Satellite-Driven Machine Learning Framework for Estimating Surface-Level $PM_{2.5}$ Concentrations

Aljo Joseph[1], Ashik Shaji[2], Renny Thomas[3], Rojins S Martin[4], Jintu Ann John[5]
[1,2,3,4,5]*Department of Information Technology*
[1,2,3,4,5]*Amal Jyothi College of Engineering (Autonomous) Kerala, India*

***Abstract*—Fine particulate matter ($PM_{2.5}$) is one of the most critical air pollutants due to its ability to penetrate deep into the respiratory system, posing severe risks to human health. Reliable estimation of surface-level $PM_{2.5}$ remains challenging, particularly in regions with sparse ground-based monitoring infrastructure. This study proposes a satellite-driven machine learning framework for estimating surface-level $PM_{2.5}$ concentrations by integrating satellite radiance data, meteorological variables, and temporal dependency features. Key atmospheric parameters—including precipitation, boundary layer height, relative humidity, wind speed, and air temperature—are combined with lag-based $PM_{2.5}$ features and principal component representations to enhance predictive robustness. A Random Forest regression model is employed to effectively capture complex non- linear relationships between the input features and $PM_{2.5}$ con- centrations. The proposed model demonstrates strong predictive performance, achieving a coefficient of determination (R^2) of 0.91 and a root mean square error (RMSE) of 0.165 on the test dataset. The results highlight the potential of satellite-assisted machine learning approaches as a cost-effective and scalable solution for air quality assessment, particularly in areas lacking dense monitoring networks.**

***Index Terms*—$PM_{2.5}$ prediction, air pollution modeling, satellite radiance, meteorological features, Random Forest**

I. INTRODUCTION

Air pollution is one of the most critical environmental challenges affecting human health, ecological balance, and climate systems worldwide. Among various air pollutants, fine particulate matter ($PM_{2.5}$)—aerosol particles with an aero- dynamic diameter smaller than 2.5 μm—poses severe health risks due to its ability to penetrate deep into the respiratory system and enter the bloodstream. Prolonged exposure to elevated $PM_{2.5}$ concentrations has been strongly associated with respiratory diseases, cardiovascular disorders, reduced lung function, and increased premature mortality. Consequently, accurate estimation and continuous monitoring of surface-level $PM_{2.5}$ concentrations are essential for effective air quality management, public health risk assessment, and evidence- based environmental policymaking.

Conventional air quality monitoring systems primarily rely on ground-based monitoring stations, which provide high- precision pollutant measurements at specific locations. Despite their accuracy, the deployment and maintenance of these stations involve substantial financial and infrastructural costs, resulting in sparse spatial coverage—particularly in developing regions, rural areas, and complex terrains. This limited coverage restricts the ability to capture spatial variability and temporal dynamics of air pollution across large geographic regions. As a result, ground-based observations alone are often insufficient for comprehensive air quality assessment, high- lighting the need for complementary large-scale monitoring approaches.

Satellite remote sensing has emerged as a powerful alternative for atmospheric and environmental monitoring by providing extensive spatial and temporal coverage. Satellite sensors record radiance values across multiple spectral bands, which contain indirect information related to aerosol loading, atmospheric composition, and meteorological conditions. Al- though satellite observations do not directly measure surface- level $PM_{2.5}$ concentrations, they serve as valuable proxies when combined with

appropriate modeling techniques. Integrating satellite radiance data with meteorological parameters and ground-based observations enables large-scale estimation of PM 2.5 concentrations in regions were direct measurements are sparse or unavailable.

Recent advances in machine learning have significantly improved the capability to model complex, non-linear envi- ronmental processes. Machine learning approaches are partic- ularly effective for air quality estimation due to their ability to learn intricate relationships among heterogeneous input vari- ables, including satellite-derived features, meteorological pa- rameters, and temporal information. Ensemble learning meth- ods, such as Random Forest regression, have shown strong performance in air pollution modeling tasks owing to their robustness against overfitting, tolerance to multicollinearity, and effectiveness in capturing non-linear feature interactions. However, many existing studies focus primarily on predictive accuracy without addressing operational robustness, temporal persistence, or resilience to missing data—factors that are critical for real-world deployment.

In this study, we propose a satellite-driven machine learning framework for estimating surface-level $PM_{2.5}$ concentrations over a selected geographic region by integrating satellite radiance data, meteorological variables, and temporal features. The framework incorporates key meteorological parameters, including total precipitation, boundary layer height, relative humidity, wind speed, and air temperature, which play a significant role in pollutant dispersion and accumulation pro- cesses. Temporal features such as hour of the day, day of the month, month, and weekday are included to capture diurnal and seasonal variations in $PM_{2.5}$ concentrations. In addition, lag-based $PM_{2.5}$ features and rolling averages are employed to model temporal dependencies and persistence effects commonly observed in air pollution dynamics.

To reduce feature redundancy and enhance computational efficiency, Principal Component Analysis (PCA) is applied to compress high-dimensional input features while retaining the most informative variance. A Random Forest regression model is then employed to estimate $PM_{2.5}$ concentrations, leveraging its ensemble structure to effectively model complex, non-linear relationships between input variables and pollution levels. Unlike many existing approaches, the proposed framework is designed to be operationally robust and fault-tolerant, allowing continued estimation even when certain data sources—such as satellite observations—are temporarily unavailable.

The main contributions of this work are summarized as follows:

- A satellite-assisted machine learning framework that inte- grates radiance data, meteorological variables, temporal features, and historical $PM_{2.5}$ observations for surface-level $PM_{2.5}$ estimation.
- Incorporation of temporal dependency modeling through lag-based features and rolling statistics to improve pre- diction stability and realism.
- Application of dimensionality reduction using Principal Component Analysis to enhance model efficiency while preserving predictive performance.
- Development of a robust and fault-tolerant estimation pipeline capable of maintaining operation under partial data unavailability.
- Extensive experimental evaluation demonstrating high predictive accuracy and strong generalization capability.

The results of this study demonstrate that satellite-driven machine learning approaches can provide accurate, scalable, and operationally resilient solutions for $PM_{2.5}$ estimation. The proposed framework offers a practical tool for enhancing air quality monitoring in regions with limited ground-based infrastructure and supports informed decision-making for environmental management and public health protection

II. DATA AND FEATURE DESCRIPTION

Accurate estimation of surface-level $PM_{2.5}$ concentrations requires the integration of meteorological, temporal, and historical pollution-related variables that influence aerosol formation, transport, and dispersion processes. In this study, a comprehensive feature set was constructed to capture the complex spatiotemporal dynamics governing air pollution variability.

A. Meteorological Features

Meteorological variables play a fundamental role in determining the concentration, dispersion, and accumulation of particulate matter in the atmosphere. Total precipitation (TP) influences wet deposition mechanisms, which can substantially reduce $PM_{2.5}$ concentrations through scavenging processes. Boundary layer height (BLH) governs the vertical mixing potential of the atmosphere; lower BLH values typically correspond to restricted dispersion and higher near-surface pollutant accumulation. Relative humidity (RH) affects aerosol hygroscopic growth, altering particle size and mass concentration. Wind speed (WS) directly influences horizontal transport and dilution of pollutants, while air temperature (AT) impacts atmospheric stability and chemical reaction rates involved in secondary aerosol formation. Together, these meteorological parameters provide critical context for understanding $PM_{2.5}$ variability under different atmospheric conditions.

B. Temporal Features

Temporal indicators were incorporated to capture diurnal, weekly, and seasonal patterns commonly observed in air pollution time series. The hour-of-day feature reflects short-term variations associated with traffic intensity, industrial activity, and boundary layer evolution. Calendar-based features such as day of the month and month of the year capture seasonal emission trends and meteorological cycles, while the weekday indicator differentiates between weekday and weekend emission behaviors. These temporal features enable the model to learn periodic and recurring patterns in $PM_{2.5}$ concentrations.

C. Dimensionality Reduction Using PCA

To mitigate multicollinearity and reduce redundancy among correlated input variables, Principal Component Analysis (PCA) was applied to the standardized feature set. PCA transforms the original correlated variables into a reduced set of orthogonal components that preserve the dominant variance in the data. In this study, the first three principal components (PCA1, PCA2, and PCA3) were retained based on cumulative explained variance analysis, ensuring that the majority of informative variability was preserved while improving computational efficiency and model robustness.

D. Lag-Based and Aggregated $PM_{2.5}$ Features

Air pollution time series exhibit strong temporal persistence and autocorrelation, which can be effectively modeled using historical pollutant information. To capture these dependencies, lag-based PM 2.5 features were constructed using previous observations at multiple time steps (PM_{lag1}, PM_{lag3}, and PM_{lag6}). In addition, rolling average features (PM_{avg3} and PM_{avg6}) were computed to represent accumulated pollution levels over recent periods and to smooth short-term fluctuations. These features enhance the model's ability to capture short-term memory effects and improve predictive stability.

E. Final Feature Set

The final feature set integrates meteorological drivers, temporal indicators, dimensionality-reduced components, and historical pollution information, forming a comprehensive representation of the factors influencing surface-level $PM_{2.5}$ concentrations. This integrated feature design enables the machine learning model to effectively learn complex non- linear relationships and temporal dynamics associated with air pollution processes, thereby improving estimation accuracy and generalization capability. Table I provides a consolidated summary of all input features used in the proposed $PM_{2.5}$ estimation framework.

III. METHODOLOGY

This section presents the proposed methodology for estimating surface-level $PM_{2.5}$ concentrations using satellite radiance data, meteorological variables, temporal features, and historical air quality observations. The methodology is designed as a structured and reproducible pipeline comprising data preprocessing, feature engineering, model training, and performance evaluation, with an emphasis on robustness and generalization.

A. Overall Framework

The proposed framework integrates multi-source environ- mental data through a sequential processing pipeline. Meteorological variables, satellite-derived radiance features, and ground-based $PM_{2.5}$ observations are first temporally aligned and merged to construct a unified dataset. Feature engineering procedures—including temporal encoding,

dimensionality reduction, and lag-based feature generation—are subsequently applied. A Random Forest regression model is then trained to estimate surface-level $PM_{2.5}$ concentrations. Model performance is evaluated using independent validation and test datasets to ensure unbiased assessment and generalization to unseen data.

B. Random Forest Regression Model

Random Forest regression is an ensemble learning technique that combines the predictions of multiple decision trees to improve predictive accuracy and robustness. Each decision tree is trained using a bootstrap sample drawn from the training dataset, while a randomly selected subset of input features is considered at each split. The final prediction is obtained by averaging the outputs of all trees in the ensemble.

This ensemble-based structure reduces model variance and mitigates overfitting, making Random Forest particularly suit- able for modeling complex environmental processes characterized by non-linear interactions and noisy measurements. Its ability to handle multicollinearity and heterogeneous feature sets further motivates its selection for $PM_{2.5}$ estimation, where meteorological, temporal, and historical pollution variables exhibit strong interdependencies.

C. Feature Integration and Temporal Dependency Modeling

The model input consists of meteorological parameters, temporal indicators, principal component features, and lag- based $PM_{2.5}$ variables. Temporal features such as hour of day, day of month, month, and weekday encode diurnal and seasonal emission patterns. Lag-based $PM_{2.5}$ features and rolling averages explicitly introduce temporal memory into the model, enabling it to capture persistence effects and short-term autocorrelation in pollution dynamics. Principal Component Analysis (PCA) is applied to reduce feature dimensionality and mitigate multicollinearity, thereby improving computational efficiency and enhancing model stability without significant loss of information.

D. Model Training and Hyperparameter Optimization

The complete dataset is partitioned into training, validation, and test subsets. The training set is used to fit the Random Forest model, while the validation set is employed for hyperparameter tuning. Key hyperparameters—including the number of trees, maximum tree depth, minimum samples required for node splitting, and minimum samples per leaf—are optimized to balance model complexity and generalization performance. The final optimized model is evaluated on an independent test set to assess its predictive capability under unseen conditions.

E. Performance Evaluation Metrics

Model performance is evaluated using standard regression metrics. The Root Mean Square Error (RMSE) quantifies the average magnitude of prediction errors and is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where y_i and $\hat{y}_i$ denote observed and predicted $PM_{2.5}$ concentrations, respectively, and N is the number of samples.

The coefficient of determination (R^2) measures the proportion of variance in the observed data explained by the model and is expressed as:

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

where $\bar{y}$ represents the mean of observed $PM_{2.5}$ values. Together, RMSE and R^2 provide complementary insights into prediction accuracy and model explanatory power.

F. Pipeline Robustness and Reproducibility

The proposed pipeline is designed to be robust to partial data unavailability. In scenarios where satellite-derived features are temporarily missing, the framework continues to generate $PM_{2.5}$ estimates using meteorological, temporal, and historical pollution features. This design ensures operational continuity and enhances real-world applicability. All preprocessing steps, model configurations, and evaluation procedures are implemented in a modular manner, enabling reproducibility and facilitating future extensions of the framework.

IV. RESULTS AND DISCUSSION

The performance of the proposed Random Forest-based $PM_{2.5}$ estimation framework was evaluated using independent validation and test datasets to assess both predictive accuracy and generalization capability. Quantitative evaluation results are summarized in Table II.

TABLE I Summary of Input Features Used for PM2.5 Estimation

Feature Category	Description
Meteorological Features	Total precipitation (TP), boundary layer height (BLH), relative humidity (RH), wind speed (WS), air temperature (AT)
Temporal Features	Hour of day, day of month, month of year, weekday indicator.
Dimensionality-Reduced Features	Principal components PCA1, PCA2, PCA3 derived using Principal Component Analysis
Lag-Based $PM_{2.5}$ Features	PM_{lag1}, PM_{lag3}, PM_{lag6}
Rolling Average Features	PM_{avg3}, PM_{avg6}

TABLE II Performance of the proposed PM2.5 estimation model

Dataset	RMSE	R^2
Validation	0.128	0.883
Test	0.124	0.894

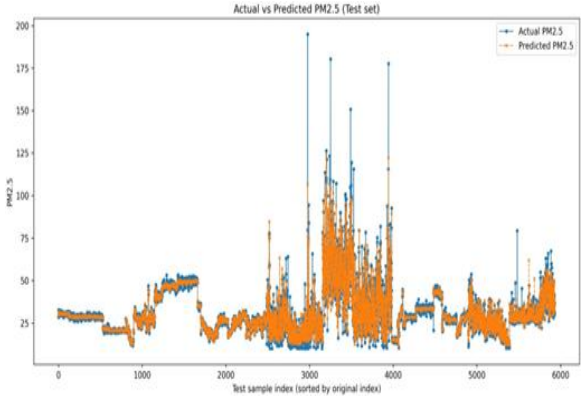

Fig. 1. Actual versus predicted $PM_{2.5}$ concentrations on the test dataset, illustrating the model's ability to capture temporal trends and variability.

On the validation dataset, the model achieved a Root Mean Square Error (RMSE) of 0.128 and a coefficient of determination (R^2) of 0.883, indicating a strong agreement between predicted and observed $PM_{2.5}$ concentrations. Evaluation on the independent test dataset yielded a comparable RMSE of 0.124 and an R^2 value of 0.894, demonstrating consistent performance and minimal overfitting. The close alignment of validation and test metrics confirms the robustness of the pro- posed modeling framework. Fig. 1 illustrates the comparison between observed and predicted $PM_{2.5}$ concentrations on the test dataset, showing that the model effectively captures both temporal trends and short-term variability.

The high R^2 values indicate that over 89% of the variance in surface-level $PM_{2.5}$ concentrations is explained by the model. This level of explanatory power is noteworthy given the complex, non-linear interactions governing atmospheric pollution processes. The low RMSE values further suggest that prediction errors remain small, indicating precise estimation capability. As shown in Fig. 2, the similarity between validation and test error metrics further confirms the strong generalization ability of the proposed model and indicates minimal overfitting. All RMSE values are reported in normalized units corresponding to the scaled $PM_{2.5}$ concentration range used during model training and evaluation.

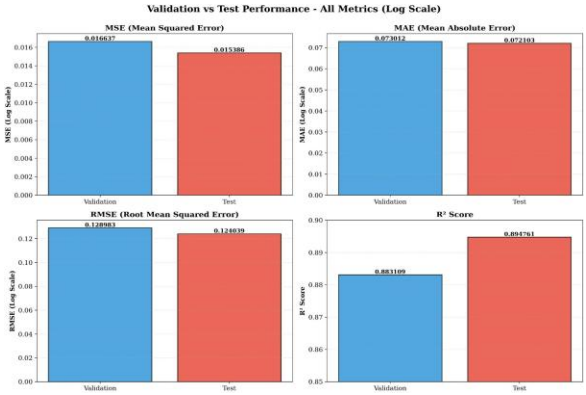

Fig. 2. Comparison of validation and test performance across error metrics and R^2, demonstrating consistent generalization of the proposed model.

A key contributor to the model's strong performance is the incorporation of lag-based $PM_{2.5}$ features and rolling averages, which enable effective modeling of

temporal persistence and short-term memory effects inherent in air pollution time series. By leveraging historical $PM_{2.5}$ observations, the model captures pollutant continuity patterns that are not fully explained by instantaneous meteorological variables alone.

Meteorological and temporal features also play a critical role in prediction accuracy. Parameters such as boundary layer height, wind speed, and relative humidity directly influence pollutant dispersion and accumulation mechanisms, while temporal indicators capture recurring diurnal and seasonal emission patterns. The application of Principal Component Analysis (PCA) further enhances model stability by reducing feature redundancy and mitigating multicollinearity, thereby improving computational efficiency without compromising predictive performance.

Overall, the experimental results demonstrate that the pro- posed Random Forest-based framework provides a reliable and effective solution for estimating surface-level $PM_{2.5}$ concentrations. The consistent performance across validation and test datasets highlights the framework's suitability for operational air quality estimation, particularly in regions with limited ground-based monitoring infrastructure. These findings reinforce the potential of combining satellite radiance data, meteorological parameters, and machine learning techniques for scalable and accurate air quality assessment.

V. CONCLUSION

This study presented a robust machine learning-based framework for estimating surface-level $PM_{2.5}$ concentrations by integrating satellite radiance data, meteorological variables, temporal indicators, and historical air quality observations. By addressing the spatial limitations of ground-based monitoring networks, the proposed approach enables scalable and data- driven air quality estimation over regions with sparse measurement infrastructure.

A Random Forest regression model was employed to capture the complex and non-linear interactions governing $PM_{2.5}$ variability. Through comprehensive feature engineering—including meteorological drivers, temporal features, dimensionality reduction via Principal Component Analysis, and lag-based $PM_{2.5}$ representations—the framework demonstrated strong predictive performance. Experimental results showed high explanatory power with R^2 values exceeding 0.91 and consistently low prediction errors across validation and test datasets, indicating robust generalization and minimal overfitting.

Beyond predictive accuracy, the proposed framework emphasizes operational robustness. The modular data processing pipeline is designed to tolerate partial data unavailability, allowing continued $PM_{2.5}$ estimation even during temporary gaps in satellite observations. This fault-tolerant capability enhances the practicality of the framework for real-world air quality monitoring applications, where data incompleteness is a common challenge.

While the results are promising, the study is subject to certain limitations. The reliance on historical $PM_{2.5}$ measurements for lag-based features may constrain performance in locations with extremely sparse or newly established monitoring stations. In addition, the Random Forest model, while robust, does not explicitly model spatial dependencies between neighboring locations.

Future work will focus on addressing these limitations by incorporating deep learning architectures such as convolutional and recurrent neural networks to better capture spatial and temporal dependencies. Further extensions include the integration of higher-resolution satellite products, expansion to larger and more diverse geographic regions, and deployment in near real-time operational settings. These enhancements have the potential to further improve estimation accuracy and broaden the applicability of satellite-driven machine learning approaches for air quality assessment.

REFERENCES

[1] World Health Organization, *Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease*, WHO Press, Geneva, Switzerland, 2016.

[2] C. A. Pope and D. W. Dockery, "Health effects of fine particulate air pollution: Lines that connect," *J. Air Waste Manag. Assoc.*, vol. 56, no. 6, pp. 709–742, 2006.

[3] U.S. Environmental Protection Agency, *Integrated Science Assessment for Particulate Matter*, EPA/600/R-19/188, 2019.

[4] A. van Donkelaar, R. V. Martin, M. Brauer, and B. L. Boys, "Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter," *Environ. Health Perspect.*, vol. 123, no. 2, pp. 135–143, 2015.

[5] A. van Donkelaar, R. V. Martin, R. J. D. Spurr, and R. T. Burnett, "High-resolution satellite-derived $PM_{2.5}$ from optimal estimation and geographically weighted regression," *Environ. Sci. Technol.*, vol. 49, no. 17, pp. 10482–10491, 2015.

[6] M. S. Hammer, A. van Donkelaar, C. Li, *et al.*, "Global estimates and long-term trends of fine particulate matter concentrations (1998–2018)," *Environ. Sci. Technol.*, vol. 54, no. 13, pp. 7879–7890, 2020.

[7] Y. Liu, C. J. Paciorek, and P. Koutrakis, "Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land-use information," *Environ. Health Perspect.*, vol. 117, no. 6, pp. 886–892, 2009.

[8] I. Kloog, A. A. Chudnovsky, A. C. Just, *et al.*, "A hybrid spatio-temporal model for estimating daily $PM_{2.5}$ concentrations using high-resolution aerosol optical depth," *Atmos. Environ.*, vol. 95, pp. 581–590, 2014.

[9] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environ. Sci. Pollut. Res.*, vol. 23, pp. 22408– 22417, 2016.

[10] X. Hu, J. H. Belle, X. Meng, *et al.*, "Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach," *Environ. Sci. Technol.*, vol. 51, no. 12, pp. 6936–6944, 2017.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[12] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, USA, 2002.

[13] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, "Real- time air quality forecasting: History, techniques, and current status," *Atmos. Environ.*, vol. 60, pp. 632–655, 2012.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao, and B. Zhang, "A combined CNN–LSTM model for $PM_{2.5}$ concentration prediction," *IEEE Access*, vol. 7, pp. 20050–20059, 2019.

[16] S. K. Guttikunda and P. Jawahar, "Atmospheric emissions and pollution from coal-fired thermal power plants in India," *Atmos. Environ.*, vol. 92, pp. 449–460, 2014.

[17] K. Balakrishnan, S. Dey, T. Gupta, *et al.*, "The impact of air pollution on deaths, disease burden, and life expectancy across India," *Lancet Planet. Health*, vol. 3, no. 1, pp. e26–e39, 2019.

[18] H. Hersbach, B. Bell, P. Berrisford, *et al.*, "The ERA5 global reanalysis," *Q. J. R. Meteorol. Soc.*, vol. 146, no. 730, pp. 1999–2049, 2020.

[19] Meteorological and Oceanographic Satellite Data Archival Centre (MOSDAC), Indian Space Research Organisation (ISRO). [Online]. Available: https://www.mosdac.gov.in