IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

SmartAQI: Air Quality Prediction Using Machine Learning and Open Data

Mrs K. Nithiya, Yazhini Amirthavarshini S, Nithyasri R, Padma Priya N, Yogeshwari C
Department of Computer Science and Engineering
Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Tamil Nadu, India
yazhini.sivabalanofficial@gmail.com

Abstract: *Air pollution is one of the most critical environmental challenges affecting public health and economic sustainability. Continuous monitoring and accurate air quality prediction are essential for effective environmental management. This paper presents SmartAQI, an intelligent air quality monitoring and prediction system that integrates environmental data analytics with machine learning techniques. The proposed system utilizes historical air pollution datasets, India district-level geographic information, and real-time meteorological data obtained through the OpenWeather API to analyze air quality conditions across different regions of India. Major pollutants, including PM2.5, PM10, NO□, SO□, CO, and O□, are processed to compute the Air Quality Index (AQI) according to the guidelines established by the Central Pollution Control Board. To improve prediction capability, multiple machine learning models such as Linear Regression, Random Forest, and Time Series forecasting are implemented and evaluated. The system further provides an interactive visualization dashboard that displays live AQI values, pollutant concentration levels, predicted AQI values, and geographic air quality distribution through charts and maps. Results demonstrate that integrating environmental data with machine learning models enables reliable AQI prediction and effective spatial analysis of pollution trends. The SmartAQI framework offers a practical solution for environmental monitoring and can support researchers, policymakers, and the public in understanding and managing air quality conditions more effectively, which is reliable.*

Keywords: Air Pollution Prediction, Air Quality Index (AQI), Linear Regression, Machine Learning, OpenWeather API, Random Forest, Time Series Forecasting

I. INTRODUCTION

Air pollution has become a most significant environmental challenge affecting human health and economic sustainability worldwide. Due to rapid urbanization, industrialization, and increased vehicle usage, air quality in many urban regions has deteriorated.Poor air quality is associated with several health problems, including respiratory diseases, cardiovascular disorders, and reduced life span. The Air Quality Index (AQI) is widely used as a standard indicator to measure pollution levels and communicate their potential health impacts [20], [21].

Traditional air quality monitoring systems rely on physical monitoring stations that measure pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, CO, and O_3. Although these stations provide reliable measurements, their installation and maintenance costs are high, and their limited spatial coverage restricts large-scale monitoring [11].

Recent advancements in machine learning and data analytics have enabled the development of intelligence in air quality prediction systems. Machine learning algorithms can analyze large environmental datasets and identify complex relationships among pollutant concentrations and meteorological variables [4], [11]. Machine learning techniques such as Random Forest, Linear Regression, and Time Series forecasting have been widely used for air quality prediction due to their capability to model nonlinear and temporal patterns in environmental data [1], [7], [15].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, April 2026 Impact Factor: 8.2

Furthermore, the integration of real-time environmental data through web APIs has significantly improved the performance of modern monitoring systems. Real-time meteorological parameters such as temperature, wind speed, and atmospheric pressure influence pollutant dispersion and can improve prediction accuracy [12].

Motivated by these developments, this paper proposes SmartAQI, an intelligent air quality monitoring and prediction framework that integrates historical pollution data, district geographic coordinates, and real-time weather data obtained from the OpenWeather API. The system applies multiple machine learning models to predict AQI values and provides visualizations for monitoring air quality trends across different regions in India.

II. LITERATURE REVIEW

Air quality prediction has received significant attention due to increasing environmental pollution and its impact on human health. Various machine learning and statistical techniques have been proposed to analyze environmental data and forecast AQI values.

Several studies have used **Random Forest** and other ensemble learning methods for air pollution prediction. These algorithms effectively capture nonlinear relationships between pollutant concentrations and meteorological parameters, resulting in improved prediction accuracy compared to traditional statistical models **[1], [7], [17]**.

Regression-based models such as **Linear Regression** have also been widely used for AQI prediction due to their simplicity and interpretability **[3], [11]**. These models provide baseline prediction performance and help evaluate the effectiveness of machine learning techniques.

Another important approach is **time series forecasting**, which analyzes temporal patterns in historical pollution data to predict future air quality levels **[15]**. Time series methods are particularly useful for capturing seasonal trends and short-term variations in pollutant concentrations.

Recent environmental monitoring systems have also integrated big data for large-scale air quality prediction **[13]**. By combining multiple environmental datasets, machine learning models can improve the reliability of AQI prediction systems.

However, many existing studies focus on single prediction models or limited datasets. Therefore, integrated frameworks that combine multiple machine learning models and real-time environmental data are required. The proposed **SmartAQI system** addresses this limitation by integrating Random Forest, Linear Regression, and Time Series models with real-time data obtained through the OpenWeather API.

III. METHODOLOGY

In this section, the proposed **SmartAQI framework** for air quality monitoring and prediction is presented in Fig 1. The system integrates historical air pollution datasets, district geographic coordinates, and real-time information to estimate AQI values and predict future air quality conditions.

The overall architecture consists of several modules, including **data collection, data preprocessing, AQI calculation, machine learning, AQI prediction, and visualization**.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

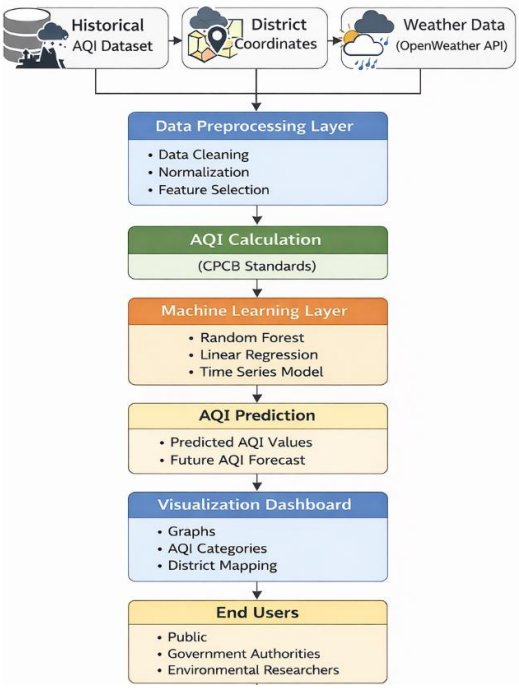

Fig. 1. System Architecture of SmartAQI System

A. Data Collection

The datasare collected from historical air pollution datasets, geographic datasets containing district latitude and longitude coordinates of India, and real-time data retrieved using the OpenWeather API.

Meteorological variables such as temperature, wind speed, and atmospheric pressure are included as model inputs because they influence pollutant dispersion and atmospheric conditions [12].

B. Data Preprocessing

Raw environmental datasets may contain missing values and inconsistencies. Therefore, preprocessing is performed to ensure data quality. The preprocessing steps include the removal of missing and duplicate records and theintegration of pollution datawith geographic coordinates and feature selection for relevant environmental variables and data normalization improves model performance by scaling data into a uniform range [19].

C. AQI Calculation

The Air Quality Index (AQI) is calculated according to Central Pollution Control Board (CPCB) guidelines [21] Table 1. The AQI value is determined by computing the sub-index for each pollutant and selecting the maximum value.

AQI = max ($IAQI_1$, $IAQI_2$, …, IAQI□)

The calculation considers six major pollutants:

PM2.5, PM10, NO_2, SO_2, CO, and O_3.

$$AQI = \max\left(IAQI_{SO2}, IAQI_{NO2}, IAQI_{O3}, IAQI_{PM10}, IAQI_{PM2.5}, IAQI_{CO}\right)$$

DOI: 10.48175/IJARSCT-32851

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

TABLE1. **CPCB AQI Classification Table**

AQI Range	Category	Health Impact
0–50	Good	Minimal impact
51–100	Satisfactory	Minor breathing discomfort
101–200	Moderate	Breathing discomfortfor sensitive groups
201–300	Poor	Breathing discomfort on prolonged exposure
301–400	Very Poor	Respiratory illness on prolonged exposure
401–500	Severe	Serious health effects

This classification helps interpret the predicted AQI values in terms of their health implications.

The AQI calculation is based on pollutant concentration breakpoint ranges defined by CPCB Table 2. These breakpoints determine the AQI category corresponding to pollutant concentrations.

Table 2. **Pollutant Breakpoint values**

PM10 (µg/m^3)	PM2.5 (µg/m^3)	AQI	Category
0–54	0–15.4	0–50	Good
55–154	15.5–40.4	51–100	Moderate
155–254	40.5–65.4	101–150	Unhealthy for sensitive groups
255–354	65.5–150.4	151–200	Unhealthy
355–424	150.5–250.4	201–300	Very unhealthy
425–504	250.5–350.4	301–400	Hazardous
505–604	350.5–500.4	401–500	

These pollutant breakpoint values are used to convert pollutant concentration data into standardized AQI values.

D. Machine Learning Prediction

To predict future AQI levels, three machine learning models, Linear Regression, Random Forest,and Time Series Forecasting is implemented.

The dataset is divided into **training and testing sets**. The models learn relationships between pollutant concentrations, meteorological parameters, and AQI values to generate predictions **[1], [3], [15]**.

Random Forest is particularly effective because it combines multiple decision trees to improve prediction accuracy and reduce overfitting **[17]**.

Here, mainly the time series forecast model is used, and it is trained in five folds, where each testing data will be 15%, and the trained data is 15% and is added after each testing data.

The trained datas are stored as .plk for further processing in the system.

E. Visualization Framework

The predicted AQI values are displayed through an interactive visualization dashboard. The dashboard provides live AQI values with major pollutants, which cause pollution, graphical charts, AQI indicators, and geographic maps, which are catergorized by CPCB standards that help users interpret air quality conditions across different regions.

IV. RESULTS AND DISCUSSION

This section presents the experimental results of the proposed SmartAQI system for air quality monitoring and prediction. The system was implemented using historical air pollution datasets, district geographic coordinates of India, and real-time data obtained through the OpenWeather API. The experiments aim to analyze pollutant concentration levels, compute the Air Quality Index (AQI), and evaluate the performance of machine learning models in predicting air quality conditions.

DOI: 10.48175/IJARSCT-32851

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

A. Dataset Description

The proposed SmartAQI system was conducted using multiple datasets that include historical air pollution data and geographic information of Indian districts. These datasets provide the necessary input features for AQI computation and machine learning–based prediction.

The dataset consists of historical air quality records containing pollutant concentration values for major air pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 and provides temporal information about pollutant levels. The pollutant concentration values are used to predict AQI value compute the Air Quality Index (AQI) based on the guidelines defined by the **Central Pollution Control Board**.

In addition to the pollution dataset, a geographical dataset containing the latitude and longitude coordinates of districts across India was used. This dataset enables spatial visualization of AQI values and supports the generation of geographic air quality maps within the SmartAQI system.

Real-time meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure are retrieved using the **OpenWeather API**.

By integrating historical pollution data, geographic coordinates, and meteorological information, the SmartAQI system creates a comprehensive dataset that supports accurate air quality monitoring, prediction, and visualisation across different regions of India.

B. AQI Calculation Results

The Air Quality Index (AQI) is calculated to convert pollutant concentration values into a single numerical indicator that represents the overall air quality level. In this project, AQI values are calculated based on the guidelines defined by the **Central Pollution Control Board**.

The calculation process begins by determining the **sub-index value** for each pollutant using its concentration level and the corresponding breakpoint values defined in the CPCB standards. The AQI sub-index for each pollutant is computed using the following formula:

$$AQI = \frac{(I_{HI} - I_{LO})}{(BP_{HI} - BP_{LO})} \times (C_p - BP_{LO}) + I_{LO}$$

where:

C_p represents the observed concentration of the pollutant

BP_{HI} and BP_{LO} represent the upper and lower breakpoint concentrations

I_{HI} and I_{LO} represent the AQI values corresponding to the breakpoints

For each observation, individual sub-indices are calculated for major pollutants, including **PM2.5, PM10, NO_2, SO_2, CO, and O_3**. After computing the sub-index values, the **maximum sub-index** among all pollutants is considered as the final AQI value for that location.Fig. 2 illustrates the pollutant concentration levels used for AQI computation in the SmartAQI system.

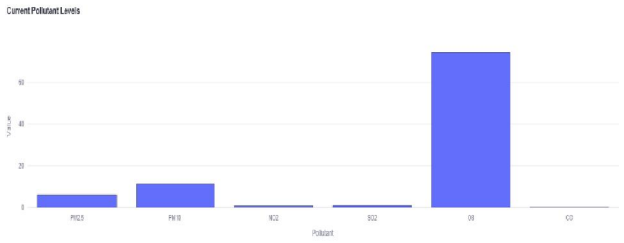

Fig. 2. Pollutant concentration levels of major air pollutants used for AQI calculation in the SmartAQI system.

The AQI values obtained from this calculation are further used as input for the machine learning models in the SmartAQI system to perform air quality prediction and analysis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

C. Machine Learning Prediction Results

To predict future air quality, the SmartAQI system uses machine learning models that analyze historical pollutant concentrations and meteorological parameters. The processed dataset obtained after preprocessing and AQI calculation is used to train the prediction models.

In this study, three machine learning techniques were implemented. These models are trained using historical air pollution data along with weather parameters obtained from the OpenWeather API. The dataset was divided into training and testing sets, where the training data was used to build the models and the testing data was used to evaluate prediction performance.

The Linear Regression model establishes a linear relationship between pollutant concentrations, meteorological variables, and AQI values and provides a simple baseline approach for predicting air quality levels.

The Random Forest model is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and robustness, and is particularly effective for nonlinear relationships between air pollutants and weather conditions.

As a result, Random Forest generally provides more stable and reliable AQI predictions compared to simple regression models.

A time series forecasting model was used to analyze temporal patterns in air pollution data. Since air quality levels often follow time-dependent trends, time series analysis helps in predicting future AQI values based on historical observations.

To evaluate the effectiveness of the proposed SmartAQI prediction system, a comparison between the live AQI value obtained from the OpenWeather API and the predicted AQI generated by the trained machine learning model was performed. The comparison results are illustrated in Fig. 3. This comparison demonstrates the capability of the proposed system to estimate AQI trends and provide early predictions that can assist in environmental monitoring and public awareness.

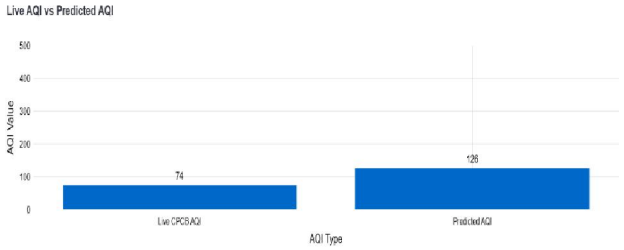

Fig. 3. Comparison between live AQI obtained from the API and AQI predicted.

To analyze short-term air quality trends, a time-series forecasting model was employed to predict AQI values for the next 24 hours. Fig. 4 illustrates the predicted AQI values across the upcoming 24-hour period generated by the SmartAQI system. The results indicate stable AQI levels around the moderate range during the forecasting window. Such forecasts enable early awareness of potential air quality changes.

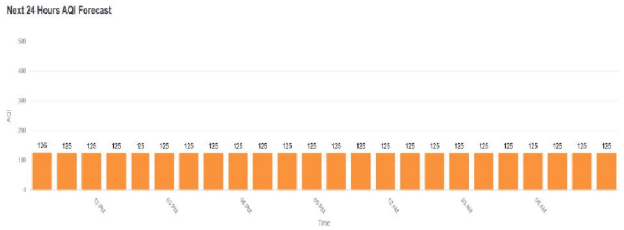

Fig. 4. Predicted AQI values for the next 24 hours generated using the time-series forecasting model.

DOI: 10.48175/IJARSCT-32851

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

The performance of the trained Random Forest model was evaluated to determine its prediction reliability for Air Quality Index (AQI) forecasting. The model achieved a confidence level of **87.3%**, indicating a high level of prediction accuracy in estimating AQI values based on pollutant concentrations and historical air quality data. Fig. 5 illustrates the model confidence visualization used in the SmartAQI dashboard.A higher confidence score indicates stronger agreement between predicted values and historical patterns learned by the model.

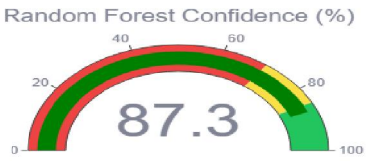

Fig. 5. Random Forest model confidence visualization indicating the reliability of AQI prediction

The experimental results show that the combination of historical pollutant data and meteorological parameters significantly improves prediction capability.

The predicted AQI values are further used in the SmartAQI system to generate visualization dashboards and geographic air quality maps for easier interpretation of air pollution levels.

D. Visualization Dashboard Results

The SmartAQI system includes an interactive visualization dashboard that presents air quality information in a clear and user-friendly format. The dashboard displays pollutant concentration levels, computed AQI values, and predicted air quality conditions using graphical representations and geographic maps, as shown in Fig 6.

The data are processed by the system and displayed through visual components such as charts, indicators, and maps to improve interpretability.

Fig. 6. Live AQI status with Major Air Pollutants.

The visualization interface provides several important featureswith their corresponding air quality categories defined by the **Central Pollution Control Board**, such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

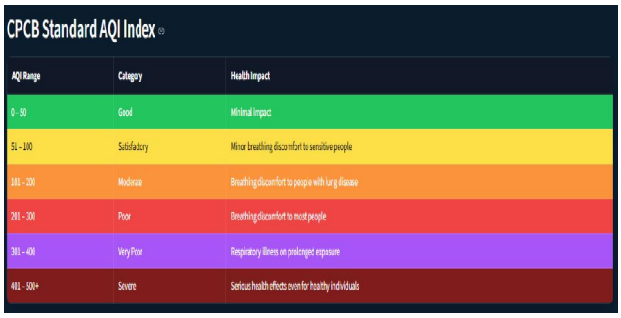

Fig. 7. CPCB Standard AQI Index.

In addition, the system provides geographic visualization of AQI values across different regions of India. The map uses color-coded indicators to represent different AQI levels, enabling users to quickly identify areas with higher pollution levels, as shown in Fig 8. This feature helps in understanding the spatial distribution of air quality conditions.

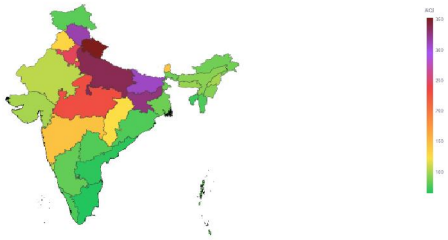

Fig. 8. Geographic visualization of AQI levels across Indian states using the SmartAQI system.

To further analyze the variation of AQI levels across different states, a comparative visualization of state-wise AQI values was generated. This leaderboard highlights the regions with the highest pollution levels based on the calculated AQI values. Such a ranking helps identify pollution hotspots and regions requiring immediate environmental monitoring. The state-wise AQI comparison generated by the SmartAQI system is illustrated in **Fig. 9**.

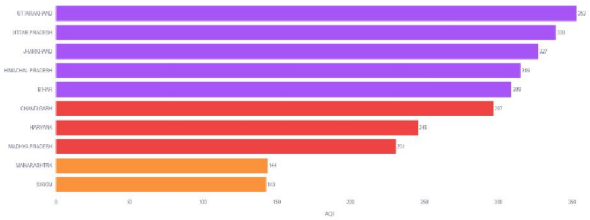

Fig. 9. State-wise AQI leaderboard showing the most polluted regions identified by the SmartAQI system.

The experimental results show that the visualization dashboard effectively presents complex environmental data in an accessible format. By integrating pollutant analysis, AQI computation, and prediction results into a single interface, the SmartAQI dashboard enhances the usability of the system and supports better environmental monitoring.

V. CONCLUSION AND FUTURE WORK

This paper presented SmartAQI, an intelligent air quality monitoring and now casting system to analyze and estimate Air Quality Index (AQI) levels using environmental data and machine learning techniques. The system integrates historical pollution datasets, geographic information of districts in India, and real-time meteorological parameters obtained through the OpenWeather API to provide an analysis of air quality conditions.

The proposed framework calculates AQI values according to the Central Pollution Control Board (CPCB) guidelines, where pollutant concentration levels are converted into standardized AQI categories. These classifications allow users

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, April 2026
Impact Factor: 8.2

to easily interpret air quality conditions and understand the potential health risks associated with different pollution levels.

To estimate AQI values, several machine learning techniques were implemented, including Linear Regression, Random Forest, and Time Series models. In this study, the time series model was primarily used for AQI nowcasting, as it can effectively capture temporal patterns in environmental data and estimate current air quality conditions using recent observations. The comparison between predicted AQI values and real-time AQI measurements indicates that the SmartAQI system can successfully track pollution trends and generate reliable short-term estimations.

The system also includes various visualization components, such as pollutant concentration charts, geographic AQI maps, and interactive dashboards. These visualizations help users monitor pollution levels and identify areas with higher air pollution. Results demonstrate that the SmartAQI framework can support real-time air quality monitoring, nowcasting, and environmental awareness.

Future work may explore the use of advanced deep learning techniques, such as Long Short-Term Memory (LSTM) networks or hybrid models, to further improve prediction accuracy. Additionally, incorporating more environmental variables such as traffic data, industrial emissions, and satellite observations could enhance model performance. Additionally, by using a time series model, the seasonal changes were calculated to predict the AQI values. And adding alerts/ notifications when the AQI values exceed 300, which are harmful, by sending SMS or Emails. Developing a mobile or web-based platform for public access could further improve the usability and practical impact of the SmartAQI system.

Overall, the SmartAQI framework demonstrates the potential of combining machine learning, real-time environmental data integration, and visualization technologies to support effective air quality monitoring and nowcasting.

REFERENCES

[1] R. Yu, Y. Yang, L. Yang, G. Han, and O. A. Move, “RAQ—A random forest approach for predicting air quality in urban sensing systems,” *Sensors*, vol. 16, no. 1, pp. 1–17, 2016.

[2] C. Zhang and D. Yuan, “Fast fine-grained air quality index level prediction using random forest algorithm,” in *Proc. IEEE Int. Conf. Ubiquitous Intelligence and Computing*, 2015, pp. 929–934.

[3] A. Verma and L. Bhatia, “Enhancing accuracy in urban air quality prediction: A comparative study of predictive algorithms,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, pp. 45–55, 2023.

[4] H. Wu, T. Yang, H. Li, and Z. Zhou, “Air quality prediction model based on machine learning techniques,” Scientific Reports, vol. 13, 2023.

[5] M. Rajesh, R. Ganesh Babu, and U. Moorthy, “Machine learning-driven framework for real-time air quality assessment,” Scientific Reports, vol. 15, 2025.

[6] S. Singh, A. Yadav, and A. Kumar, “Prediction of air pollution using random forest,” EasyChair Preprint, 2021.

[7] X. Yang, “Air quality forecasting using random forest regression,” Journal of Innovation and Development, vol. 5, no. 2, pp. 112–120, 2024.

[8] M. N. Afif Al Arsy and A. M. Yasir, “PM2.5 concentration prediction model using random forest algorithm,” Journal of Computation Physics and Earth Science, vol. 3, no. 1, 2024.

[9] X. Su et al., “Research on urban air quality prediction system based on improved random forest modeling,” Ecological Chemistry and Engineering, vol. 32, 2025.

[10] A. Kumar and P. Goyal, “Forecasting of air quality index using artificial neural networks,” Atmospheric Environment, vol. 44, no. 8, pp. 101–110, 2010.

[11] J. Chen, Y. Li, and X. Zhang, “Air quality prediction using machine learning methods,” Environmental Science and Pollution Research, vol. 26, pp. 1–12, 2019.

[12] T. Grange, C. Carslaw, and D. Lewis, “Random Forest meteorological normalisation models for air quality data,” Atmospheric Environment, vol. 200, pp. 1–9, 2018.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, April 2026 Impact Factor: 8.2

[13] Y. Zheng, F. Liu, and H. Hsieh, "U-Air: When urban air quality inference meets big data," in Proc. ACM SIGKDD, 2013, pp. 1436–1444.
[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[15] G. Box and G. Jenkins, Time Series Analysis: Forecasting and Control. Hoboken, NJ, USA: Wiley, 2015.
[16] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY, USA: Springer, 2009.
[17] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[18] J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[19] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Burlington, MA, USA: Morgan Kaufmann, 2012.
[20] World Health Organization, "Ambient air pollution: A global assessment of exposure and burden of disease," WHO Report, 2016.
[21] Central Pollution Control Board (CPCB), "National air quality index," Government of India Report, 2022.