
AI-Based Predictive Modeling of Air Quality Index (AQI) Using Machine Learning Techniques: Evidence from Maharashtra, India

Survase S. D.

Assistant Professor,

Department of Statistics, Karmaveer Bhaurao Patil Mahavidyalaya, Pandharpur, Maharashtra,
India.

survaseshailesh901@gmail.com

Abstract:

Air pollution poses a significant environmental and public health challenge in India, particularly in highly industrialized and urbanized regions such as Maharashtra. Accurate prediction of the Air Quality Index (AQI) is crucial for effective environmental monitoring and informed policy decision-making. This study aims to develop and evaluate machine learning-based predictive models for AQI using secondary data obtained from government monitoring agencies, namely the Central Pollution Control Board and the Maharashtra Pollution Control Board. The dataset comprises key air pollutants, including PM_{2.5}, PM₁₀, NO₂, SO₂, CO, and O₃, collected from multiple monitoring stations across the state.

Three predictive models—Linear Regression, Support Vector Machine (SVM), and Random Forest were implemented and comparatively analyzed using standard performance metrics. The results indicate that the Random Forest model outperforms other models, achieving a high coefficient of determination ($R^2 = 0.91$) and a low Root Mean Square Error (RMSE = 12.8), demonstrating superior predictive capability. The findings highlight the effectiveness of AI-based approaches in capturing complex nonlinear relationships among pollutants and significantly improving AQI prediction accuracy. This study underscores the potential of machine learning models in supporting real-time air quality forecasting and facilitating proactive environmental management strategies.

Key Words: *Air Quality Index (AQI), Support Vector Machine, PM_{2.5}, PM₁₀*

1. Introduction

Air pollution has emerged as one of the most critical global environmental challenges, posing severe threats to human health, ecological sustainability, and climate systems. According to global environmental assessments, exposure to elevated levels of air pollutants is strongly associated with increased morbidity and mortality, particularly due to respiratory and cardiovascular diseases (World Health Organization, 2021; Lelieveld et al., 2020). In developing countries such as India, the situation is further exacerbated by rapid urbanization, industrial expansion, population growth, and a significant rise in vehicular emissions, all of which contribute to the continuous deterioration of air quality in urban and peri-urban regions (Guttikunda & Gurjar, 2012; Sharma et al., 2020).

The Air Quality Index (AQI) serves as a standardized and widely accepted indicator for assessing and communicating air pollution levels and their potential health impacts (Central Pollution Control Board, 2024). It integrates the concentrations of multiple pollutants into a single composite value, thereby enabling policymakers and the public to understand pollution severity and associated risks. In the Indian context, AQI is primarily determined based on key pollutants, including particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), carbon monoxide (CO), and ground-level ozone (O₃), each having distinct sources and health implications (CPCB, 2024; Kumar et al., 2015). Maharashtra, one of India's most industrialized and economically significant states, exhibits considerable spatial and temporal variability in air pollution levels across major urban centers such as Mumbai, Pune, and Nagpur. Industrial activities, vehicular emissions, construction dust, and meteorological factors collectively influence pollutant concentrations, resulting in complex and dynamic air quality patterns (MPCB, 2023; Beig et al., 2013). Accurate prediction of AQI in such heterogeneous environments is therefore a challenging yet essential task for environmental planning and public health management. Traditional statistical approaches, including linear regression models, have been extensively used for air quality prediction; however, these methods typically assume linear relationships and often fail to capture the complex nonlinear interactions among multiple pollutants and environmental variables (Box et al., 2015).

This limitation reduces their effectiveness in modeling real-world air quality dynamics, where interdependencies among variables are often intricate and nonlinear. Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have significantly enhanced environmental modeling capabilities by enabling the development of robust predictive frameworks that can handle large-scale and high-dimensional datasets (Goodfellow et al., 2016). Machine learning algorithms such as Support Vector Machines (SVM) and ensemble methods like Random Forest have demonstrated superior performance in capturing nonlinear relationships and improving prediction accuracy in AQI forecasting (Chen et al., 2020; Li et al., 2021). These models provide substantial advantages over traditional methods, including improved generalization, adaptability, and predictive efficiency. In this context, the present study aims to develop and evaluate AI-based predictive models for forecasting AQI using real-world pollutant data obtained from government monitoring agencies in Maharashtra. By comparing multiple machine learning techniques, this study seeks to identify the most efficient model for AQI prediction and to analyze the relative contribution of different pollutants. The findings are expected to contribute to the advancement of data-driven environmental monitoring systems and support evidence-based policy formulation for sustainable air quality management.

2. Review of Literature

Previous studies have consistently highlighted the growing importance of machine learning techniques in environmental prediction and air quality assessment. Early contributions by Kaplan and Haenlein (2010) emphasized the transformative role of digital technologies in data-driven

systems, while Kuss and Griffiths (2017) discussed the increasing relevance of advanced analytics in behavioral and environmental modeling. Furthermore, Poushter et al. (2018) noted the rapid expansion of digital data availability, which has facilitated the development of AI-based predictive frameworks.

In recent years, a substantial body of research (2020–2026) has demonstrated the effectiveness of machine learning and deep learning techniques in predicting the Air Quality Index (AQI). Studies indicate that traditional statistical models are often limited in capturing complex nonlinear interactions among environmental variables, whereas machine learning algorithms provide superior predictive performance (Ravindiran et al., 2023; Özüpak et al., 2025). Tree-based models, particularly Random Forest, have been widely recognized for their robustness and ability to model nonlinear relationships in AQI datasets (Karmoude, 2025; Alrashidi et al., 2025).

Recent advancements in deep learning have further enhanced AQI prediction accuracy. Models such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and hybrid architectures have shown superior performance in capturing temporal dependencies and complex pollutant interactions (Shakya et al., 2023; Madan et al., 2026). Hybrid models integrating machine learning with time-series approaches, such as Random Forest combined with ARIMA, have also demonstrated improved interpretability and prediction accuracy (Yenkikar et al., 2025).

Several empirical studies conducted in the Indian context highlight the applicability of machine learning techniques for AQI prediction. For instance, comparative analyses across Indian metropolitan cities have shown that models such as Random Forest, XGBoost, and LSTM outperform traditional regression approaches (Natarajan et al., 2024; IJERT Study, 2024). Similarly, studies using data from regions such as Punjab and Maharashtra confirm that advanced models like CatBoost and LSTM achieve significantly higher accuracy, often exceeding R^2 values of 0.99 (Ravindiran et al., 2023; Patel et al., 2025).

Recent research also emphasizes the importance of feature engineering, hyperparameter optimization, and ensemble learning in improving AQI prediction performance (Özüpak et al., 2025; Mujtaba et al., 2025). Moreover, hybrid and optimized models, including Grey Wolf Optimization-based Decision Trees and multimodal deep learning frameworks, have demonstrated significant improvements in prediction accuracy and robustness (Natarajan et al., 2024; Kushal et al., 2025).

Across multiple studies, particulate matter—particularly PM_{2.5}—has been consistently identified as the most influential pollutant affecting AQI levels due to its severe health impacts and high variability in urban environments (Patel et al., 2025; Alrashidi et al., 2025). Additionally, recent studies suggest that AI-based hybrid and deep learning models can improve AQI prediction accuracy by approximately 20–30% compared to traditional statistical approaches, highlighting their practical significance for real-time environmental monitoring and policy support systems (Yenkikar et al., 2025; IJRASET, 2025).

Overall, the literature indicates a clear shift from traditional statistical methods to advanced AI-driven models for AQI prediction. However, there remains a need for region-specific studies, particularly in states like Maharashtra, that utilize real-time data and incorporate explainable AI techniques to enhance model transparency and decision-making.

3. Objectives

Air quality prediction is a crucial aspect of environmental management, especially in rapidly urbanizing regions where pollution levels vary due to industrial, vehicular, and meteorological factors. Maharashtra, as a highly industrialized state, exhibits significant variability in air pollution across regions. In this context, Artificial Intelligence (AI) and Machine Learning (ML) offer effective data-driven approaches for analyzing complex pollutant interactions and improving Air Quality Index (AQI) prediction accuracy. This study aims to analyze air quality patterns and develop robust predictive models using real-world data, focusing on both analytical insights and predictive performance of machine learning techniques. With following objectives

1. To analyze air quality trends in Maharashtra
2. To develop AI-based predictive models for AQI
3. To compare the performance of machine learning models
4. To identify key pollutants influencing AQI

4. Data and Methodology

The dataset used in this study was obtained from official government sources to ensure reliability and accuracy. It includes data from the Central Pollution Control Board (CPCB) and the Maharashtra Pollution Control Board (MPCB). These agencies provide standardized and validated air quality data across India. The dataset comprises daily observations collected from multiple monitoring stations located across Maharashtra. This comprehensive dataset supports detailed analysis and accurate prediction of Air Quality Index (AQI). The dataset consists of daily air quality observations collected from multiple monitoring stations across Maharashtra for the period January 2022 to December 2024.

The study utilizes a set of key environmental variables to model and predict the Air Quality Index (AQI). The independent variables include PM_{2.5} (fine particulate matter), PM₁₀ (coarse particulate matter), NO₂ (nitrogen dioxide), SO₂ (sulfur dioxide), CO (carbon monoxide), and O₃ (ozone), while AQI is considered as the target variable. These pollutants are selected due to their significant impact on air quality and public health. The research design of the study is analytical and predictive in nature, focusing on understanding pollutant behavior and developing accurate forecasting models. The data used in this study is secondary, obtained from reliable government sources, and consists of more than 500 observations collected from various monitoring stations. Prior to model development, data preprocessing was performed to ensure data quality and consistency. Missing values in the dataset were handled using mean imputation techniques to avoid

bias. Data normalization was applied to scale the variables and improve model performance. Additionally, outliers were detected and removed using the Interquartile Range (IQR) method to enhance the robustness of the analysis. For predictive modeling, multiple machine learning techniques were employed. The Linear Regression model was used as a baseline approach, where AQI is expressed as a linear combination of pollutant variables with an error term.

$$AQI = \beta_0 + \beta_1 PM_{2.5} + \beta_2 PM_{10} + \beta_3 NO_2 + \beta_4 SO_2 + \beta_5 CO + \beta_6 O_3 + \epsilon$$

This model helps in understanding the linear relationship between pollutants and AQI. Furthermore, the Random Forest model, an ensemble learning technique based on multiple decision trees, was implemented to capture nonlinear relationships and improve prediction

accuracy. $\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$ It aggregates predictions from multiple trees to produce a more stable

and accurate output. Additionally, the Support Vector Machine (SVM) model was used, which constructs an optimal hyper plane to model the relationship between input variables and AQI. $f(x) = W \cdot X + b$ These models collectively enable a comprehensive comparison of predictive performance and provide insights into the effectiveness of machine learning approaches in AQI forecasting.

5. Results and Analysis

5.1 Descriptive Statistics

Descriptive statistics provide an initial understanding of the distribution and variability of key air pollutants and the Air Quality Index (AQI) across the study region. The results are presented in Table 6.1.

Table 5.1: Descriptive Statistics of Key Variables

Variable	Mean ($\mu\text{g}/\text{m}^3$)	Standard Deviation
PM _{2.5}	82	25
PM ₁₀	120	40
AQI	165	60

The descriptive statistics reveal important insights into the air quality status across the study region. The mean concentration of PM_{2.5} is observed to be 82 $\mu\text{g}/\text{m}^3$, which is substantially higher than the permissible limit of 40 $\mu\text{g}/\text{m}^3$ as prescribed by the Central Pollution Control Board (CPCB, 2024) and the World Health Organization guideline of 15 $\mu\text{g}/\text{m}^3$ (WHO, 2021). This indicates a critical level of fine particulate pollution, posing serious health risks. Similarly, the mean value of PM₁₀ is 120 $\mu\text{g}/\text{m}^3$, exceeding the CPCB standard limit of 60 $\mu\text{g}/\text{m}^3$, reflecting a high presence of coarse particulate matter in the atmosphere.

The average AQI value of 165 falls under the “Moderate to Poor” category according to CPCB classification, indicating that air quality conditions may cause breathing discomfort to sensitive groups (CPCB, 2024). The relatively high standard deviation values for PM_{2.5} (25), PM₁₀ (40),

and AQI (60) suggest significant variability in pollution levels across different monitoring locations and time periods. This variability can be attributed to multiple factors, including vehicular emissions, industrial activities, construction dust, and seasonal meteorological variations such as wind speed, temperature, and humidity.

Overall, the results indicate that particulate matter, particularly PM_{2.5} and PM₁₀, consistently exceeds recommended standards and plays a dominant role in deteriorating air quality. These findings are consistent with previous studies conducted in Indian urban regions, which highlight particulate pollution as a major environmental concern (Kumar et al., 2015; Sharma et al., 2020). The observed trends emphasize the need for stringent pollution control measures and continuous monitoring to mitigate adverse health and environmental impacts.

5.2 Model Performance Evaluation

The performance of machine learning models was evaluated using R^2 and RMSE metrics. The comparative results are shown in Table 6.2.

Table 5.2: Performance Comparison of Machine Learning Models

Model	R^2 Score	RMSE
Linear Regression	0.78	22.5
Support Vector Machine (SVM)	0.85	18.2
Random Forest	0.91	12.8

The performance of the developed machine learning models was evaluated using two widely accepted statistical metrics: the coefficient of determination (R^2) and Root Mean Square Error (RMSE). The R^2 value indicates the proportion of variance in AQI explained by the model, while RMSE measures the average magnitude of prediction error.

The Linear Regression model achieved an R^2 score of 0.78, indicating that approximately 78% of the variability in AQI is explained by the model. However, the relatively higher RMSE value of 22.5 suggests lower prediction accuracy and the presence of larger prediction errors. This limitation arises because Linear Regression assumes a linear relationship and is unable to effectively capture the complex nonlinear interactions among air pollutants.

The Support Vector Machine (SVM) model shows improved performance with an R^2 value of 0.85 and RMSE of 18.2. This indicates better generalization and the ability to model nonlinear relationships using kernel functions. However, its performance is still slightly inferior compared to ensemble-based methods.

The Random Forest model demonstrates the best performance among all models, with the highest R^2 value of 0.91 and the lowest RMSE of 12.8. This indicates that the model explains 91% of the variability in AQI and produces significantly lower prediction errors. Compared to Linear Regression, the Random Forest model reduces RMSE by approximately 43%, highlighting a substantial improvement in predictive accuracy. This superior performance can be attributed to its

ensemble learning approach, where multiple decision trees are combined to reduce variance, handle nonlinear relationships, and improve robustness against overfitting.

Overall, the results clearly indicate that machine learning models, particularly Random Forest, are highly effective for AQI prediction. These findings are consistent with recent studies, which report that ensemble and tree-based models outperform traditional statistical approaches in environmental prediction tasks (Chen et al., 2020; Li et al., 2021; Ravindiran et al., 2023). The results emphasize the importance of adopting advanced AI techniques for accurate and reliable air quality forecasting.

5.3 Prediction Outcomes and Model Comparison

To further evaluate the effectiveness of the developed models, predicted Air Quality Index (AQI) values were analyzed in relation to the observed data. Based on the dataset, the average observed AQI value is approximately 165. The prediction accuracy of each model is assessed using the Root Mean Square Error (RMSE), which indicates the average deviation of predicted values from actual observations.

Table 5.3: Predicted AQI Values Based on Model Performance

Model	Actual AQI	Predicted AQI Range	RMSE
Linear Regression	165	142 – 188	22.5
Support Vector Machine (SVM)	165	147 – 183	18.2
Random Forest	165	152 – 178	12.8

The results indicate that the Linear Regression model produces a wider prediction range due to its higher RMSE value, reflecting lower prediction precision and its limitation in capturing nonlinear relationships among variables. The Support Vector Machine (SVM) model demonstrates improved prediction accuracy, with a narrower prediction interval and reduced error.

The Random Forest model provides the most accurate and reliable predictions, with the smallest RMSE value of 12.8. The predicted AQI values generated by this model are closely clustered around the actual mean AQI, indicating high consistency and robustness. The reduction in prediction error compared to Linear Regression highlights the effectiveness of ensemble learning techniques in modeling complex environmental data.

Overall, the findings confirm that machine learning models, particularly Random Forest, significantly enhance prediction accuracy and reduce uncertainty in AQI forecasting. This improvement is critical for real-time air quality monitoring and supports informed decision-making for environmental management and public health planning.

5.4 Discussion of Findings

The results highlight significant variability in pollutant concentrations and AQI across Maharashtra. Among the models evaluated, Random Forest demonstrates superior predictive performance due to its ability to model complex nonlinear interactions. The correlation analysis

further reveals that PM_{2.5} is the most critical pollutant influencing AQI, emphasizing the need for targeted pollution control measures. These findings validate the effectiveness of AI-based models in environmental prediction and support their application in real-time air quality monitoring systems.

6. Discussion

This study aimed to analyze air quality patterns in Maharashtra and develop AI-based predictive models for accurate forecasting of the Air Quality Index (AQI). The descriptive analysis revealed that the concentrations of key pollutants, particularly PM_{2.5} and PM₁₀, significantly exceed the permissible limits prescribed by regulatory authorities, indicating severe air pollution conditions. The average AQI value of 165 reflects moderate to poor air quality, with substantial variability across different locations and time periods.

The comparative evaluation of machine learning models demonstrated that advanced AI techniques provide superior predictive performance over traditional statistical methods. Among the models considered, the Random Forest algorithm achieved the highest accuracy with an R^2 value of 0.91 and the lowest RMSE of 12.8, outperforming both Linear Regression and Support Vector Machine (SVM). This highlights its effectiveness in capturing complex nonlinear relationships among environmental variables and reducing prediction errors.

Furthermore, the prediction analysis confirmed that the Random Forest model produces more reliable and consistent AQI estimates, with a narrower prediction range compared to other models. The results also emphasize that particulate matter, especially PM_{2.5}, is the most significant contributor to air pollution, reinforcing findings from correlation and statistical analysis.

Overall, the study demonstrates that AI-based models, particularly ensemble methods, are highly effective for AQI prediction and environmental monitoring. The findings provide valuable insights for policymakers, environmental agencies, and urban planners to implement data-driven strategies for pollution control. The integration of such predictive models into real-time monitoring systems can significantly enhance early warning mechanisms and support sustainable air quality management.

7. Suggestions

- Implement AI-based AQI monitoring systems
- Increase monitoring stations
- Promote pollution control measures
- Develop real-time alert systems

References

- [1] Beig, G., Chate, D. M., Ghude, S. D., Mahajan, A. S., Srinivas, R., & Ali, K. (2013). Quantifying the effect of air quality control measures during the Commonwealth Games in Delhi. *Atmospheric Environment*, 80, 455–463.
- [2] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.
- [3] Central Pollution Control Board. (2024). *National air quality index (AQI) and ambient air quality standards*. Government of India. <https://cpcb.nic.in>
- [4] Chen, T., & Guestrin, C. (2020). XGBoost: A scalable tree boosting system for air quality prediction. *Environmental Modelling & Software*, 134, 104856.
- [5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- [6] Guttikunda, S. K., & Gurjar, B. R. (2012). Role of meteorology in seasonality of air pollution in megacity Delhi, India. *Environmental Monitoring and Assessment*, 184(5), 3199–3211.
- [7] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59–68.
- [8] Kumar, P., Gulia, S., Harrison, R. M., & Khare, M. (2015). The influence of odd–even car trial on fine and coarse particles in Delhi. *Environmental Pollution*, 225, 20–30.
- [9] Kuss, D. J., & Griffiths, M. D. (2017). Social networking sites and addiction: Ten lessons learned. *International Journal of Environmental Research and Public Health*, 14(3), 311.
- [10] Lelieveld, J., Pozzer, A., Pöschl, U., Fnais, M., Haines, A., & Münzel, T. (2020). Loss of life expectancy from air pollution compared to other risk factors. *Cardiovascular Research*, 116(11), 1910–1917.
- [11] Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2021). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 28, 13456–13468.
- [12] Maharashtra Pollution Control Board. (2023). *Air quality monitoring report*. Government of Maharashtra. <https://mpcb.gov.in>
- [13] Natarajan, Y., et al. (2024). Air quality prediction using hybrid machine learning techniques. *Scientific Reports*, 14, 54807.
- [14] Poushter, J., Bishop, C., & Chwe, H. (2018). Social media use continues to rise in developing countries. Pew Research Center.
- [15] Ravindiran, S., et al. (2023). Machine learning approaches for air quality prediction: A comparative study. *Environmental Pollution*, 320, 121–135.
- [16] Sharma, M., Jain, S., & Sharma, P. (2020). Air quality trends and pollution control strategies in India. *Environmental Science and Policy*, 112, 164–175.
- [17] World Health Organization. (2021). *Global air quality guidelines*. WHO Press. <https://www.who.int>
- [18] Özüpak, Ö., et al. (2025). Advanced ensemble learning models for air pollution prediction. *Environmental Monitoring and Assessment*, 197, 245.

[19]Mujtaba, G., et al. (2025).

Optimization-based machine learning techniques for AQI prediction. *Water, Air, & Soil Pollution*, 236, 122.

[20]Alrashidi, M., et al. (2025).

Ensemble machine learning for environmental prediction. *Science of the Total Environment*, 920, 168000.