the TTZ area.

H. Calculation of Performance Matrix for each Machine Learning Algorithm

The accuracy of the AQI is assessed using a performance matrix that includes the R^2 score, Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE). Each of these matrices is described individually:

- The degree of agreement between the regression model and the actual data is shown by R^2. A better model has a higher R^2 value.
- MSE is a measure of how closely a fitted line resembles a set of data points; the closer the number is to the line, the better.
- Data density along the line of best fit is indicated by RMSE. The model can predict the data with reasonable accuracy if the RMSE values fall between 0.2 and 0.5.
- MAE calculates how far the observations deviate from the regression line's predictions in absolute terms. This should be a lower value.

I. Comparison between Machine Learning Algorithms

In this step, the datasets from three air monitoring stations of CPCB in Agra are compared in a performance matrix for each machine learning algorithm. All training and testing datasets are examined in comparison tabulation. A variety of datasets, both with and without SMOTE, are used to calculate accuracy. The best-fitting curve for each machine learning algorithm is shown in a scatter plot.

J. Final Comparative Study

The next step after tabulating the numbers is to compare each matrix value and identify the most accurate machine learning technique. For every air quality monitoring station, XGBoost is the most effective and precise machine learning algorithm for AQI prediction in the TTZ area.

IV. RESULTS AND ANALYSIS

Figure 9 shows the correlation matrix of AQI and pollutants. This correlation matrix helps select pertinent features for AQI forecasting models and assists in identifying the pollutants with the greatest impact on AQI in the specific area. The more influenced the pollutant, the higher the value on the red color scale; the less influenced the pollutant, the lower the value on the blue color. In this figure, PM10 has the biggest impact on AQI in the TTZ area. Higher correlation coefficient pollutants—whether positive or negative—are prime candidates for feature engineering since they are probably going to have a bigger impact on AQI.

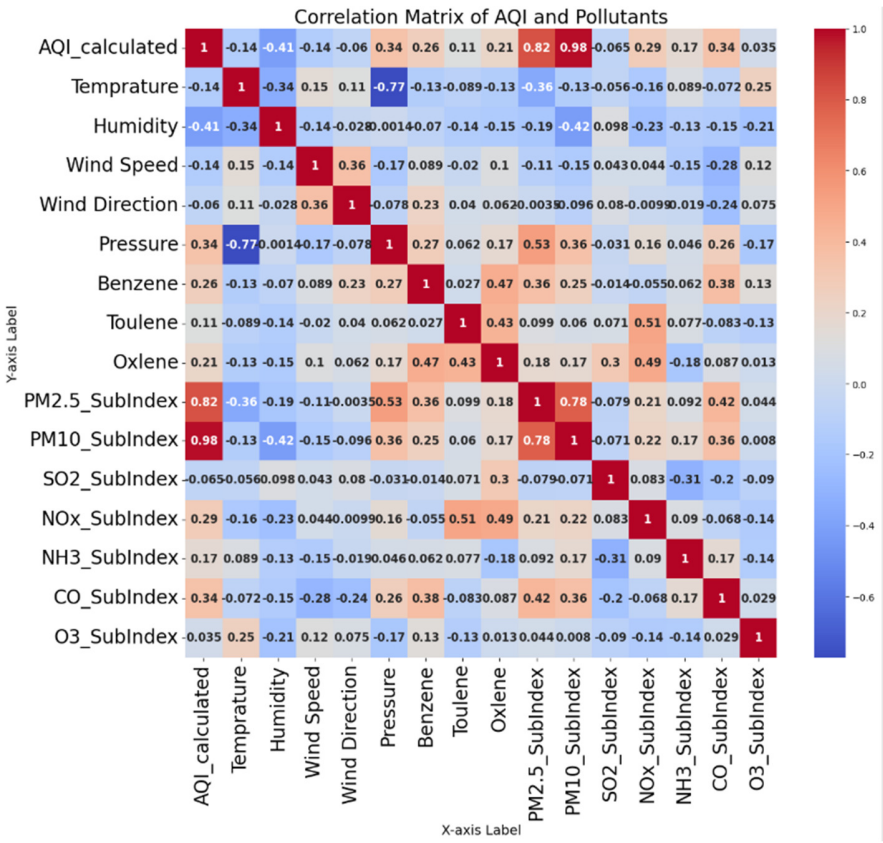

Fig. 9. Correlation matrix of AQI and pollutants.

AdaBoost, XGBoost, CatBoost, and LightGBM are the four algorithms used to get the experimental outcome. The data used by these algorithms were gathered from three air monitor stations and were analyzed separately for each station. Then, the data were combined and analyzed as a whole, obtaining the same results. The outcome demonstrates that the XGBoost machine learning algorithm is the best method for predicting AQI in the TTZ area.

Tables III and IV present a comparison of the four metrics for the training and testing datasets. From these tables, it can be concluded that XGBoost is the most effective and precise machine learning algorithm for AQI prediction in the TTZ area, as it has the lowest values of MSE, RMSE, and MAE, as well as the highest value of R^2.

TABLE III. COMPARISON OF FOUR MODELS WHEN AIR MONITORING STATIONS ARE COMBINED DURING TRAINING

Model	MSE	RMSE	R^2	MAE
AdaBoost	5.041434	2.245314	0.99494	1.803694
XGBoost	1.405921	1.185715	0.998589	0.848872
CatBoost	6.933324	2.633121	0.993041	1.937783
LightGBM	1.943207	1.39399	0.99805	0.969427

TABLE IV. COMPARISON OF FOUR MODELS WHEN AIR MONITORING STATIONS ARE COMBINED DURING TESTING

Model	MSE	RMSE	R^2	MAE
AdaBoost	7.662593	2.768139	0.992143	2.055372
XGBoost	4.794782	2.189699	0.995083	1.312817
CatBoost	10.49724	3.239945	0.989236	2.265891
LightGBM	5.213221	2.283248	0.994654	1.378861

Figures 10, 11, 12, and 13 illustrate the scatter plots for the four algorithms: AdaBoost, XGBoost, CatBoost, and LightGBM, respectively. In a scatter plot, the "line of best fit", also known as a trend line, is a straight line that most accurately depicts the connection between two variables in a collection of data points. XGBoost provides the most accurate "line of best fit" for AQI prediction compared to the other options.

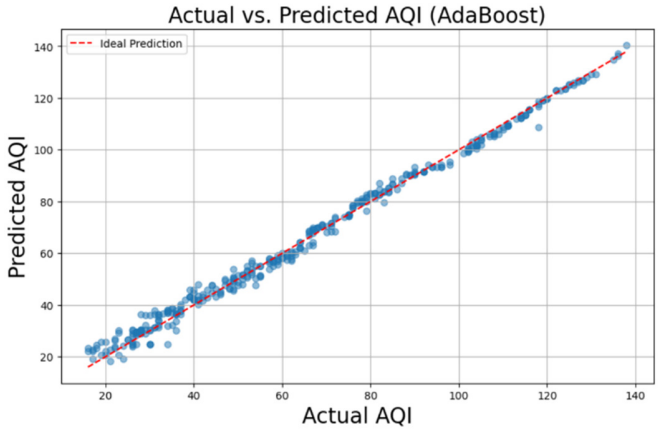

Fig. 10. Scatter plot for the AdaBoost algorithm.

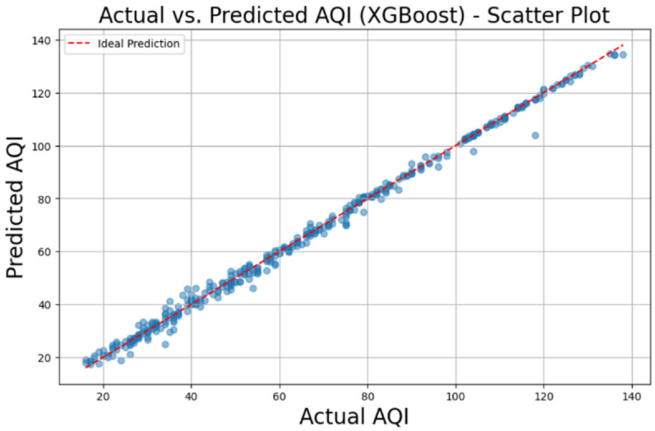

Fig. 11. Scatter plot for the XGBoost algorithm.

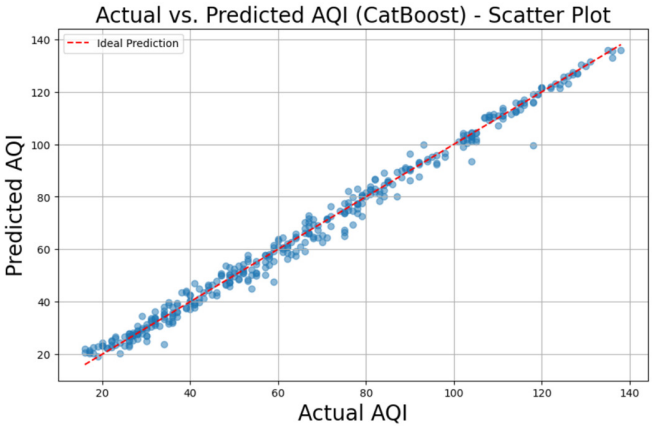

Fig. 12. Scatter plot for the CatBoost algorithm.

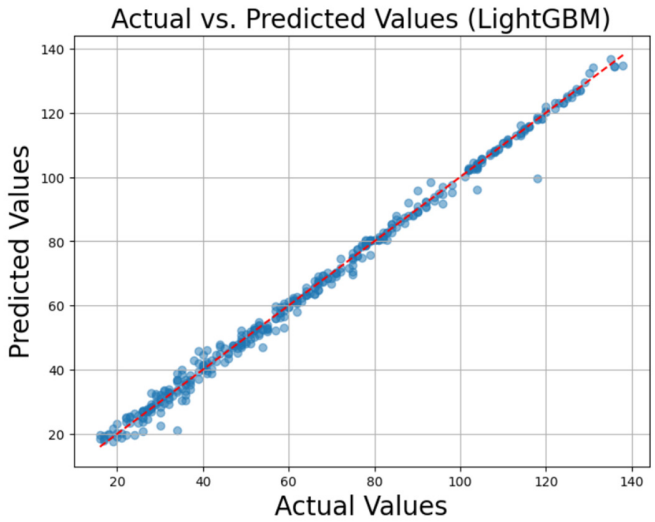

Fig. 13. Scatter plot for the LightGBM algorithm.

Figure14 shows a bar graph comparing the four algorithms, indicating that XGBoost the XGBoost algorithm performs better than the others. Consequently, the XGBoost algorithm stands out as the most efficient model selection in comparison with other models examined in this analysis.

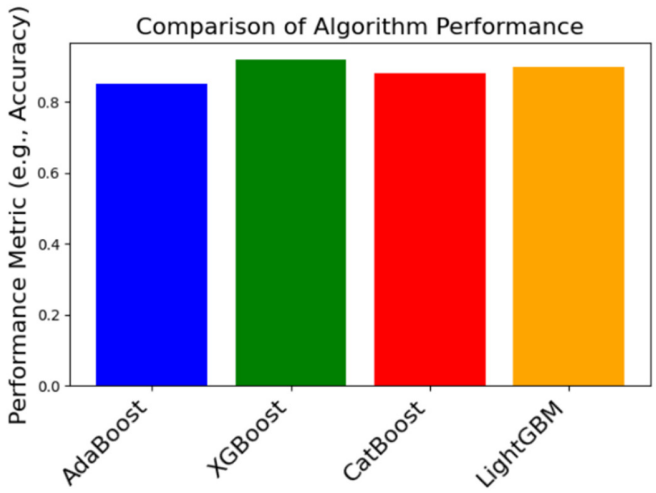

Fig. 14. Comparison of the performance of the four machine learning algorithms.

V. LIMITS OF THE STUDY

Currently, the study only uses data from Agra. To confirm these results, it is necessary to integrate data from other major cities in the TTZ area, such as Mathura and Firozabad. Combining these results with those from the Rajasthan desert/semi-arid area may provide a clearer picture for improved AQI prediction in the TTZ area.

VI. CONCLUSION

In light of the recent mandate by the Commission for Air Quality Management (CAQM) requiring a three-day advance escalation of the Graded Response Action Plan (GRAP) stage, this study focuses on predicting the Air Quality Index (AQI) to minimize economic losses through timely and appropriate GRAP implementation. The research examined AQI prediction in the Agra area using machine learning algorithms. A real-time dataset, spanning three years, was collected from the Central Pollution Control Board (CPCB) for the Agra area, comprising key contaminants and meteorological factors. The study also aimed to identify the most dominant pollutants in the Taj Trapezium Zone (TTZ).

Four machine learning algorithms —AdaBoost, XGBoost, CatBoost, and LightGBM—where employed to assess AQI prediction accuracy. Each model's performance was evaluated using four statistical metrics: R^2 score, Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE). The results demonstrate that the XGBoost algorithm achieved the highest prediction accuracy and overall performance.

The findings further reveal that PM10 is the most significant pollutant influencing AQI in the area. Therefore, it is essential to develop micro-area strategies focused on the control and reduction of PM10 emissions in the TTZ. Such targeted interventions will support the creation of localized, data-driven AQI control measures in the future.

ACKNOWLEDGMENT

The cooperation of the staff at Invertis University is acknowledged with deep appreciation.

REFERENCES

[1] "National Air Quality Index." Central Pollution Control Board. https://airquality.cpcb.gov.in/AQI_India/.

[2] "Stage-IV Of GRAP In Delhi." Pwonlyias. https://pwonlyias.com/current-affairs/stage-iv-of-grap-in-delhi/.

[3] "Continuous Stations Status." Central Pollution Control Board. https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard/caaqm-landing.

[4] A. Ansari and A. R. Quaff, "Advanced Machine Learning Techniques for Precise hourly Air Quality Index (AQI) Prediction in Azamgarh, India," *International Journal of Environmental Research*, vol. 19, no. 1, Nov. 2024, Art. no. 15, https://doi.org/10.1007/s41742-024-00684-5.

[5] G. Sharma, S. Khurana, N. Saina, Shivansh, and G. Gupta, "Comparative Analysis of Machine Learning Techniques in Air Quality Index (AQI) prediction in smart cities," *International Journal of System Assurance Engineering and Management*, vol. 15, no. 7, pp. 3060–3075, July 2024, https://doi.org/10.1007/s13198-024-02315-w.

[6] H. Jing and Y. Wang, "Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost," *E3S Web of Conferences*, vol. 165, May 2020, Art. no. 02014, https://doi.org/10.1051/e3sconf/202016502014.

[7] 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, no. 5, pp. 5333–5348, May 2023, https://doi.org/10.1007/s13762-022-04241-5.

[8] Q. Liu, B. Cui, and Z. Liu, "Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling," *Atmosphere*, vol. 15, no. 5, May 2024, Art. no. 553, https://doi.org/10.3390/atmos15050553.

[9] V. Devasekhar and P. Natarajan, "Prediction of Air Quality and Pollution using Statistical Methods and Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, pp. 927–937, 39/29 2023, https://doi.org/10.14569/IJACSA.2023.01404103.

[10] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, "Machine Learning-Based Prediction of Air Quality," *Applied Sciences*, vol. 10, no. 24, Dec. 2020, Art. no. 9151, https://doi.org/10.3390/app10249151.

[11] M. Emeç and M. Yurtsever, "A novel ensemble machine learning method for accurate air quality prediction," *International Journal of Environmental Science and Technology*, vol. 22, no. 1, pp. 459–476, Jan. 2025, https://doi.org/10.1007/s13762-024-05671-z.

[12] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi, "Air-pollution prediction in smart city, deep learning approach," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, Art. no. 161, https://doi.org/10.1186/s40537-021-00548-1.

[13] C.-C. Wei and W.-J. Kao, "Establishing a Real-Time Prediction System for Fine Particulate Matter Concentration Using Machine-Learning Models," *Atmosphere*, vol. 14, no. 12, p. 1817, Dec. 2023, https://doi.org/10.3390/atmos14121817.

[14] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Scientific Reports*, vol. 14, no. 1, Mar. 2024, Art. no. 6795, https://doi.org/10.1038/s41598-024-54807-1.

[15] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," *Big Data and Cognitive Computing*, vol. 2, no. 1, Mar. 2018, Art. no. 5, https://doi.org/10.3390/bdcc2010005.

[16] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *Journal of Environmental and Public Health*, vol. 2023, no. 1, 2023, Art. no. 4916267, https://doi.org/10.1155/2023/4916267.

[17] S. Simu *et al.*, "Air Pollution Prediction using Machine Learning," in *2020 IEEE Bombay Section Signature Conference*, Mumbai, India, 2020, pp. 231–236, https://doi.org/10.1109/IBSSC51096.2020.9332184.

[18] K. M. O. Nahar, M. A. Ottom, F. Alshibli, and M. M. A. Shquier, "Air quality index using machine learning – A Jordan case study," *Compusoft: An International Journal of Advanced Computer Technology*, vol. 9, no. 9, pp. 3831–3840, Sept. 2020.

[19] P. Bhalgat, S. Pitale, and S. Bhoite, "Air Quality Prediction using Machine Learning Algorithms," *International Journal of Computer Applications Technology and Research*, vol. 8, no. 9, pp. 367–370, Sept. 2019, https://doi.org/10.7753/IJCATR0809.1006.

[20] T. Madan, S. Sagar, and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking*, Greater Noida, India, 2020, pp. 140–145, https://doi.org/10.1109/ICACCCN51052.2020.9362912.

[21] H. A. Al-Jamimi, S. Al-Azani, and T. A. Saleh, "Supervised machine learning techniques in the desulfurization of oil products for environmental protection: A review," *Process Safety and Environmental Protection*, vol. 120, pp. 57–71, Nov. 2018, https://doi.org/10.1016/j.psep.2018.08.021.

[22] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California," *Complexity*, vol. 2020, no. 1, Aug. 2020, Art. no. 8049504, https://doi.org/10.1155/2020/8049504.

[23] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, "A Machine Learning Model for Air Quality Prediction for Smart Cities," in *2019 International Conference on Wireless Communications Signal Processing and Networking*, Chennai, India, 2019, pp. 452–457, https://doi.org/10.1109/WiSPNET45539.2019.9032734.

[24] D. Sanjeev, "Implementation of Machine Learning Algorithms for Analysis and Prediction of Air Quality," *International Journal of Engineering Research & Technology*, vol. 10, no. 3, pp. 433–538, Mar. 2021, https://doi.org/10.17577/IJERTV10IS030323.

[25] M. R. Delavar *et al.*, "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran," *ISPRS International Journal of Geo-Information*, vol. 8, no. 2, Feb. 2019, Art. no. 99, https://doi.org/10.3390/ijgi8020099.

[26] V. M. Madhuri, G. G. H. Samyama, and S. Kamalapurkar, "Air Pollution Prediction Using Machine Learning Supervised Learning Approach," *International Journal of Scientific and Technology Research*, vol. 9, no. 4, pp. 118–123, Apr. 2020.