Enhancing ML Performance on Small Datasets Applied to Forecasting Urban Air Pollution Spikes in Jaipur

Aditya Kumar Jha , Riddhi Sharma
Department of Computer Science and Engineering
Manipal University Jaipur
Jaipur, India

***Abstract*—Urban air pollution, especially $PM_{2.5}$ and PM_{10}, has emerged as a serious concern in rapidly urbanizing Indian cities such as Jaipur. Machine learning algorithms can help predict short term air quality, but they are not a perfect solution. Their usefulness is often limited because there is not enough monitoring data, especially in cities other than Tier-1. This paper presents a small-data-oriented framework that combines domain-knowledge-driven feature engineering with virtual sample generation (VSG) using SMOTE for regression to improve $PM_{2.5}$ and PM_{10} forecasting performance. Using 8-hourly CPCB measurements collected from 2019 to 2025, we first develop a baseline Random Forest regression model and then enhance its performance by integrating domain-specific features such as Diwali and Lohri periods, along with synthetic oversampling of rare extreme-pollution episodes. The enhanced $PM_{2.5}$ model achieves a mean absolute error of 13.09 $\mu g/m^3$ and an R^2 of 0.503 on the held-out 2024–2025 test set, and also demonstrates reliable performance during the Diwali 2025 festival period. The results indicate that virtual sample generation can effectively mitigate small-data limitations and enable useful early-warning air-quality predictions for Indian cities.**

***Index Terms*—Air pollution forecasting, $PM_{2.5}$, PM_{10}, Random Forest, SMOTE for regression, small datasets**

I. Introduction

One of the most challenging environmental issues of the twenty-first century is urban air pollution, and this is further aggravated by the rapidly increasing rate of urbanization and industrialization in developing countries. Jaipur, being one of the most prominent cities in the state of Rajasthan with a population of over three million, experiences a wide range of activities, from industrial production to agriculture, which have a direct impact on air quality. These activities contribute to increased concentrations of fine particulate matter ($PM_{2.5}$ and PM_{10}). During cultural events such as Diwali and winter festivals like Lohri, brief but intense spikes in $PM_{2.5}$ and PM_{10} trigger very poor or severe Air Quality Index (AQI) categories and increase risks of severe respiratory issues [16]–[18].

Air quality prediction models based on data analytics use past air quality and meteorological data to predict future air quality and provide early warnings [10]–[13]. However, most of the existing work is focused on data-rich cities such as Mumbai, Delhi, and Bengaluru, which have dense air quality monitoring stations and rich historical data [10]–[15]. Jaipur, on the other hand, has a remarkably small air quality monitoring dataset, which is mostly limited to a few monitoring stations and a few years of data. The small dataset causes over-optimistic bias in the model, fails to generalize to new years, and performs poorly on small and rare events [2]–[5]. Thus, we point out a particular research gap: developing and testing a lightweight and interpretable ML model for Jaipur that can handle around 7,670 cleaned 8-hourly data points while still being able to model the festival-related peaks.

Air pollution patterns in Jaipur are not random. They show daily and weekly trends, as well as seasonal trends, including spikes during festivals such as Diwali and Lohri. However, we have very limited data for such seasonal trends. This makes Jaipur a suitable example for studying machine learning methods on small environmental datasets.

Therefore, we incorporate techniques such as SMOTE to increase the representation of rare high-pollution events using virtual sample generation. In addition, we include temporal features and domain-knowledge features (festival indicators), hereby creating a framework that can be applied to different cities.

II. Related Work

Research studies show that forecasting in Indian cities, particularly Tier-1 cities such as Delhi and Mumbai, attains higher accuracies given the high densities of weather monitoring infrastructure across these cities [10]–[15]. Models like Random Forests and gradient boosting, as well as neural networks and, more recently, deep learning architectures have been employed to identify non-linear relationships between weather variables, emission indicators, and historical pollutant levels [10]–[15]. Most of the studies are based on extensive monitoring of the networks. Therefore, they are not applicable in Tier-2 cities such as Jaipur due to the lack of data availability.

In recent times, there has been increasing research on the difficulties of using machine learning on small datasets and how it can be prone to overfitting [2]–[5]. Several approaches have been proposed for solving small sample datasets have been presented by data scientists, such as SMOTE [6]. This paper fills the gap in the identified problem by proposing a machine learning approach that is suited to the data limitations of Jaipur. The primary aim is to combine domain knowledge-driven feature engineering with SMOTE and using virtual

sample generation to synthesize the occurrence of less frequent extreme-pollution levels [6]–[9].

III. Materials and Methods

A. Data collection and preprocessing

Air quality data was taken from the Central Pollution Control Board (CPCB) monitoring station at the Police Commissionerate, Jaipur. This data spans from January 2019 to December 2025 and is recorded every 8 hours. The data includes concentrations of $PM_{2.5}$, PM_{10}, NO, NO_2, NH_3, SO_2, CO, and ozone. Raw files from each year were merged and column names were made consistent. Negative pollutant values were removed, which were considered data-entry errors. Missing entries, about 0.4 to 1.0% per column, were filled using forward-fill and linear interpolation to maintain the timeline. After cleaning, there were 7,670 valid observations left.

B. Feature engineering

We developed three types of features:

Temporal Features: From the timestamps, three types of features have been extracted: hour of the day (0, 8, 16), day of the week (0–6), and month (1–12). For smooth daily patterns, the hour feature was encoded using sine and cosine transformations [10]–[13].

Domain-knowledge features: We used binary indicators and continuous variables for major festivals. For Diwali, we created a binary flag for all dates within a window of ± 21 days of the Diwali date each year, and added a continuous variable *days_since_diwali* to calculate the distance from the Diwali date. Similarly, for Lohri, dates within ± 7 days of January 13 each year were considered [16]–[18].

Autoregressive features: Past pollution values were used to predict current pollution levels. We included lag-1 and lag-2 values of $PM_{2.5}$ and PM_{10} as predictors to represent pollution concentrations from the previous and second-previous time steps [10]–[13].

C. Model development

We used a series of Random Forest regression models to train separate models for the prediction of $PM_{2.5}$ and PM_{10} under two configurations [10]–[13]:

A. Baseline model: It is based purely on temporal, meteorological, and lagged pollutant features. The training data includes observations from 2019 to 2023 (about 5,247 samples), while the test set includes data from 2024 to 2025 (around 2,423 samples).

B. Enhanced model: This model is similar to the baseline model but with additional features such as Diwali and Lohri indicators. It applies SMOTE for Regression to oversample high-pollution periods. Specifically, training samples with $PM_{2.5}$ over a certain level (e.g., 150 μg/m^3) are considered minority samples. Synthetic examples are generated by interpolating feature vectors and target values within their vicinity. This method increases the number of extreme events without simply duplicating data [6]–[9].

The Random Forest was selected due to its ability to handle nonlinear relationships, robustness to multicollinearity, and strong performance on small-to-medium datasets. Moreover, unlike deep learning models, Random Forest is less susceptible to overfitting when the training data is small.

D. Model Configuration

For the Random Forest models, we used the scikit-learn library. We used 150 decision trees ($n_{\text{estimators}} = 150$), a maximum tree depth of 10, a fixed random state of 42 to ensure reproducibility. 2019-2023 data were used for training and data from 2024-2025 were reserved for testing. We controlled model complexity by limiting tree depth, which helped reduce overfitting, especially under small data conditions.

E. Evaluation Metrics

Model performance was evaluated using Mean Absolute Error (MAE) and the coefficient of determination (R^2):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2}$$

IV. Experiments and Findings

A. Overall test performance (2024–2025)

On the 2024-2025 test set, the enhanced Random Forest model showed consistent performance for both $PM_{2.5}$ and PM_{10} after adding lag-2 terms and Lohri features [10]–[13]. The final model for $PM_{2.5}$ resulted in an MAE of 13.09 μg/m^3 and an R^2 of 0.503, whereas the model for PM_{10} resulted in an MAE of 30.53 μg/m^3 and an R^2 of 0.460. These performances are slightly better than the earlier baseline Random Forest model without the lag features (MAE for $PM_{2.5}$ is 13.87 μg/m^3 and R^2 is 0.455; MAE for PM_{10} is 30.82 μg/m^3 and R^2 is 0.457), verifying that the addition of lag and festival indicators enhances model performance when a small dataset is present [2]–[5],[10]–[13].

TABLE I
Performance comparison of baseline and enhanced Random Forest models

Model	PM2.5		PM10	
	MAE ($\mu g/m^3$)	R^2	MAE ($\mu g/m^3$)	R^2
Baseline RF	13.87	0.455	30.82	0.457
Enhanced RF	13.09	0.503	30.53	0.460

B. Diwali 2025 performance

In the period around Diwali 2025, $PM_{2.5}$ concentrations exceeded 300 μg/m^3, causing AQI to be around 300 to 400 [16]–[18]. In these severe conditions, the evaluation process yields a 2025 estimate for MAE values for $PM_{2.5}$ of approximately 19.27 μg/m^3 with $R^2 \approx 0.45$, while PM_{10} MAE values for this period are around 40.48 μg/m^3 with $R^2 \approx 0.21$. The model tracks the timing and approximate magnitude of the spike to

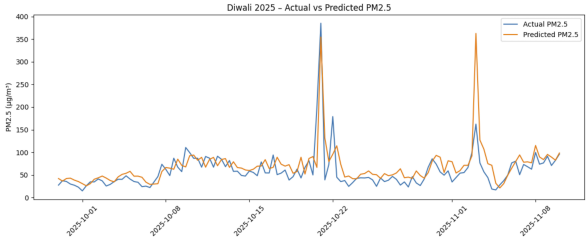

Fig. 1. Observed vs. predicted $PM_{2.5}$ concentrations during the Diwali period (2025).

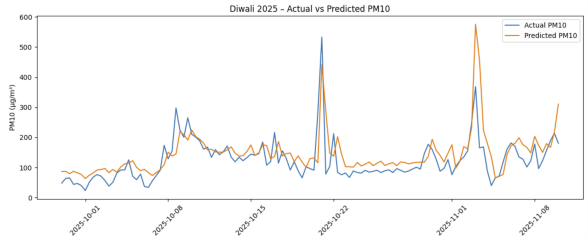

Fig. 2. Observed vs. predicted PM_{10} concentrations during the Diwali period (2025).

distinguish between moderate, poor, and severe AQI levels [10]–[15]. During earlier Diwali seasons from 2019 to 2024, the R^2 for $PM_{2.5}$ in the Diwali period generally falls between about 0.63 and 0.88, with MAE ranging from 13 to 17 $\mu g/m^3$, indicating its effectiveness in capturing this phenomenon [16]–[18].

C. Lohri 2025 performance

While Lohri 2025 had a very low and stable pollution level, winter still makes air quality worse than normal. During this period, the model achieves $PM_{2.5}$ MAE typically in the range 9–16 $\mu g/m^3$ with R^2 between about 0.56 and 0.76 for 2019–2023, and PM_{10} MAE around 20–26 $\mu g/m^3$ with R^2 between about 0.42 and 0.80. For the years 2024 and 2025, performance during the Lohri period remains stable, with a $PM_{2.5}$ MAE of approximately 19.67 $\mu g/m^3$ and an R^2 of approximately 0.32 for 2024. For 2025, the $PM_{2.5}$ MAE is about 17.11 $\mu g/m^3$, with an R^2 of about -0.03. The PM_{10} MAE is around 33.39 and 30.88 $\mu g/m^3$, with R^2 values of approximately 0.24 and 0.14, respectively. Combined with the Diwali performance, these results indicate that the framework works well across both extreme (Diwali) and moderate (Lohri) pollution levels in Jaipur.

D. Error Distribution

Fig. 5 shows the distribution of prediction errors during the Diwali and Lohri period. The residuals are centered around zero , which means that model does not overpredict or underpredict during the period. Larger errors occur primarily during extreme spikes, where pollution is harder to predict due to sudden increase in pollutant levels during festivals.

Performance stays consistent from 2019 to 2023, with R^2 mostly above 0.7. However, it drops in 2024 and 2025. This

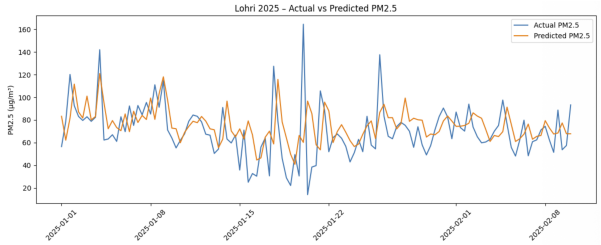

Fig. 3. Observed vs. predicted $PM_{2.5}$ concentrations during the Lohri period (2025).

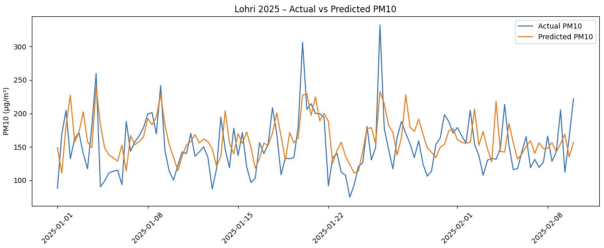

Fig. 4. Observed vs. predicted PM_{10} concentrations during the Lohri period (2025).

shows more changes in distribution during the latest festival seasons.

V. Limitations and Future Work

While the proposed framework has shown good results, it has several limitations that need to be acknowledged. This study relies on data from a single CPCB monitoring station in Jaipur i.e. Police Commissionerate, Jaipur. This may not fully capture the differences across the city. Importantly, the model does not include meteorological forecasting variables such as wind speed, humidity and temperature which may

TABLE II
Year-wise Model Performance During Diwali Period

Year	PM2.5 R^2	PM2.5 MAPE (%)	PM10 R^2	PM10 MAPE (%)
2019	0.852	16.7	0.823	12.6
2020	0.723	17.8	0.762	14.3
2021	0.772	21.8	0.822	16.0
2022	0.855	22.1	0.874	15.6
2023	0.667	17.5	0.729	11.9
2024	-0.068	18.7	0.405	15.9
2025	0.426	38.0	0.291	39.1

TABLE III
Year-wise Model Performance During Lohri Period (4-Week Window)

Year	PM2.5 MAPE (%)	PM10 R^2	PM10 MAPE (%)
2019	18.4	0.425	14.4
2020	18.4	0.510	15.5
2021	19.6	0.682	16.2
2022	28.4	0.805	19.7
2023	20.7	0.697	16.1
2024	24.2	0.244	16.5
2025	32.6	0.145	22.7

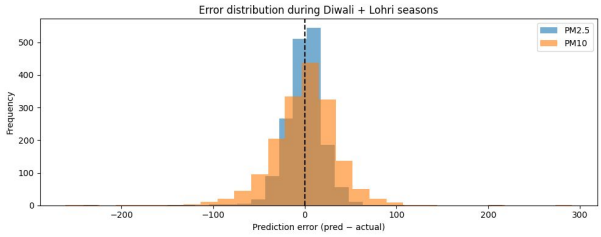

Fig. 5. Distribution of prediction errors for $PM_{2.5}$ and PM_{10} on the 2024–2025 test set.

influence pollutant dispersion patterns. Future research may involve multi-station modeling, adding meteorological forecasting inputs and use this framework across all Indian cities. By incorporating local domain knowledge, such as information about local festivals, industrial patterns, or seasonal crop burning, and using the same strategy. This makes the framework relevant and applicable to Tier-2 and Tier-3 cities where limited monitoring infrastructure is available.

VI. Conclusion

This study presents a framework for forecasting urban air pollution "spikes" in Jaipur using applicable machine learning strategies under realistic small-dataset and imbalanced-data constraints [2]–[7],[20]. The enhanced Random Forest model, by combining two autoregressive features (two lags of $PM_{2.5}$ and PM_{10}) and appropriate festival indicators for Diwali and Lohri, was able to achieve reasonable performance for both $PM_{2.5}$ and PM_{10}, indicating its potential for use during the most critical days by providing early support warnings [6–9,10–15]. The proposed framework can be applied to other Indian cities that face seasonal or festival-related pollution spikes [10]–[15],[17]–[19].

References

[1] World Health Organization, *WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*, Geneva, Switzerland: WHO Press, 2021.

[2] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, Art. no. e0224365, 2019.

[3] S. Steinert, V. Ruf, D. Dzsotjan, N. Großmann, A. Schmidt, J. Kuhn, and S. Küchemann, "A refined approach for evaluating small datasets via binary classification using machine learning," *PLOS ONE*, vol. 19, 2024.

[4] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang, and G.-W. Wei, "Machine learning methods for small data challenges in molecular science," *npj Computational Materials*, vol. 9, 2023.

[5] G.-W. Wei et al., "Small data, big challenges: Machine- and deep-learning strategies for data-limited drug discovery," *Advanced Drug Delivery Reviews*, vol. 229, Feb. 2026, Art. no. 115762.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[7] L. Camacho, G. Douzas, and F. Bacao, "Geometric SMOTE for regression," *Expert Systems with Applications*, vol. 192, Art. no. 116368, 2022.

[8] J. Sivakumar, K. Ramamurthy, M. Radhakrishnan, and D. Won, "GenerativeMTD: A deep synthetic data generation framework for small datasets," *Knowledge-Based Systems*, vol. 280, Art. no. 110956, Sep. 2023.

[9] M. Kannan, D. Umamaheswari, B. Manimekala, I. P. S. Mary, P. M. Savitha, and J. Rozario, "An enhancement of machine learning model performance in disease prediction with synthetic data generation," *Scientific Reports*, vol. 15, Art. no. 33482, 2025.

[10] T. Sreenivasulu and G. M. Rayalu, "Air pollution forecasting using advanced machine learning techniques and ensemble stacking in Delhi," *Environmental Health Engineering and Management Journal*, vol. 12, p. 1370, 2025.

[11] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: A case study of Indian cities," *International Journal of Environmental Science and Technology*, vol. 20, no. 1, 2022.

[12] G. Sathvika, P. Poojitha, K. Rakesh, L. Tadepalli, and K. Chaitanya, "Air quality prediction using Random Forest regression," *Journal of Emerging Technologies and Innovative Research (JETIR)*, 2024.

[13] Y. R. Dubey, P. Pushpender, R. Gupta, S. K. Srivastava, and R. Jafri, "Air quality prediction model using Random Forest," *International Journal of Research Publication and Reviews*, vol. 6, no. 5, 2023.

[14] K. S. Rautela and M. K. Goyal, "Transforming air pollution management in India with artificial intelligence and machine learning technologies," *Scientific Reports*, vol. 14, no. 1, 2024.

[15] S. Lakshmi and A. Krishnamoorthy, "Deep transfer learning and attention-based $PM_{2.5}$ forecasting in Delhi using long-term winter season data," *Scientific Reports*, vol. 15, Art. no. 31787, 2025.

[16] D. Ghei and R. Sane, "Estimates of air pollution in Delhi from the burning of firecrackers during the festival of Diwali," *PLOS ONE*, vol. 13, no. 8, Art. no. e0201311, 2018.

[17] A. Chanchpara, M. Muduli, V. Prabhakar, A. K. Madhava, R. B. Thorat, S. Haldar, and S. Ray, "Pre- to post-Diwali air quality assessment and particulate matter characterization of a western coastal place in India," *Environmental Monitoring and Assessment*, vol. 195, Art. no. 413, 2023.

[18] A. S. Pipal, S. P. Singh, T. Tripathi, and A. Taneja, "Variations in black carbon and particulate matters (PM_1, $PM_{2.5}$, and PM_{10}) during firecrackers bursting episodes and biomass burning: A case study during the Diwali festival," *Journal of Air Pollution and Health*, vol. 7, no. 4, 2022.

[19] P. Sharma *et al.*, "Assessment of air quality index of Jaipur city based on major criteria pollutants," *International Journal of Scientific Research in Applied Sciences*, vol. 11, no. 4, pp. 1–10, 2023.

[20] O. Kothari, N. K. Sah, K. V. S. H. Kumar, P. C. Nair, and N. Sampath, "Forecasting India's air quality: A machine learning approach for comprehensive analysis and prediction," in *Proc. 2024 4th International Conference on Air Quality Forecasting*, 2024.

[21] Y. Li, Y. Yang, P. Song, L. Duan, and R. Ren, "An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space," *Scientific Data*, 2025.