ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

AN INTEGRATED MACHINE LEARNING FRAMEWORK FOR AIR QUALITY MONITORING AND ASSESSMENT

*** Suryaprakash Upadhyay**

** Assistant Professor, Department of Information Technology, Ramanand Arya D.A.V College (Autonomous)*

Abstract:

Air pollution poses significant risks to human health and the environment, particularly in rapidly urbanizing metropolitan regions. Accurate monitoring and prediction of air quality parameters are essential for effective policy formulation and public health interventions. This study proposes an integrated machine learning framework for monitoring and assessing air quality in the Mulund region of Mumbai using real-world sensor data collected from February 2025 till 23 January 2026. The dataset comprises pollutant concentrations including PM2.5, PM10, and O_3 obtained from the Maharashtra Pollution Control Board through the OpenAQ platform. After data preprocessing and temporal feature engineering, predictive models based on Linear Regression and Random Forest algorithms were developed to estimate PM2.5 concentrations. Experimental results demonstrate that the Random Forest model achieves superior performance with an R^2 value of 0.85 and a mean absolute error of 2.07 µg/m³, significantly outperforming Linear Regression. The proposed framework effectively captures nonlinear relationships and temporal patterns in air quality data, offering a reliable and scalable approach for real-time air quality assessment and decision-making in urban environments.

Keywords: *Air Quality Monitoring, Machine Learning, PM2.5 Prediction, Random Forest, OpenAQ, Temporal Modeling, Urban Air Pollution*

Introduction:

Air pollution has emerged as one of the most critical environmental challenges in urban regions worldwide, particularly in rapidly developing countries such as India. The increasing concentration of atmospheric pollutants has been strongly associated with adverse health outcomes, including respiratory illnesses, cardiovascular diseases, and premature mortality. Among various air pollutants, fine particulate matter with an aerodynamic diameter less than 2.5 micrometers (PM2.5) is considered one of the most harmful due to its ability to penetrate deep into the respiratory system and bloodstream. Consequently, continuous monitoring and accurate prediction of air quality parameters have become essential for effective public health management, environmental regulation, and urban planning.

Urban environments such as Mumbai experience complex pollution dynamics influenced by traffic density, industrial emissions, meteorological conditions, and seasonal variations. Traditional air quality monitoring systems rely primarily on fixed monitoring stations and statistical forecasting techniques. While these approaches provide valuable insights, they often fail to adequately capture the nonlinear and time-dependent relationships among

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

multiple pollutants and environmental variables. As a result, the predictive performance of conventional methods remains limited, particularly under rapidly changing urban conditions.

Recent advances in machine learning have introduced powerful data-driven techniques capable of modeling complex nonlinear patterns and temporal dependencies in environmental data. Machine learning algorithms such as Linear Regression, Random Forest, and ensemble methods have demonstrated significant potential in improving the accuracy of air quality prediction and assessment. These models can effectively integrate multiple pollutant parameters and temporal features, enabling more reliable forecasting and real-time decision support.

In this study, an integrated machine learning framework is proposed for air quality monitoring and assessment in the Mulund region of Mumbai using real-world sensor data obtained from the OpenAQ platform. The framework focuses on predicting PM2.5 concentrations by leveraging pollutant measurements including PM10 and O_3 along with temporal features. By comparing the performance of Linear Regression and Random Forest models, the proposed approach aims to demonstrate the effectiveness of ensemble learning techniques in capturing nonlinear relationships and temporal patterns in urban air quality data. The outcomes of this work contribute to the development of reliable and scalable systems for real-time air quality assessment and environmental management in metropolitan regions.

Dataset Description:

The dataset was collected from the OpenAQ platform for the Mulund West monitoring station in Mumbai, operated by the Maharashtra Pollution Control Board. Monthly datasets from February 2025 to 23 January 2026 were merged to form a comprehensive annual dataset. The parameters considered include PM2.5, PM10, and O_3 concentrations, along with temporal information. After preprocessing, missing values were removed and the dataset was transformed into a structured time-series format suitable for machine learning analysis.

Literature Review:

Air quality monitoring and prediction have emerged as critical research domains due to the increasing health and environmental impacts of atmospheric pollution, particularly fine particulate matter (PM2.5). Conventional statistical approaches, while widely used, often struggle to model the complex nonlinear interactions between meteorological variables and pollutant concentrations. Consequently, machine learning (ML) and deep learning techniques have gained increasing attention for accurate air quality assessment and forecasting.

Nguyen et al. (2024) investigated the application of statistical and machine learning methods for estimating PM2.5 concentrations in Vietnam. Their comparative analysis demonstrated that ensemble learning models significantly outperformed traditional regression techniques, highlighting the importance of incorporating multiple environmental variables for improved predictive performance.

Singh et al. (2025) conducted a comprehensive spatiotemporal analysis of air quality in Indian urban regions using machine learning-based prediction models. Their study emphasized the relevance of city-scale monitoring and revealed that ensemble models are particularly effective in capturing seasonal variations and spatial heterogeneity of air pollutants. This work is highly relevant for Indian metropolitan cities, where pollution dynamics are influenced by traffic congestion, industrial emissions, and complex meteorological conditions.

In a study published in *Scientific Reports* (2025), researchers proposed a virtual monitoring station framework based on machine learning for PM2.5 concentration prediction. The proposed model

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

demonstrated strong generalization capability in regions with sparse physical monitoring infrastructure, suggesting that data-driven approaches can effectively complement traditional air quality monitoring networks.

A comparative investigation reported in *Water, Air, & Soil Pollution* (2025) evaluated multiple machine learning algorithms and ensemble strategies for air quality forecasting. The authors found that ensemble learning methods consistently achieved higher prediction accuracy and robustness than individual models, underscoring the advantages of combining multiple learners for reliable air quality assessment.

Earlier foundational work by *Jamal and Nabizadeh Nodehi (2017)* examined various machine learning techniques for forecasting PM2.5 mass concentrations. Their results confirmed the superiority of nonlinear models, including neural networks and tree-based algorithms, over linear regression methods, thereby establishing an important methodological basis for subsequent studies in this field.

Further advancements were reported in *Aerosol and Air Quality Research* (2020), where the XGBoost algorithm was applied for PM2.5 prediction in Shanghai. The study demonstrated that gradient boosting methods effectively capture nonlinear interactions among predictors and significantly reduce forecasting errors.

More recent studies published in *Earth Science Informatics* (2025) and *Scientific Reports* (2024) explored deep learning and hybrid machine learning frameworks for urban air quality prediction. These investigations revealed that integrating meteorological variables with temporal features substantially enhances prediction accuracy, particularly in highly dynamic urban environments.

Overall, the existing literature clearly establishes the effectiveness of machine learning and ensemble techniques for PM2.5 prediction and air quality assessment. However, limited research has focused on city-level real-time prediction using publicly available datasets such as OpenAQ for Indian metropolitan regions. Moreover, few studies have systematically evaluated ensemble-based frameworks using long-term real-world sensor data. Therefore, the present study aims to develop and evaluate an integrated machine learning framework for PM2.5 concentration prediction using OpenAQ data from Mumbai, contributing to improved urban air quality monitoring and decision support systems.

Research Gaps:

Although existing studies demonstrate the effectiveness of machine learning and integrated frameworks for air quality prediction, several important research gaps remain and motivate the present work:

1. **Localized Monitoring Frameworks:** Most existing air quality prediction frameworks are developed using national- or regional-scale datasets. Limited attention has been given to highly localized suburban monitoring stations, where pollution dynamics may differ significantly from city-wide averages. This creates a need for station-level predictive frameworks tailored to specific urban subregions.
2. **Integration of Temporal Features:** While many studies utilize pollutant concentrations as primary predictors, relatively few have systematically integrated temporal features such as hour, day, and month to capture diurnal and seasonal variations in localized urban environments. Incorporating temporal modeling remains an underexplored area in suburban air quality assessment.
3. **Indian Urban Context Beyond Major Metropolises:** In the Indian context, most machine learning-based air quality studies focus predominantly on highly

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

polluted cities such as Delhi. There is limited research addressing suburban regions of other major metropolitan areas, including Mumbai, using long-term real-world sensor data with comprehensive temporal coverage.

4. **Comparative Evaluation of Linear and Ensemble Models:**
   Although ensemble methods have shown promising performance, systematic comparisons between traditional linear models and advanced ensemble techniques on long-term datasets from individual monitoring stations remain scarce. A detailed comparative evaluation is necessary to assess model robustness and generalization capability under real-world conditions.

Addressing these gaps, the present study develops an integrated machine learning framework for localized air quality monitoring in the Mulund region of Mumbai, incorporating temporal feature engineering and a comparative evaluation of linear and ensemble learning models using long-term OpenAQ sensor data.

Methodology:

This section describes the dataset, preprocessing steps, feature engineering process, model development, and evaluation metrics employed in the proposed integrated machine learning framework for air quality monitoring and assessment.

1. Data Preprocessing

The raw air quality data were obtained from the OpenAQ platform, sourced from the Maharashtra Pollution Control Board monitoring station located in the Mulund region of Mumbai. The dataset initially contained multiple pollutant measurements recorded at irregular time intervals. To ensure data consistency and relevance, only key pollutants relevant to urban air quality assessment—PM2.5, PM10, and O_3—were retained for further analysis.

A pivot transformation was applied to restructure the dataset into a multivariate time-series format, where each timestamp corresponds to a single record containing simultaneous measurements of the selected pollutants. Missing values were handled by removing incomplete records to maintain data integrity. Timestamp information was converted into a standardized datetime format, and the data were sorted chronologically to preserve temporal continuity.

To capture temporal variability and seasonal effects, temporal features were extracted from the timestamp, including hour of day, day of month, and month of year. These features enable the model to learn diurnal cycles, short-term fluctuations, and seasonal trends in pollutant concentrations.

2. Feature Selection

Based on prior literature and exploratory data analysis, a set of predictor variables was selected to estimate PM2.5 concentrations. The input feature set consists of:

- PM10 concentration
- O_3 concentration
- Hour of day
- Day of month
- Month of year

The target variable for prediction is PM2.5 concentration. This selection allows the model to capture both pollutant interdependencies and temporal variations, thereby enhancing predictive accuracy.

3. Model Development

Two regression-based machine learning models were implemented and evaluated in this study:

- **Linear Regression**, serving as a baseline model to assess linear relationships between predictors and the target variable.
- **Random Forest Regression**, an ensemble learning technique capable of modeling nonlinear relationships and complex interactions among input features.

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

The complete dataset was randomly divided into training and testing subsets, with 70% of the data used for model training and 30% reserved for performance evaluation. Model hyperparameters for the Random Forest algorithm were selected empirically to balance model complexity and generalization performance.

4. Performance Evaluation

Model performance was assessed using standard regression evaluation metrics, including:

- **Mean Absolute Error (MAE)**, which measures the average magnitude of prediction errors.
- **Root Mean Square Error (RMSE)**, which penalizes larger errors and reflects overall prediction accuracy.
- **Coefficient of Determination (R^2)**, which indicates the proportion of variance in PM2.5 concentration explained by the model.

These metrics provide a comprehensive assessment of model accuracy, robustness, and explanatory power.

Results and Discussion:

The performance comparison of the developed models is presented in Table 1.

Table 1: Model Performance Comparison

Model	MAE (µg/m³)	RMSE (µg/m³)	R^2
Linear Regression	5.14	6.95	0.62
Random Forest	2.07	4.33	0.85

The Random Forest model significantly outperformed Linear Regression across all evaluation metrics. The high R^2 value of 0.85 indicates strong predictive capability and effective modeling of non-linear relationships among pollutants and temporal variables. The results demonstrate the importance of incorporating temporal features and ensemble learning methods for air quality prediction.

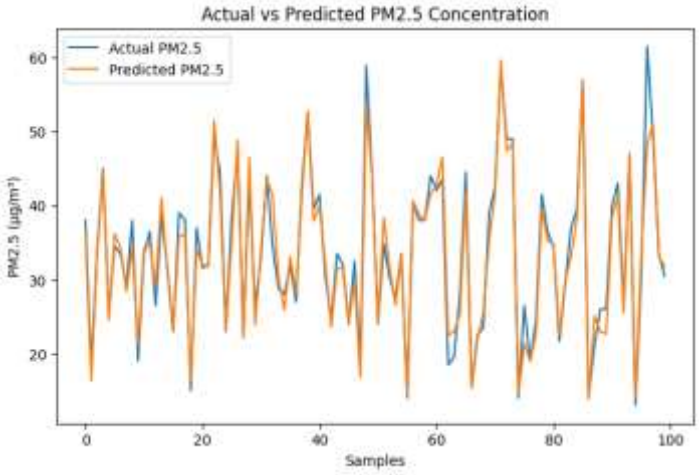

Figure X

Figure X illustrates the comparison between actual and predicted PM2.5 concentrations for the test dataset. The predicted values closely follow the actual measurements, indicating that the Random Forest model effectively captures the temporal and pollutant-driven variations in PM2.5 levels.

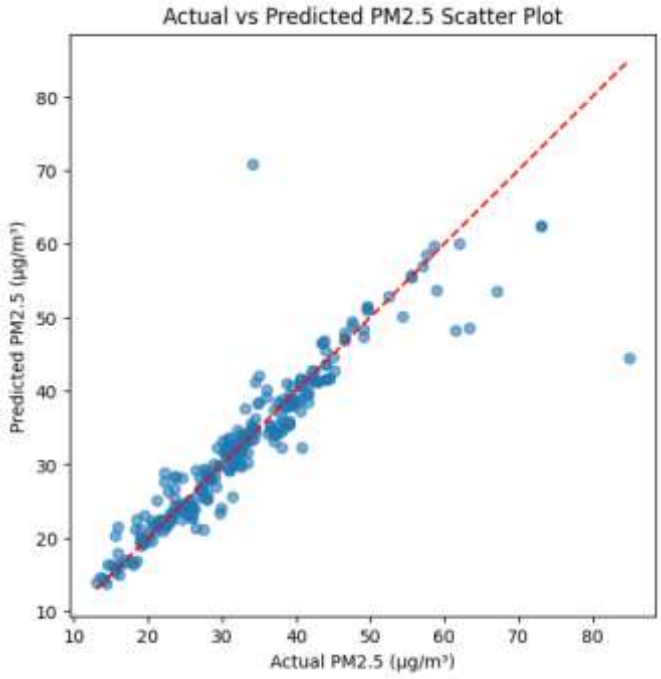

Figure Y

Figure Y presents the scatter plot between actual and predicted PM2.5 concentrations using the Random Forest model. Most data points are closely distributed

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(a) Jan – Feb, 2026

OPEN ACCESS Original Research Article

along the reference diagonal line, indicating strong agreement between predicted and observed values. The limited dispersion confirms the high predictive accuracy and robustness of the proposed framework.

Conclusion:

This study presented an integrated machine learning framework for air quality monitoring and assessment using long-term real-world sensor data collected from the Mulund region of Mumbai through the OpenAQ platform. By incorporating pollutant measurements and temporal features, the proposed framework effectively captured both nonlinear relationships and time-dependent patterns in urban air pollution.

Experimental results demonstrated that the Random Forest model significantly outperformed Linear Regression, achieving a high coefficient of determination ($R^2 = 0.85$) and low prediction errors. These findings confirm the suitability of ensemble learning techniques for modeling complex air pollution dynamics and highlight the importance of temporal feature integration for accurate PM2.5 prediction.

The proposed framework provides a reliable and scalable approach for localized air quality assessment and can support data-driven decision-making for environmental management and public health protection. The results indicate strong potential for deploying such models in real-time air quality monitoring systems and smart city applications.

Future Work:

Future work may incorporate additional meteorological parameters such as temperature, humidity, and wind speed, as well as advanced deep learning models such as LSTM networks for long-term air quality forecasting.

References:

1. *Singh, S. K., R. Jain, D. Palaniappan, K. Parmar, T. Premavathi, & J. Gothania, Spatiotemporal analysis and machine learning-based prediction of air quality in Indian urban cities, Environmental Research and Technology*
2. *Nguyen, T. N. T., T. D. Trinh, P. C. L. T. Vu, & P. T. Bao, Statistical and machine learning approaches for estimating pollution of fine particulate matter (PM2.5) in Vietnam, Journal of Environmental Engineering and Landscape Management*
3. *PM2.5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations, Scientific Reports, 2025.*
4. *Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction, Water, Air, & Soil Pollution.*
5. *Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai, Aerosol and Air Quality Research, 20:128–138, 2020.*
6. *Predicting air quality index based on meteorological data: a comparison of regression analysis, artificial neural networks and decision tree ,Akram Jamal, Ramin Nabizadeh Nodehi.*

Cite This Article:

Upadhyay S. (2026). *An Integrated Machine Learning Framework for Air Quality Monitoring and Assessment*. **In Aarhat Multidisciplinary International Education Research Journal**: Vol. XV (Number I, pp. 114–119). **Doi: https://doi.org/10.5281/zenodo.18609164**