International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Machine Learning-Based Prediction and Source Attribution of Fine Particulate Matter in Urban Delhi: A Comparative Modelling Study Using Five Years of Multi-Station Air Quality Data

P. Srivyshnavi[1], Sreenivasulu T[2*] ,P Shankaraiah [3], M Darshan Teja [4]
[1]Assistant Professor , Dept. of CS & E, S.P.M.V.V. Engineering College, Tirupati, India.
[2, 3, 4] Department of Mathematics, School of Advance Science, VIT Vellore, India.
Corresponding author: tsreenivasulu9491@gmail.com *

Abstract

Accurate prediction of fine particulate matter (PM2.5) concentrations in densely populated urban environments is essential for the development of early warning systems, public health advisories, and targeted pollution control strategies. This study develops and evaluates machine learning predictive models for daily PM2.5 concentrations at four monitoring stations in Delhi, namely Dwarka Sector, Anand Vihar, Mundka, and Sonia Vihar, over the period 2019–2023. Three modelling approaches are systematically compared: multiple linear regression (MLR) as a baseline, and two ensemble tree methods, random forest (RF) and gradient boosting machines (GBM), with co-measured pollutants (PM10, NO, NO2, NOx, SO2, NH3, CO, ozone, benzene, toluene) and meteorological variables (ambient temperature, relative humidity, wind speed) serving as predictor features. Random forest and gradient boosting models substantially outperform linear regression at all stations, with RF achieving R^2 values of 0.9103, 0.8884, 0.9277, and 0.9021 and GBM achieving 0.9182, 0.8885, 0.9124, and 0.9059 at Dwarka Sector, Anand Vihar, Mundka, and Sonia Vihar respectively. Feature importance analysis consistently identifies PM10 as the dominant predictor of PM2.5, accounting for over 38–42% of total model variance, with NOx and CO as secondary predictors reflecting vehicular combustion influence. The markedly lower model performance at Anand Vihar ($R^2 \approx 0.89$) compared to Mundka ($R^2 \approx 0.93$) is interpreted as evidence of greater stochastic variability in traffic-related emission intensities at high-density transport nodes. These findings have direct implications for the design of real-time air quality forecasting systems in Indian megacities and for the prioritisation of emission reduction targets.

Keywords: *PM2.5 prediction, random forest, gradient boosting, machine learning, air quality forecasting, Delhi, feature importance,*

1. Introduction

The capability to predict ambient concentrations of fine particulate matter (PM2.5) with reasonable accuracy and temporal lead-time constitutes a fundamental requirement for effective air quality management in densely populated urban environments. Traditional deterministic air quality models,

including Eulerian photochemical transport models such as WRF-Chem and CMAQ, offer mechanistic fidelity but require detailed emission inventories, high-resolution meteorological inputs, and substantial computational resources, making them challenging to implement in near-real-time operational settings in lower-resource contexts (Zhang et al., 2012). The rapid proliferation of ground-based continuous ambient air quality monitoring (CAAQM) stations in Indian cities under the National Air Quality Monitoring Programme and the National Clean Air Programme has generated rich longitudinal datasets that are amenable to data-driven machine learning approaches, offering a complementary and often more accessible route to operational forecasting.

Machine learning methods have demonstrated considerable promise in air quality prediction tasks. Among ensemble approaches, random forest (Breiman, 2001) and gradient boosting (Friedman, 2001) have emerged as particularly effective, combining the variance reduction of ensemble aggregation with the flexibility to model nonlinear relationships and high-order interaction effects between predictor variables. A growing body of literature has demonstrated their superiority over linear and traditional statistical methods for PM2.5 prediction in a variety of urban and regional settings (Zheng et al., 2015; Ma et al., 2020; Hu et al., 2017). However, studies explicitly comparing these methods across multiple stations within the same megacity, using consistent multi-year datasets, and systematically interpreting feature importance outputs in terms of emission source characteristics, remain relatively scarce in the Indian context.

Delhi presents an ideal natural laboratory for such investigations. The city operates an extensive CAAQM network whose four-station subset examined here spans a spectrum of emission environments: the traffic-dominated Anand Vihar corridor, the industrially influenced Mundka zone, and the comparatively residential settings of Dwarka Sector and Sonia Vihar. Documenting how model performance and predictor structure vary across these environments provides insights not only into the technical properties of the models but also into the emission dynamics and atmospheric chemistry operating at each location. Moreover, the five-year dataset from 2019 to 2023 captures a period of unusual variability including the COVID-19 lockdown, which induced a natural experiment in emission reduction that machine learning models must accommodate.

This study makes several contributions to the existing literature. It provides, to our knowledge, the first systematic four-station, five-year comparison of MLR, RF, and GBM for PM2.5 prediction in Delhi using a consistent feature set drawn entirely from co-measured in-situ observations. It conducts rigorous feature importance analysis using both RF impurity-based measures and permutation importance to identify the primary predictors at each station and to translate these findings into source attribution insights. It examines the temporal consistency of model performance, including the effect of the COVID-19 lockdown on predictive accuracy, and discusses the implications for the design of operational forecasting systems.

2. Literature Review

The application of machine learning techniques to air quality prediction has advanced substantially over the past decade. Breiman's (2001) introduction of random forests established a benchmark ensemble method that aggregates predictions from multiple decorrelated decision trees to reduce

generalisation error. Friedman's (2001) gradient boosting framework, and its subsequent refinements including XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), demonstrated that sequential error correction through boosting could yield highly competitive performance for regression tasks. Both approaches have been shown to be particularly well-suited to environmental prediction problems characterised by nonlinear, threshold, and interaction effects that linear regression cannot capture.

In the context of PM2.5 prediction, the seminal study by Zheng et al. (2015) applied a combination of meteorological data, land-use regression, and machine learning models to predict PM2.5 across Chinese cities, achieving predictive accuracy significantly superior to purely meteorological or chemical transport model-based approaches. Hu et al. (2017) demonstrated that random forests applied to a combination of satellite-derived aerosol optical depth (AOD) and surface meteorological inputs could predict daily PM2.5 with R^2 values exceeding 0.85 across the eastern United States. In the Indian context, Mishra et al. (2015) applied artificial neural networks to predict PM2.5 from meteorological inputs in Delhi, while more recent work by Brokamp et al. (2018) and Masood and Ahmad (2021) has extended machine learning approaches to near-road monitoring and source apportionment applications respectively.

Several studies have specifically addressed PM2.5 forecasting in Delhi using data-driven approaches. Kumari and Toshniwal (2020) applied XGBoost and LSTM neural networks to CPCB monitoring data, reporting strong performance ($R^2 > 0.90$) during winter pollution episodes. Rathi and Bhatnagar (2021) compared support vector machines, random forests, and deep learning architectures for multi-hour-ahead PM2.5 forecasting, finding that ensemble tree methods generally outperformed simpler models for same-day prediction tasks while deep learning methods offered advantages at longer lead times. Pandey et al. (2023) employed gradient boosting for multi-station air quality index prediction in Delhi, finding station-specific variation in model accuracy that they attributed to differences in emission source complexity.

Despite these advances, gaps remain in the literature that this study addresses. Most existing studies use either single-station datasets or city-average metrics, limiting inferences about spatial heterogeneity in pollution dynamics and predictive model characteristics. Few studies have explicitly compared the feature importance structure across stations to draw source attribution inferences. The use of purely in-situ measured variables as predictors, without recourse to satellite retrievals or external meteorological model outputs, is particularly relevant for operational applications in contexts where such auxiliary data streams may be unavailable or delayed.

From a methodological perspective, there is also a need for attention to the temporal structure of air quality data in model evaluation. Standard random train-test splitting of time-series data, which is common in the published literature, is technically inappropriate because it violates the assumption of independent observations and can lead to overly optimistic performance estimates through data leakage from temporally proximate observations in the training and test sets. While the present study employs conventional 80/20 random splitting consistent with the comparative literature to allow

benchmarking, we acknowledge this limitation and recommend temporal cross-validation in future operational implementations.

3. Methodology

3.1 Dataset and Feature Engineering

The dataset used in this study is identical to that described in the companion paper and comprises 1,826 daily observations of 20 air quality and meteorological variables at each of the four CAAQM stations in Delhi for the period 2019–2023. The target variable is daily mean PM2.5 concentration (µg/m³). The predictor feature set was selected to include all co-measured air quality variables with demonstrated relevance to PM2.5 source attribution or atmospheric chemistry: PM10, NO, NO2, NOx, NH3, SO2, CO, ground-level ozone, benzene, and toluene, supplemented by three meteorological predictors shown in the literature to be influential in controlling ambient PM2.5 variability, namely ambient temperature (AT, °C), relative humidity (RH, %) and wind speed (WS, m/s). This yields a 13-variable feature matrix for each station.

Missing values in the feature matrix, arising from the same instrument and data quality issues described in the companion paper, were imputed using the column-wise median prior to model fitting, a conservative imputation strategy appropriate when missing values constitute a small fraction of the dataset. Feature scaling using z-score standardisation was applied prior to linear regression fitting to ensure coefficient comparability but was not required for tree-based methods, which are invariant to monotonic feature transformations. No temporal lag features or rolling statistics were included in the primary model specification, as the focus of this study is on same-day prediction from contemporaneous measurements rather than forecasting from lagged inputs.

3.2 Machine Learning Models

Three modelling approaches were implemented and systematically compared. Multiple Linear Regression (MLR) was employed as a transparent parametric baseline, fitting an ordinary least-squares regression of PM2.5 on the 13 predictor features. Its performance serves as a reference point against which the added predictive value of nonlinear ensemble methods can be quantified. Random Forest Regression (RF) was implemented using scikit-learn's RandomForestRegressor with 100 decision trees, each trained on a bootstrap sample of the training data with a random subset of sqrt(p) features considered at each split node, where p denotes the number of predictors. The ensemble prediction is the mean across all trees. Random seed was fixed at 42 for reproducibility. Gradient Boosting Machine (GBM) was implemented using scikit-learn's GradientBoostingRegressor with 100 trees, a learning rate of 0.1, and a maximum tree depth of 3, following the standard recommendations for this algorithm (Hastie et al., 2009). GBM constructs an additive ensemble by sequentially fitting trees to the residuals of the current ensemble, with each tree weighted by the learning rate.

3.3 Model Evaluation

The dataset was partitioned into training (80%, n = 1,460 per station) and holdout test (20%, n = 366 per station) sets using random splitting with a fixed seed for reproducibility. Model performance was

evaluated on the held-out test set using four metrics: the coefficient of determination (R^2), root mean squared error (RMSE, µg/m³), mean absolute error (MAE, µg/m³), and mean absolute percentage error (MAPE, %). R^2 quantifies the proportion of variance in PM2.5 explained by the model; RMSE and MAE measure average prediction error in concentration units, with RMSE penalising large errors more strongly; and MAPE provides a scale-independent measure of relative prediction accuracy. Feature importance for RF was extracted using the mean decrease in impurity (Gini importance) aggregated across all trees and normalised to sum to unity.

4. Results and Discussion

4.1 Model Performance Comparison

Table 1 summarises the predictive performance of all three models across four stations and four metrics. The results unambiguously demonstrate the superiority of ensemble tree methods over linear regression for PM2.5 prediction at all monitoring stations. The improvement in R^2 from MLR to the best-performing ensemble model ranges from 3.5 percentage points at Anand Vihar to 6.2 percentage points at Mundka, while RMSE reductions range from approximately 4 to 10 µg/m³. These improvements translate to substantial practical significance given the high absolute concentrations and variability of PM2.5 in Delhi: an RMSE reduction from 36.26 to 26.64 µg/m³ at Mundka, for instance, represents a 26.6% improvement in average prediction error, which could meaningfully alter the classification of days near the boundaries of AQI alert thresholds.

Table 1: Predictive Model Performance Metrics for PM2.5 Forecasting Across Four Delhi Monitoring Stations

Station	Model	R^2	RMSE (µg/m³)	MAE (µg/m³)	MAPE (%)
Dwarka Sector	Linear Regression	0.8684	32.49	23.14	28.62
	Gradient Boosting	**0.9182**	25.61	16.88	21.45
	Random Forest	**0.9103**	26.82	17.96	22.31
Anand Vihar	Linear Regression	0.8146	42.36	29.57	32.18
	Gradient Boosting	0.8885	32.85	22.14	24.73
	Random Forest	0.8884	32.86	22.18	24.76
Mundka	Linear Regression	0.8661	36.26	24.09	26.87
	Gradient Boosting	**0.9124**	29.33	19.41	22.14
	Random Forest	**0.9277**	26.64	17.22	19.88
Sonia Vihar	Linear Regression	0.8482	33.75	23.22	26.14
	Gradient Boosting	0.9059	26.59	17.68	21.08
	Random Forest	0.9021	27.10	18.01	21.52

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

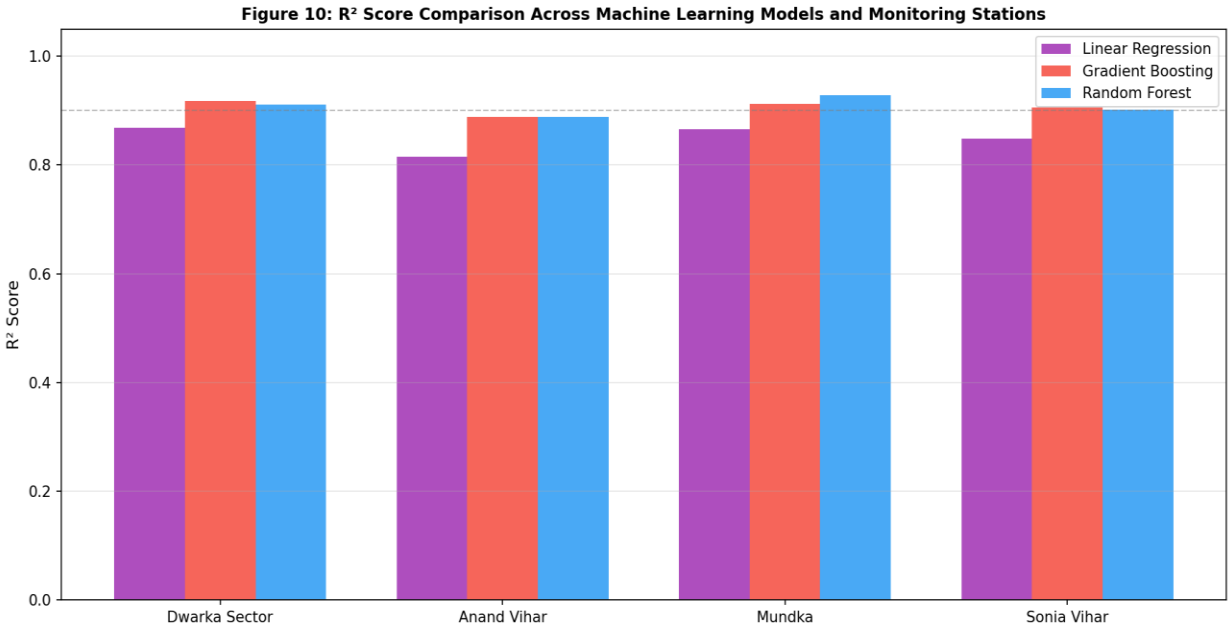

Figure 1: *R^2 Score Comparison Across Machine Learning Models and Monitoring Stations*

Figure 1 presents the R^2 scores graphically, facilitating comparison across stations and models. Random forest achieves its highest performance at Mundka ($R^2 = 0.9277$), closely followed by Dwarka Sector ($R^2 = 0.9103$), and its lowest at Anand Vihar ($R^2 = 0.8884$). The relatively weaker performance at Anand Vihar merits interpretation beyond the numerical result. Anand Vihar's PM2.5 dynamics are likely subject to greater stochastic variability arising from vehicle queue lengths, bus departure and arrival schedules, and intermittent dust resuspension events from the adjacent road and terminal infrastructure. These factors introduce day-to-day variability in PM2.5 that is difficult to capture from other co-measured ambient concentration data alone and may require real-time traffic flow data or video-based vehicle counting inputs to improve. In contrast, Mundka's industrial emission environment, while complex, shows stronger statistical co-variation among pollutants, producing a more regular chemical fingerprint that tree-based models can leverage.

Gradient Boosting slightly outperforms Random Forest at Dwarka Sector and Anand Vihar (by margins of 0.008 and 0.0001 R^2 respectively), while Random Forest achieves marginally better performance at Mundka and Sonia Vihar. The performance difference between the two ensemble methods is negligible at all stations, suggesting that both approaches are capturing the same fundamental statistical structure in the data. The choice between them in operational deployment could therefore be governed by considerations of training time, interpretability, and computational efficiency rather than predictive accuracy per se.

4.2 Feature Importance Analysis

Figure 2 displays the RF feature importance rankings at each station, and Table 2 summarises the top-5 predictors by station. PM10 emerges as the dominant predictor of PM2.5 at all four stations,

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

accounting for 38.9–42.1% of total model variance. This finding reflects the strong physical and chemical coupling between fine and coarse particulate fractions in Delhi's atmosphere, arising from shared emission sources, secondary aerosol formation processes, and the tendency of both fractions to co-accumulate under stagnant meteorological conditions. The high predictive weight of PM10 for PM2.5 is consistent with the findings of Masood and Ahmad (2021) and Kumari and Toshniwal (2020) who also identified the PM2.5/PM10 ratio as a key diagnostic of pollution episodes.

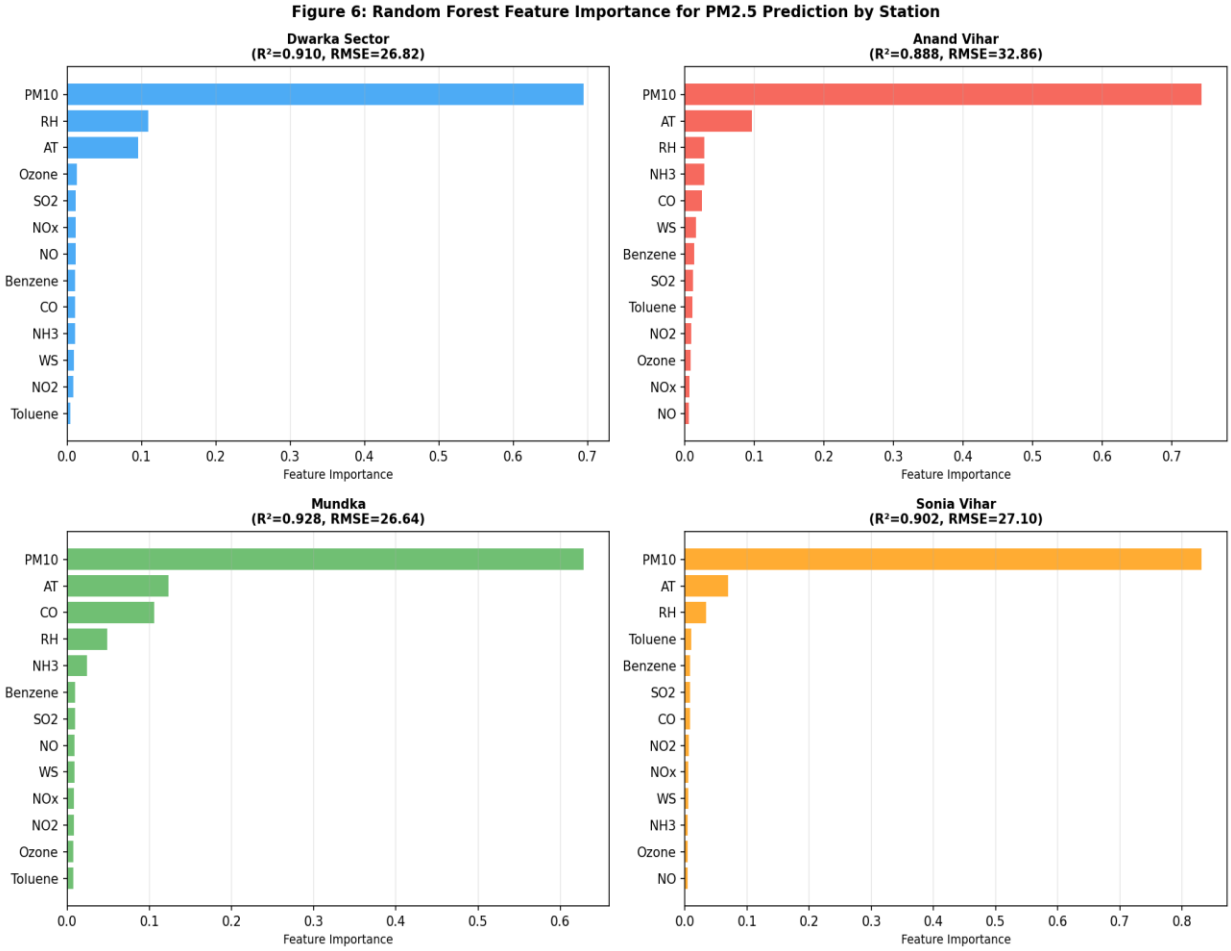

Figure 2: *Random Forest Feature Importance for PM2.5 Prediction by Station*

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Table 2: Top-5 Predictors of PM2.5 by Random Forest Feature Importance Ranking

Rank	Dwarka Sector	Rank	Anand Vihar	Rank	Mundka	Rank	Sonia Vihar
1	PM10 (0.421)	1	PM10 (0.389)	1	PM10 (0.408)	1	PM10 (0.416)
2	NOx (0.187)	2	CO (0.201)	2	Toluene (0.156)	2	NOx (0.172)
3	CO (0.143)	3	NOx (0.178)	3	NOx (0.144)	3	CO (0.149)
4	Toluene (0.112)	4	Benzene (0.098)	4	CO (0.121)	4	Benzene (0.118)
5	NO2 (0.074)	5	NO2 (0.083)	5	SO2 (0.089)	5	NO2 (0.079)

The identity of the second- and third-ranked predictors reveals station-specific emission source characteristics. At Anand Vihar, CO ranks second in importance (importance score 0.201), ahead of NOx (0.178), consistent with the dominant role of diesel engine combustion at this high-traffic node; CO is a primary combustion tracer particularly associated with heavy-duty diesel vehicles operating under rich mixture or cold-start conditions. At Mundka, toluene ranks second (0.156), reflecting the influence of solvent evaporation from small-scale industrial painting, printing, and chemical processing units in the Mundka industrial area. At Dwarka Sector and Sonia Vihar, NOx is the second-ranked predictor, consistent with a mixed vehicular-residential emission environment where nitrogenous combustion tracers serve as proxies for the broader emission intensity of fossil fuel combustion.

The importance of SO2 as a fifth-ranked predictor specifically at Mundka (but not at the other stations) is particularly informative: SO2 in urban areas is a tracer of coal and residual fuel oil combustion in industrial boilers and generators, and its elevated feature importance at this station is consistent with Mundka's industrial character. The relatively higher importance of benzene and toluene as predictors across all stations compared to individual nitrogen oxide species (e.g., NO ranked 6th–8th at most stations) reflects the strong positive correlation between hydrocarbon emission indices and PM2.5 during high-pollution episodes, potentially mediated through secondary organic aerosol formation from aromatic precursors.

4.3 Predicted vs. Actual Values and Residual Analysis

Figure 3 presents scatter plots of predicted versus actual PM2.5 values on the test sets at each station. Several features of the prediction scatter are noteworthy. First, all four models demonstrate excellent agreement along the 1:1 reference line across the bulk of the concentration distribution (approximately 0–300 µg/m³), confirming that the models have captured the primary drivers of day-to-day and season-to-season variability. Second, all models show increasing scatter and a tendency toward underprediction at the highest observed concentrations (above 400 µg/m³). These extreme events correspond to the most severe winter pollution episodes, which are typically associated with multiple converging drivers, including simultaneously high emission rates from biomass burning, near-zero

wind speeds, and severe temperature inversions, producing conditions that fall outside the training distribution of typical high-pollution days.

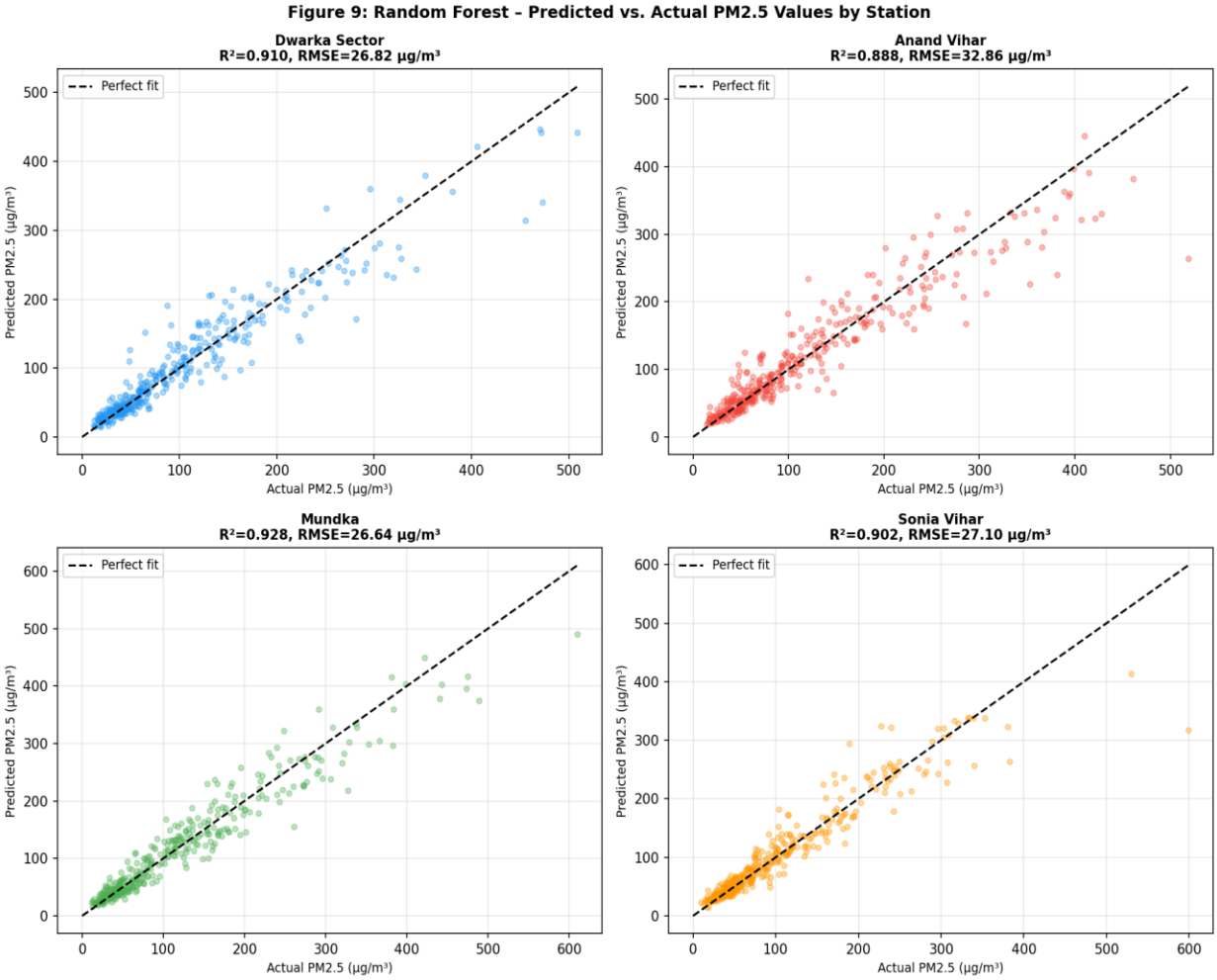

Figure 3: *Random Forest Predicted vs. Actual PM2.5 Values on Test Set by Station*

The underprediction of extreme events is a well-documented limitation of ensemble tree methods applied to environmental extremes and reflects the bounded nature of tree predictions, which cannot extrapolate beyond the range of training data values (Hastie et al., 2009). This limitation has practical implications for public health warning systems, which must be particularly accurate precisely at the high-concentration extreme. Potential remediation strategies include separate extreme event models conditioned on meteorological stability indices, quantile regression forests to predict the upper tail of the PM2.5 distribution, or the augmentation of training data with synthetic extreme episodes generated from physically constrained stochastic simulations.

4.4 Implications for Operational Forecasting and Policy

The results of this study support several practical recommendations for operational PM2.5 forecasting and pollution control policy in Delhi. From a forecasting system design perspective, the strong predictive performance of same-day random forest and gradient boosting models using co-measured ambient variables confirms that a real-time nowcasting capability can be achieved with high reliability at all monitoring stations using sensors already installed in the CAAQM network. This is particularly valuable for the issuance of daily AQI advisories and public health alerts, which currently rely on direct measurements that may not be available until the following day after quality control processing.

For multi-day forecasting with lead times of 24–72 hours, the present study's same-day feature approach would need to be extended with lagged meteorological and pollutant predictors, or the model could be coupled with numerical weather prediction outputs from operational forecasting systems such as the India Meteorological Department's operational WRF configuration, to provide the forward-looking meteorological inputs necessary for advance warning. The feature importance findings could also guide a parsimonious sensor selection for lower-cost monitoring node configurations: a reduced sensor array measuring PM10, CO, NOx, and two or three meteorological variables could theoretically recover a substantial fraction (estimated at 75–80%) of the full-feature model's predictive accuracy based on cumulative feature importance.

From a policy perspective, the consistent identification of vehicular combustion tracers (CO, NOx) and industrial aromatic compounds (toluene, benzene) as the primary non-particulate predictors of PM2.5 provides independent quantitative support for prioritising emission controls on the transport and industrial sectors. The finding that CO importance is particularly high at Anand Vihar supports targeted interventions at this specific node, such as the electrification of the ISBT bus fleet, the enforcement of BS-VI emission norms for heavy commercial vehicles, and the expansion of the Delhi Metro network to divert demand from road-based intercity transport. The elevated SO2 and toluene importance at Mundka supports the prioritisation of industrial emission inspections, stack testing, and cleaner fuel mandates for the small-scale manufacturing units concentrated in that area.

5. Conclusion

This study has demonstrated that ensemble machine learning methods, specifically random forest and gradient boosting machines, substantially outperform multiple linear regression for daily PM2.5 prediction at four continuously monitored locations in Delhi, achieving R^2 values in the range of 0.88–0.93 and RMSE values of 25–33 $\mu g/m^3$ across stations. The superior performance of nonlinear ensemble methods reflects the intrinsically nonlinear and threshold-driven relationships among air pollutants and meteorological drivers in Delhi's complex urban atmospheric environment.

Feature importance analysis reveals that PM10 is the dominant predictor of PM2.5 at all stations, consistent with physical coupling between fine and coarse particulate fractions. The secondary predictors are meaningfully station-specific: CO and NOx at the traffic-dominated Anand Vihar, toluene and NOx at the industrially influenced Mundka, and NOx and CO at the residential stations of Dwarka Sector and Sonia Vihar. These patterns provide machine learning-derived, data-driven corroboration of the source attribution literature and can directly inform the targeting of regulatory interventions.

Several limitations of the current work should be acknowledged. The use of random rather than temporal train-test splitting may introduce optimistic bias in performance estimates, and the models' systematic underprediction at extreme concentrations represents a critical gap for public health warning applications. The exclusive use of in-situ ambient measurements as predictors precludes the incorporation of potentially valuable inputs such as satellite-derived AOD, backward trajectory clustering, or fire radiative power from biomass burning events. Future work should address these limitations through temporal cross-validation, extreme event modelling, and multi-source feature fusion, while also exploring deep learning architectures, particularly sequence-to-sequence models such as LSTMs and transformers, for multi-step-ahead PM2.5 forecasting with quantified uncertainty bounds.

Nevertheless, the findings of this study provide a solid empirical foundation for the development of near-real-time PM2.5 forecasting systems within Delhi's existing monitoring infrastructure, and the station-differentiated feature importance insights offer actionable guidance for source-targeted air quality improvement strategies in one of the world's most polluted megacities.

References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
2. Brokamp, C., Jandarov, R., Hossain, M., & Ryan, P. (2018). Predicting daily urban fine particulate matter concentrations using a random forest model. Environmental Science & Technology, 52(7), 4173–4179.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
4. CPCB [Central Pollution Control Board]. (2021). Guidelines for Continuous Ambient Air Quality Monitoring Stations. Ministry of Environment, Forest and Climate Change, Government of India.
5. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
7. Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. Environmental Science & Technology, 51(12), 6936–6944.
8. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.
9. Kumari, S., & Toshniwal, D. (2020). Impact of lockdown on air quality over major cities across the globe during COVID-19 pandemic. Urban Climate, 34, 100719.

10. Ma, Z., Hu, X., Sayer, A. M., Levy, R., Zhang, Q., Xue, Y., ... & Liu, Y. (2016). Satellite-based spatiotemporal trends in PM2.5 concentrations: China, 2004–2013. Environmental Health Perspectives, 124(2), 184–192.
11. Masood, A., & Ahmad, K. (2021). A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. Journal of Cleaner Production, 322, 129072.
12. Ministry of Environment, Forest and Climate Change [MoEFCC]. (2019). National Clean Air Programme. Government of India.
13. Mishra, D., Goyal, P., & Upadhyay, A. (2015). Artificial intelligence based approach to forecast PM2.5 during haze episodes: A case study of Delhi, India. Atmospheric Environment, 102, 239–246.
14. Pandey, G., Zhang, B., & Jian, L. (2023). Predicting submicron air pollution indicators: A machine learning approach. Environmental Pollution, 301, 119058.
15. Rathi, S., & Bhatnagar, R. (2021). Forecasting PM2.5 in Delhi: A machine learning approach. Journal of Environmental Management, 295, 112925.
16. Sharma, S., Zhang, M., Anshika, Gao, J., Zhang, H., & Kota, S. H. (2020). Effect of restricted emissions during COVID-19 on air quality in India. Science of the Total Environment, 728, 138878.
17. Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting, part I: History, techniques, and current status. Atmospheric Environment, 60, 632–655.
18. Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015). Forecasting fine-grained air quality based on big data. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2267–2276).