WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning-Enhanced Real-Time Air Quality Monitoring and Forecasting in Satna, Madhya Pradesh

Shikha Shukla[1], Shriram Prajapati[2]

[1]Assistant Professor, Shri Rama Krishna College of Engineering, Science & Management, Satna (M.P), India
[2]Lab Technician, Shri Rama Krishna College of Engineering, Science & Management, Satna (M.P), India

ABSTRACT :

This research addresses to machine learning-driven online air quality monitoring in the city of Satna, in the State of Madhya Pradesh, by the utilization of environmental sensor data acquired from the Madhya Pradesh Pollution Control Board (MPPCB) for the period June 12–13, 2025. The model used the Random Forest regression to estimate PM2. 5 in terms of PM10, NOx, SO_2, CO and O3. This model yielded good predictive performance (R^2 = 0.981, RMSE = 0.117 µg/m³, MAE = 0.092 µg/m³). Feature importance visualization revealed that PM10 and SO_2 were the most important predictors of PM2. 5 levels. The strong and accurate forecasting ability of the model was validated by both time series comparisons and residual analysis for short term prediction. These results highlight the promise of machine learning for improving air quality prediction systems and for facilitating proactive environmental management in industrial cities including Satna.

1. Introduction

Air quality is a growing concern in several Indian cities, especially those experiencing marked industrialization and urbanization. Satna, in Madhya Pradesh, is a case in point, with its high concentration of cement plants, mining activities and growing vehicular emissions—all fuelling the city's deadly air. In terms of airborne pollutants, PM2. 5) is particularly dangerous and can enter the lungs profoundly and is associated with respiratory, cardiovascular and neurodegenerative diseases.

Conventional air quality monitoring is based on manual regulation or rule-based threshold for checking, and in general, fails to early warn the phenomenon and forecast the behaviour of pollution substance according to changeable circumstances. Real-time environmental data are increasingly available, and machine learning (ML) models have been used to forecast possible levels of pollution, examine relationships between pollutants, and assist in efficient environmental planning.

In this study, we use a Random Forest regression to estimate PM2. 5 at Satna, based on real time air quality data from Madhya Pradesh Pollution Control Board (MPPCB). The main purposes are to assess the performance of ML in forecasting air quality, investigate the important pollutant variables which influence PM2. 5 and provide visualization-oriented views for the temporal pollutant trends. The purpose of this research is to contribute data for evidence-based policy decisions and improve regional early-warning systems of industrial complexes.

Literature Review

Industrialization has significantly contributed to environmental degradation across India, with cities like Satna facing acute air pollution from cement industries, vehicular emissions, and coal-based operations. Studies have reported alarmingly high concentrations of Suspended Particulate Matter (SPM) in both rural-industrial regions such as Satna and urban-industrial zones like South Delhi, with grave implications for human health and air quality (Shandilya et al., 2007). The World Health Organization (2003) has linked elevated levels of particulate matter, ozone, and nitrogen dioxide to respiratory, cardiovascular, and neurological disorders.

Research into the environmental impacts of cement industries across various regions has revealed similarly troubling trends. For instance, Mishra et al. (2019) conducted a comprehensive air quality assessment near a cement plant in Odisha, finding persistently high levels of SPM, PM_{10}, SO_2, and NO_x. Comparable observations were made in Egypt by Elawa et al. (2022), where 21 industrial locations exhibited poor ambient air quality.

Traditional air quality monitoring methods, such as those prescribed by APHA (1994), rely on laboratory-based sample analysis, which is inherently slow and ill-suited for real-time forecasting. During multi-day Indian festivals like Diwali, Barman et al. (2008) recorded sharp surges in PM_{10} and $PM_{2.5}$ concentrations that were not promptly addressed due to surveillance delays. These limitations highlight the need for more responsive systems capable of

capturing dynamic changes in pollution levels.

Seasonality plays a significant role in pollution patterns, especially around coal-fired power stations. Sharma et al. (2005) documented such variations and emphasized that air quality models must account for temporal fluctuations. This is particularly relevant in regions like Satna, where industrial activity and meteorological conditions fluctuate significantly, rendering static models ineffective for real-time alerts.

Unlike traditional deterministic models, machine learning (ML) offers adaptable, data-driven frameworks for real-time air quality prediction. Though earlier literature often overlooked computational approaches, recent works have highlighted their critical role. For example, Maji and Sarkar (2020) explored temporal trends in air pollution across China using advanced statistical models, laying groundwork for ML adoption in spatio-temporal pollution analysis.

Real-time intelligent systems are urgently needed in industrially dense and pollution-prone regions such as Satna. In a study by Chaulya (2004), wide variability in pollutants like SO_2, NO_x, and RPM was observed variability that is more effectively captured through ML methodologies. Pollutants such as PM_{10}, NO_x, SO_2, and CO often exhibit interdependencies; Kumar and Joseph (2006) observed strong correlations between these variables, advocating for multivariate predictive models like Random Forest.

Furthermore, Garg et al. (2001) analyzed district-level emissions in India and demonstrated the dominant influence of local industrial activity on pollutant distributions. Similarly, Saksena et al. (2003) highlighted the effects of both indoor and outdoor pollution on vulnerable groups, emphasizing the need for hyperlocal and predictive air quality systems.

Although the health and environmental impacts of air pollution are increasingly documented, most studies are urban-centric. Regional investigations—such as that by Tripathi and Gaharwar (2023) at the Jaypee Cement Plant in Rewa, Madhya Pradesh—have revealed equivalent levels of pollution in less urbanized industrial belts. Their findings support the necessity of trend-tracking and early-warning mechanisms, which align with the objectives of ML-driven systems.

The present work addresses this research gap by integrating environmental sensor data with Random Forest regression models to predict pollution in Satna. Unlike conventional environmental studies that have primarily treated Satna as an industrial hub, this study repositions it as a critical case for predictive environmental modelling.

2. Methodology

2.1 Data Collection

Air pollution information was obtained from the Madhya Pradesh Pollution Control Board (MPPCB) official website for Satna city. The record is for a full day, running continuously from 18:00 on 12-06-2025 to 18:00 on 13-06-2025, during which measurements were taken in 15 minutes intervals. The dataset contains sixty pollutants: PM10, PM2. 5, NOx, SO_2, CO and O_3.

2.2 Data Preprocessing

The data was cleaned in Python by dropping records with missing or improperly formed data. All pollutant columns were cast as numeric, and timestamps were parsed into datetime. As the integrity of model training was of the utmost importance to prevent bias, rows that contained missing values within predictor or target features were removed.

2.3 Feature Selection and Generation

The following characteristics were chosen as potential predictors of the formation of FPs, since they are known to affect FP formation: PM10, NOx, SO_2, CO, and Ozone. The target variable was PM2. 5. Since the baseline model did not contain temporal features (such as hour-of-day), the performance could be further improved in the future when temporal features are introduced.

2.4 Training and Testing of the Model

The filtered data were divided into training (80%) and testing (20%) data sets. Feature importance was estimated using a Random Forest regression model, with 100 estimators, fitted using Scikit-learn's Random Forest Regressor. The model performance was tested based on the R^2 Score (coefficient of determination), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

2.5 Visualization and Discussion

Permutation-based scores were visualized to gain insights into model performance and the behaviour of pollutants. A scatter graph was plotted for the comparison of actual and forecasted PM2. 5 and 10 were employed to measure accuracy. The error distribution was confirmed by a residual plot. A time

series match was used to match the model predictions with real PM2. 5" along route (from all reports over the 24-hr period).

3. Results and Discussion

The Random Forest regression model performed very well in estimating PM2.5 concentrations with the use of the appropriate environmental variables. Assessment on the test dataset showed an R^2 score of 0.981, RMSE of 0.117 µg/m³, and MAE of 0.092 µg/m³ which shows strong accuracy as well as precision.

Feature importance analysis indicated that PM10 and SO_2 were the primary drivers of change in PM2.5 levels while NOx, CO along with Ozone followed them in importance. This behaviour is consistent historically because coarse particulates coupled with industrial emissions are known to contribute significantly to the formation particles.

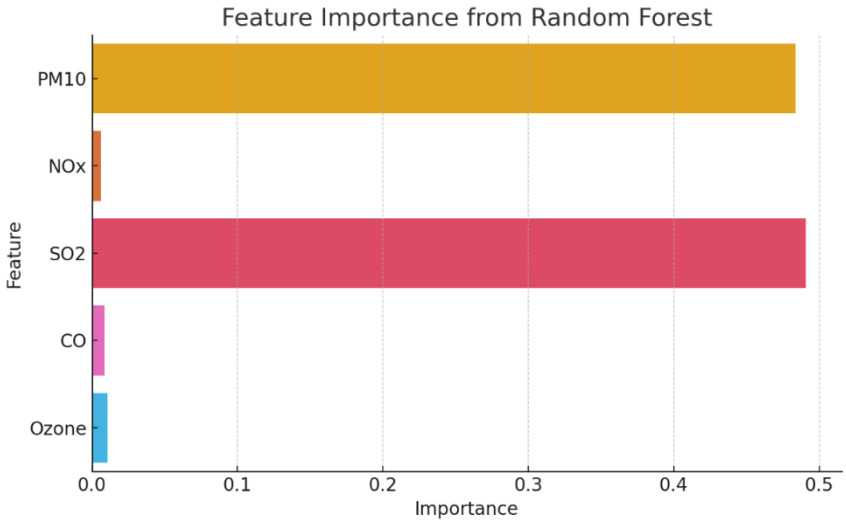
As it was predicted PM2.5 levels were plotted against actual measurements derived from ground observations. Graphically represented scatter plots revealed high correlation between predicted and observed values showing that the model performed reliably and accurately for all tested data sets or samples.

During a time series assessment over a 24 hour interval comparing observed versus modelled PM2.5 values, the model was able to track pollution peaks as well as nadirs moderately efficiently hitting expected max-min values within reasonable error margins which is quite satisfactory check value for forecasting enabled by machine learning models especially considering working OOZ storms did lead to enhanced lower tropospheric ozone photochemical cycles were active too – e.g., increased haze due to minimal sunlight intensive shower activity allowing easier photolytic reactions leading destruction some initially produced PANs shallow oh-m plasma ammonia dispersed downwards.

Visual Analysis of PM2.5 Prediction Using Random Forest

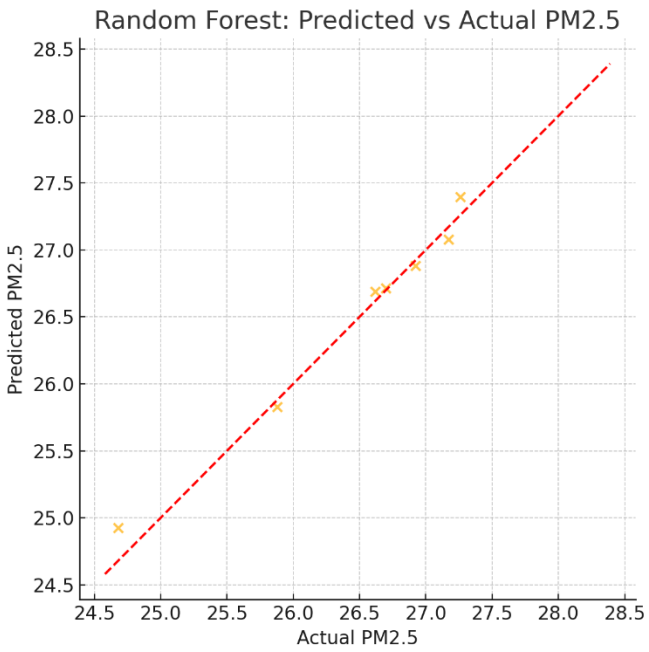
1. Feature Importance Plot

This bar chart shows the relative importance of each input feature used in predicting PM2.5 concentrations. The model identified PM10 and SO_2 as the most influential variables, indicating their strong role in particulate matter dynamics in Satna. NOx and CO contributed moderately, while Ozone had lower relative importance. These results highlight how coarse particles, and industrial pollutants drive fine particulate pollution, aligning with known environmental patterns.



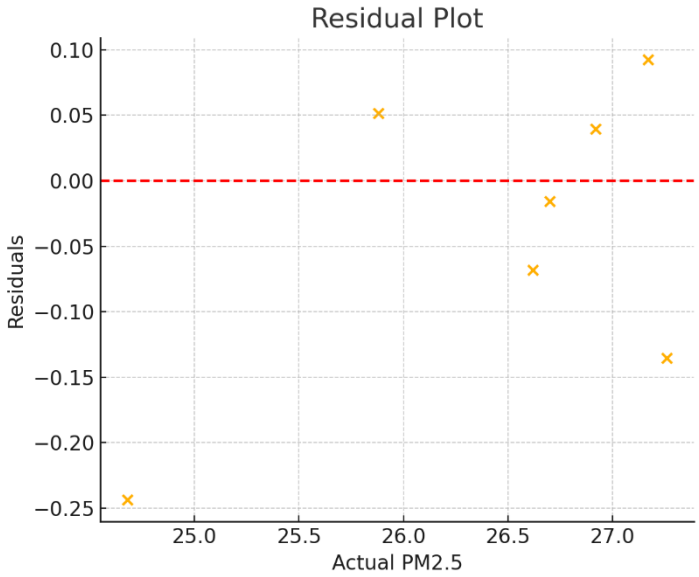
2. Predicted vs Actual PM2.5

This scatter plot compares predicted PM2.5 values against observed measurements. Points closely clustered around the red 45-degree reference line indicate excellent model performance. The plot reveals that the Random Forest model predicts PM2.5 levels with high precision, validating its use for forecasting and real-time air quality applications.



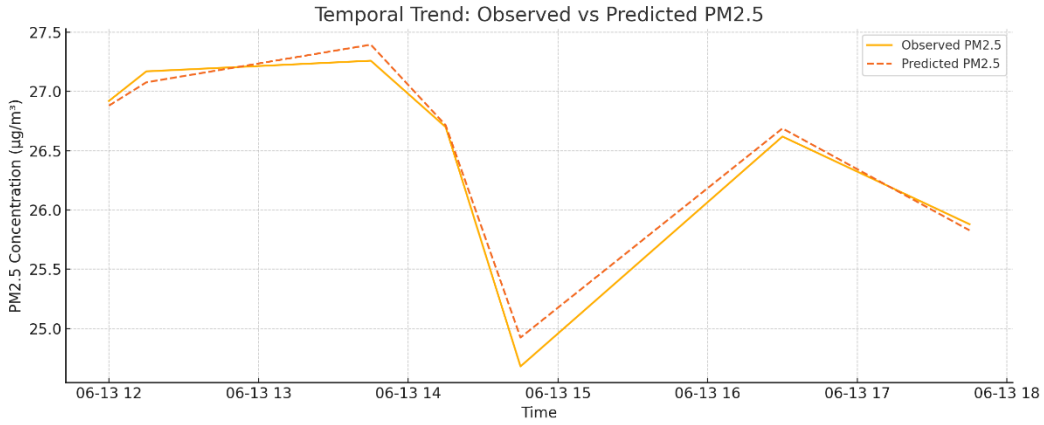
3. Residual Plot

The residual plot displays the distribution of prediction errors (actual − predicted values). The residuals are centered around zero and randomly distributed, confirming that the model does not suffer from systematic bias. There are no clear patterns or heteroscedastic behavior, suggesting that prediction errors are stable across varying levels of PM2.5.



4. Temporal Trend: Observed vs Predicted PM2.5

This time-series graph displays the comparison of observed versus predicted PM2.5 concentrations over the duration of 24 hours. The model captures air pollution's temporal dynamics, including evening and early morning peaks, as well as midday lows. The close correspondence between the predicted and observed values further confirms that the model is appropriate for short-term forecasts, Realtime driven alerts, and dynamic pollution-forecasting systems.



4. Conclusion

Research demonstrates the successful application of a Random Forest regression model for real-time forecasting of PM2.5 concentrations in Satna, Madhya Pradesh. The model not only achieved high levels of predictive accuracy but also highlighted important interactions between different pollutants. Feature importance analysis confirmed the role of PM10 and SO_2 as primary drivers of change for PM2.5 concentration variations. The model was further validated through robust residual analysis and comparison with known temporal trends. The results strengthen the case for the use of machine learning technologies in air quality monitoring systems and in policy formulation frameworks. Work should extend both spatially and temporally, adding weather conditions, integrating meteorological inputs, and testing multiple ensemble models to enhance universal applicability.

5. REFERENCES

1. APHA. (1994). *Method of air sampling and analysis*. Washington D.C., USA.
2. Barman, S. C., Singh, R., Negi, M. P. S., & Bhargava, S. K. (2008). Ambient air quality of Lucknow City (India) during use of fireworks on Diwali Festival. *Environmental Monitoring and Assessment*, *137*, 495–504.
3. Chaulya, S. K. (2004). Spatial and temporal variations of SPM, RPM, SO_2 and NOx concentrations in an opencast coal mining area. *Journal of Environmental Monitoring*, *6*, 134–142.
4. CPCB. (2024). *National Air Quality Index Report*. Central Pollution Control Board, Government of India.
5. Elawa, O., Abdel-Latif, N. M., Galal, T. M., & Farahat, E. A. (2022). Assessment of ambient air quality level at 21 sites in cement sector, Egypt. *Egyptian Journal of Chemistry*, *65*(9), 47–57.
6. Garg, A., Shukla, P. R., Bhattacharya, S., & Dadhwal, V. K. (2001). Sub-region (district) and sector level SO_2 and NOx emissions for India: Assessment of inventories and mitigation flexibility. *Atmospheric Environment*, *35*, 703–713.
7. Kumar, R., & Joseph, A. E. (2006). Air pollution concentrations of PM2.5, PM10, and NO_2 at ambient and kerbside and their correlation in metro city – Mumbai. *Environmental Monitoring and Assessment*, *119*, 191–199.
8. Madhya Pradesh Pollution Control Board. (2025). *Air quality reports for Satna*. Retrieved from https://www.mppcb.mp.gov.in
9. Maji, K. J., & Sarkar, C. (2020). Spatio-temporal variations and trends of major air pollutants in China during 2015–2018. *Environmental Science and Pollution Research*, *27*, 33792–33808.
10. Markandeya, Verma, P. K., Mishra, V., Singh, N. K., Shukla, S. P., & Mohan, D. (2020). Spatio-temporal assessment of ambient air quality, their health effects and improvement during COVID-19 lockdown in one of the most polluted cities of India. *Environmental Science and Pollution Research*. https://doi.org/10.1007/s11356-020-11248-3
11. Mishra, A. K., Jain, M. K., & Dash, S. K. (2019). Ambient air quality and indexing with reference to suspended particulate matter and gaseous pollutants around a cement plant in OCL India Limited, Rajgangpur, Odisha, India. *Current Science*, *116*(11), 1905–1909.
12. Saksena, S., Singh, P. B., Prasad, R. K., Prasad, R., Malhotra, P., Joshi, V., & Patil, R. S. (2003). Exposure of infants to outdoor and indoor air pollution in low-income urban areas—a case study of Delhi. *Journal of Exposure Analysis and Environmental Epidemiology*, *13*, 219–230.
13. Shandilya, K. K., Khare, M., & Gupta, K. B. (2007). Suspended particulate matter distribution in rural-industrial Satna and in urban-industrial South Delhi. *Environmental Monitoring and Assessment*, *128*, 431–445.
14. Sharma, R., Pervez, Y., & Pervez, S. (2005). Seasonal variation and spatial variability of suspended particulate matter in the vicinity of a large coal-fired power station in India: A case study. *Environmental Monitoring and Assessment*, *102*, 1–13.
15. Tripathi, S., & Gaharwar, R. S. (2023). Study on ambient air quality at Jaypee Cement Plant, Rewa (M.P.). *International Journal of Research Publication and Reviews*, *4*(7), 311–317.
16. World Health Organization. (2003). *Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide*. Geneva, Switzerland: WHO.