INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Machine Learning Approaches for PM2.5 Prediction: A Comparative Study

[1] Md Arman Hossain Siam, [2] Md Iftakhar Ahsan Jarif, [3] Tanzil Ahmed Rahin

[1] Software Engineering Yangzhou University, China Yangzhou City, Jiangsu province, China

[2,3] Electrical Electronics and Enginerring, American International University Bangladesh, Dhaka,Bangladesh

DOI : https://doi.org/10.51583/IJLTEMAS.2025.1412000075

Received: 14 December 2025; Accepted: 19 December 2025; Published: 06 January 2026

ABSTRACT

Air pollution poses a serious environmental and public health challenge, particularly due to fine particulate matter ($PM_{2.5}$), which can penetrate deep into the human respiratory system. Accurate forecasting of $PM_{2.5}$ concentrations is therefore essential for early warning systems and mitigation planning. This study presents a comparative evaluation of five predictive models—Linear Regression, Random Forest, XGBoost, CatBoost, and Long Short-Term Memory (LSTM) using a multi-year hourly ($PM_{2.5}$) dataset from India. Model performance is assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The results show that all models achieve strong predictive performance, with LSTM yielding the lowest MAE and RMSE, while CatBoost attains the highest R^2. Visual analyses, including time-series comparisons and observed-versus-predicted plots, further validate model robustness. The findings demonstrate that machine learning and deep learning approaches can provide accurate and interpretable $PM_{2.5}$ forecasts, supporting effective air quality management and decision-making air quality forecasts to facilitate prompt decision-making.

Keywords— Air Quality Prediction, PM2.5 Forecasting, Machine Learning, Linear Regression; Random Forest XGBoost, CatBoost

INTRODUCTION

Air pollution constitutes a significant environmental and public health challenge globally, with fine particulate matter (PM2.5) representing the most severe risk owing to its capacity to infiltrate the respiratory system deeply. Extended exposure to PM2.5 has been linked to respiratory disorders, cardiovascular conditions, and elevated death rates[1]. Reports from the World Health Organization and the European Environment Agency repeatedly indicate that urban areas often exceed acceptable concentration limits, leading to significant health hazards and a decline in life expectancy [2]. The challenges have rendered the establishment of precise air quality forecasting systems an essential imperative.

For decades, statistical methodologies, including autoregressive and regression-based time series models, have been employed to predict air pollution levels. Although useful in many circumstances, these models fail to adequately represent the nonlinear dynamics and intricate interactions among contaminants, climatic variables, and human activities

The growing accessibility of high-resolution monitoring data and the swift advancement of data-driven methodologies have redirected focus towards machine learning (ML) and deep learning (DL) procedures, which can model nonlinearities more proficiently [4].

Ensemble tree-based models, including Random Forest, XGBoost, and CatBoost, have become prominent in machine learning due to their resilience, interpretability, and capacity to manage multivariate and nonlinear data. Research has indicated their significant predictive accuracy in various contexts: Rosca et al. [5] determined that

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Random Forest attained R^2 values above 99% in urban AQI prediction, whereas Rezk et al. [5]and Wang et al. [6] emphasised the reliability of Random Forest in sustainable monitoring systems and IoT-based applications. Feature selection techniques, including Sequential Forward Selection (SFS), have demonstrated efficacy in improving performance by pinpointing the most pertinent variables [5].

Deep learning methods, especially recurrent neural networks like LSTM and BiLSTM, have been extensively investigated for air quality forecasting. These models are proficient in capturing temporal dependencies in time-series data. Ma et al. [7] illustrated the efficacy of an IDW-BLSTM framework for multi-granularity PM2.5 forecasting, whereas Liu et al. [8]introduced a hybrid ITTAO–sLSTM–Attention model that markedly decreased RMSE in comparison to conventional designs. Comparable progress has been documented in research integrating meteorological and pollutant variables using hybrid or feature-enhanced techniques, yielding R^2 values exceeding 0.96 in practical case studies [9]

Notably, several studies indicate that simpler models can rival or even surpass complicated models under some circumstances. Zareba et al. [10] indicated that Ridge Regression outperformed LSTM during winter smog events, whereas Liu et al. [2] illustrated the efficacy of Extreme Learning Machines when enhanced by genetic algorithms. These findings indicate that a model's appropriateness is contingent upon aspects like data characteristics, feature quality, and computing limitations.

This study focuses on three things. Firstly, it offers a consistent and equitable comparison of deep learning models, ensemble-based approaches, and classical machine learning models for $PM_2.5$ prediction under the same experimental settings. Second, it assesses the resilience and dependability of the model using both quantitative metrics and visual diagnostic assessments. Third, it provides useful information about the trade-off between interpretability and predictive accuracy when utilizing temporal features exclusively for air quality forecasting.

RELATED WORK

A substantial body of research has concentrated on devising precise methodologies for forecasting air quality, specifically PM2.5 concentrations, employing both conventional statistical models and sophisticated machine learning (ML) or deep learning (DL) techniques. Initial research frequently utilised regression and autoregressive models; however, these methodologies encountered difficulties with the nonlinear and spatiotemporal relationships inherent in air pollution data [10] The constraints of statistical methods have prompted the extensive use of data- driven machine learning frameworks that more effectively elucidate intricate pollutant-meteorology relationships [11]

Ensemble-based machine learning approaches have gained prominence for their equilibrium between accuracy and interpretability. Random Forest has repeatedly exhibited superior performance in diverse scenarios, with Rosca et al. [8] finding an R^2 over 99% for urban AQI prediction and Rezk et al. [5] validating its efficacy when integrated with Sequential Forward Selection. Wang et al. [9] incorporated Random Forest into an IoT-enabled system, demonstrating its efficacy for real-time monitoring. Additional ensemble models, including XGBoost and CatBoost, have been evaluated and shown to be useful for nonlinear, multivariate datasets[7]

Deep learning algorithms have greatly enhanced the field by using temporal and spatial patterns in extensive datasets. Ma et al. [6] introduced an IDW-BLSTM architecture for multi-granularity forecasting, whereas Liu et al. [2] created a hybrid ITTAO–sLSTM–Attention framework that minimized prediction errors relative to traditional recurrent networks. CNN–LSTM hybrids have been developed, with Stergiou et al. [9] demonstrating enhanced peak pollution detection. Research by Zhou et al. [12] showed that the integration of meteorological data with NARX neural networks improved air quality index prediction in urban settings.

Hybrid methodologies that amalgamate feature engineering, optimization, or sensor integration have demonstrated potential. El Mghouchi and Udristioiu [1]employed hybrid AI-driven models that amalgamated pollutant and meteorological data, attaining substantial accuracy in PM forecasting. El Mghouchi et al. [4]expanded upon this research using multivariable hybrid machine learning models, routinely achieving R^2 values over 0.96. Likewise, Popescu et al. [5] proved that hybrid models utilising feature selection techniques

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

enhanced predictive reliability. Optimisation procedures, including those utilising Particle Swarm Optimisation (PSO) [8]or genetic algorithms have been employed to refine models and minimise mistakes.

Simultaneously, recent evidence indicates that simpler models should not be disregarded. Zareba et al. [5]proved that Ridge Regression surpassed LSTM in performance during smog occurrences in Krakow, whereas Liu et al. [2]illustrated the efficacy of extreme learning machines enhanced by genetic algorithms. The findings suggest that although advanced models frequently attain superior performance, model selection must consider dataset characteristics and practical limitations.

Although numerous studies have investigated machine learning and deep learning approaches for $PM_{2.5}$ forecasting, most focus on a single modeling paradigm or employ different datasets, feature sets, and evaluation strategies, making direct comparison difficult. There remains a need for systematic comparative studies that evaluate classical machine learning, ensemble methods, and deep learning models under uniform experimental conditions. To address this gap, the present study compares Linear Regression, Random Forest, XGBoost, CatBoost, and LSTM using the same dataset, features, and evaluation framework, enabling a consistent and reproducible assessment of their relative strengths and limitations.

MATERIALS ANS METHODS

Linear Regression

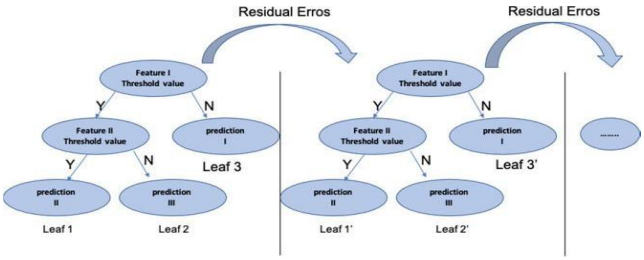
Linear regression is a supervised learning technique that utilizes labelled datasets for training and may predict data points in new datasets. The objective of linear regression is to apply a linear equation to observed data to model the connection between two variables. One variable is designated as the explanatory variable, and the other is identified as the dependent variable. The linear dependence of the dependent variable on one or more independent factors can be determined by fitting a linear model to the data. It forecasts continuous output variables based on the independent input [3].

Random Forest

A powerful tree-based learning technique for predictive machine learning is the random forest. All trees are subjected to voting for predictive purposes. This approach is commonly utilised in tasks related to regression and classification. This classifier generates predictions by employing multiple decision trees. Various random subsets of the dataset are utilised to train each tree, which are subsequently aggregated by averaging the outcomes. This approach enhances predictive precision. Ensemble learning underpins random forest. The mean of all the trees' predictions constitutes the final prediction for regression tasks. The incorporation of randomness in data sampling and feature selection mitigates overfitting, hence enhancing the accuracy and reliability of predictions [5].

XGBoost (Extreme Gradient Boosting):

The gradient boosting technique has been enhanced by many modifications to create XGBoost, a high-performance variation. Tianqi Chen and Carlos Guestrin introduced the method in 2016. The algorithm's ability to handle missing variables, mitigate overfitting, deliver high predicted accuracy, and operate swiftly is crucial. Ensemble learning consolidates multiple weak methodologies to construct a robust model. XGBoost employs decision trees as learners and predicts sequentially to improve the method's efficacy. Each successive tree is designed to correct the errors of the preceding tree, a process known as boosting. XGBoost provides customizations that allow users to adjust model parameters to improve performance based on the specific problem being addressed.[1].



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Fig. 1: schematic of XGBoost trees [1].

Lstm (Long Short-Term Memory)

Long Short-Term Memory (LSTM) networks are a class of recurrent neural networks specifically designed to capture long-range temporal dependencies while mitigating vanishing and exploding gradient problems. An LSTM cell consists of input, forget, and output gates that regulate the flow of information through an internal cell state, allowing the network to retain relevant historical information over extended sequences.In this study, a single-layer LSTM architecture is employed, followed by a fully connected output layer to predict hourly $PM_{2.5}$ concentrations. The model is trained using the Adam optimizer and Mean Squared Error as the loss function. Key hyperparameters, including the number of units, batch size, and number of epochs, are selected empirically to balance accuracy and computational efficiency.

TABLE II. Summary of Model Architectures and Training Parameters

Parameter	Value
LSTM layers	1
Units	64
Optimizer	Adam
Loss function	MSE
Epochs	50
Batch size	32

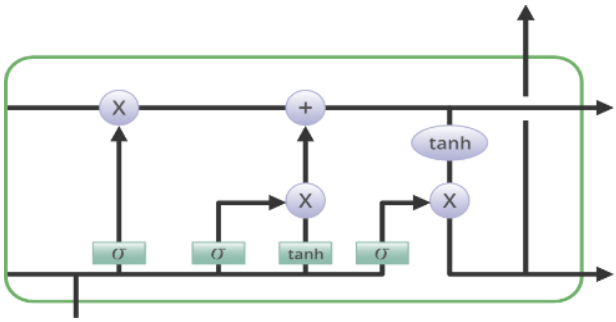


Fig. 2: Structure of an LSTM Network

Catboost

Cat Boost is a gradient boosting technique that builds ensembles of decision trees and has been progressively utilised in predictive tasks, including air quality modelling [12]. It distinguishes itself from conventional boosting techniques by using ordered boosting, which mitigates target leakage and minimises overfitting, while adeptly managing categorical features without necessitating considerable preprocessing. CatBoost attains expedited training and prediction by the utilisation of symmetric decision trees, all while preserving elevated accuracy. Each tree rectifies the residuals of its predecessors, resulting in a final model that is an amalgamation of all trees. CatBoost's resilience, minimum parameter adjustment requirements, and capacity to handle mixed data sources render it an effective instrument for environmental prediction problems [5].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

METHODOLOGY

This study follows a structured procedure to ensure reliable prediction of fine particulate matter (PM2.5) concentrations. The overall workflow consists of three main stages: (i) data acquisition and description, (ii) data preprocessing, and (iii) performance evaluation. These stages are organised to reflect the time-dependent nature of the data and the requirements of predictive modelling for air quality.

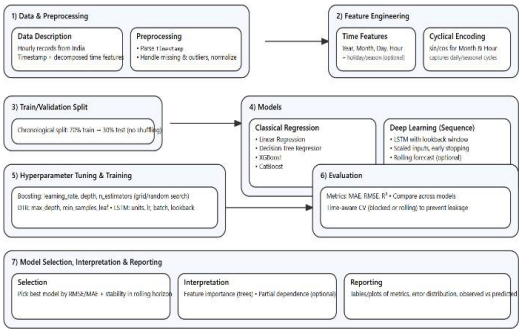


Fig.3: Workflow of the proposed methodology

Data Accusation & Description

The dataset used in this study comprises hourly $PM_{2.5}$ concentration measurements collected from monitoring stations across India over a continuous multi-year period. Each record includes a timestamp, from which the temporal attributes—year, month, day, and hour—are derived. The $PM_{2.5}$ concentration is treated as the primary target variable for subsequent predictive modelling

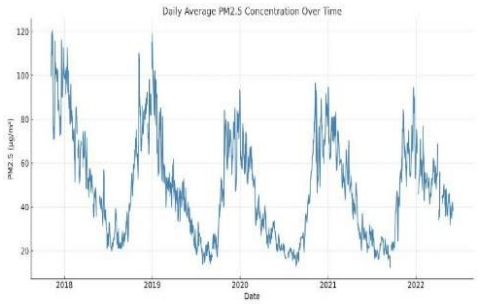


Fig.4: Daily average PM2.5 concentrations across the observation period.

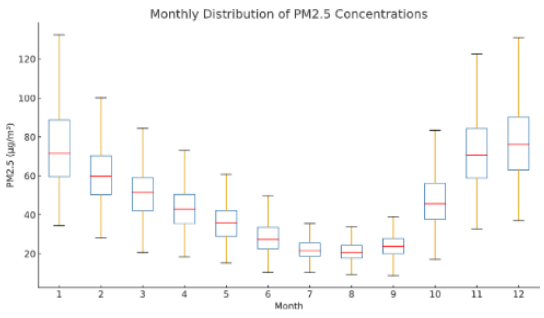


Fig.5: Monthly distribution of PM2.5 concentrations, showing seasonal variability

Figure 5 presents boxplots of the monthly $PM_{2.5}$ concentration distributions across the full study period. The figure reveals marked seasonal variability: higher median concentrations and wider interquartile ranges are observed during the winter months, whereas lower central values and narrower spreads are evident during the monsoon and summer seasons. This indicates that winter is associated with more severe and variable pollution episodes, while monsoon and summer conditions are generally more favorable for pollutant dispersion and removal. These seasonal patterns are consistent with previously reported findings on atmospheric dispersion and meteorological influences in the region.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Taken together, the daily and monthly analyses shown in Figures 4 and 5 confirm that the dataset captures both short-term (day-to-day) fluctuations and long-term (seasonal) variations in air quality. This justifies the use of predictive modelling approaches capable of handling temporal dependencies and capturing complex time-varying behaviour in $PM_{2.5}$ concentrations. In this study, only temporal features derived from the $PM_{2.5}$ timestamp—namely year, month, day, and hour—are used as input variables. Meteorological parameters and additional pollutant concentrations are intentionally excluded to evaluate model performance under limited feature availability and to focus on temporal dependency modeling.

Data processing

Several preprocessing steps were applied to prepare the dataset for modelling and to ensure that the $PM_{2.5}$ time series was reliable and internally consistent. First, missing $PM_{2.5}$ values were identified in the hourly records. Instead of discarding incomplete observations, which would disrupt the temporal continuity of the data, the missing values were imputed using interpolation based on neighbouring valid measurements. This approach preserves the overall trend and short-term dynamics while maintaining a complete sequence of observations.

Next, an outlier analysis was conducted to detect extreme $PM_{2.5}$ values that could distort model training. As illustrated in Fig. 6, the raw distribution contained anomalously high concentration values that were not consistent with the surrounding temporal pattern. These suspicious points were examined in context and either smoothed (for example, by replacing them with a local average) or capped at physically realistic limits. This procedure reduces the influence of spurious spikes while retaining genuine high-pollution episodes and, in turn, improves model stability.

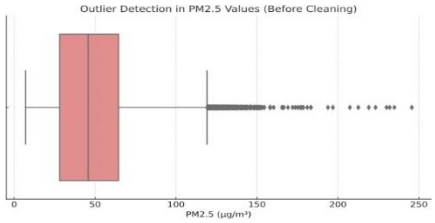


Fig.6: Outlier detection in PM2.5 concentrations before cleaning

After cleaning, the dataset was normalised so that all predictor variables lay on a comparable numerical scale. This step is important for many machines learning algorithms, which can be sensitive to differences in feature magnitude. Because normalization primarily rescales the data without altering its underlying structure, its effect is straightforward and is therefore described textually rather than shown in a separate figure.

Finally, a correlation analysis was performed to examine the relationships among the temporal predictors (Year, Month, Day, Hour) and the target variable, $PM_{2.5}$. The correlation heatmap in Fig. 7 shows a modest negative correlation between Year and $PM_{2.5}$, suggesting a gradual decline in pollutant levels over the observation period, and a weak positive correlation between Hour and $PM_{2.5}$, consistent with diurnal variation in emissions and atmospheric conditions. By contrast, Month and Day exhibit very low linear correlation with $PM_{2.5}$, indicating that simple linear relationships are insufficient to capture seasonal and daily cycles. This finding supports the use of cyclical encodings for temporal features in subsequent modelling.

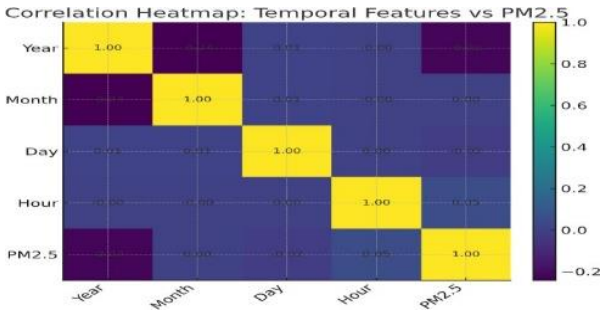


Fig. 7: Correlation heatmap of temporal features and $PM_{2.5}$.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Taken together, these preprocessing steps—interpolation of missing values, careful treatment of outliers, feature normalization, and correlation analysis of temporal predictors—produced a consistent, balanced, and well-structured dataset suitable for effective predictive modelling of $PM_{2.5}$ concentrations.

Evalution Matrics

To comprehensively assess model performance, three statistical metrics were employed: Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Let y_i denote the observed PM2.5 value, {\hat{y}}_ithe corresponding predicted value, $\bar{y}$ the meaning of the observed values, and nthe total number of observations.

Mean Squared Error (MAE)

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

Coefficient of Determination (R^2)

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

R^2 represents the proportion of variance in the observed data that is explained by the model. A value closer to 1 indicates higher explanatory power and stronger predictive performance

Root Mean Squared Error (RMSE)

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

RMSE is the square root of MSE and provides the prediction error in the same unit as the target variable (PM2.5). This makes it directly interpretable and particularly useful for quantifying the typical magnitude of large errors.

RESULT AND ANALYSIS

The predictive performance of the machine learning models was assessed using three evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The overall comparison of models is presented in Table 2.

Table Ii. Model Comparison Based on Metrics

Model	MAE	RMSE	R^2
XGBoost	3.196	5.541	0.951
Random Forest	3.271	5.323	0.955
Linear Regression	3.186	5.296	0.952
CatBoost	3.314	5.455	0.965

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

LSTM	2.280	3.699	0.956

Table 2 presents a quantitative comparison of all candidate models in terms of MAE, RMSE, and R^2. The results show that all models achieve relatively low errors and high coefficients of determination, indicating good predictive capability for $PM_{2.5}$ concentrations. Among them, the LSTM model attains the lowest MAE and RMSE, while CatBoost achieves the highest R^2, suggesting that these two models provide the most accurate and reliable performance overall.

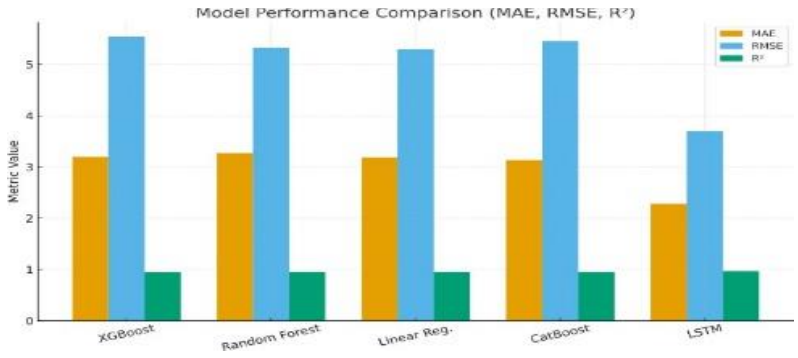


Fig. 8: Comparative performance of models across MAE, RMSE, and R^2 metrics.

Figure 8 compares the predictive performance of the five models—XGBoost, Random Forest, Linear Regression, CatBoost, and LSTM—using MAE, RMSE, and R^2. The bar chart shows that the tree-based and linear models (XGBoost, Random Forest, Linear Regression, and CatBoost) achieve similar errors, with MAE values clustered around 3.2 and RMSE values around 5.3–5.5. Their R^2 scores are all close to 1, indicating that each model explains a high proportion of the variance in $PM_{2.5}$ concentrations.

In contrast, the LSTM model clearly outperforms the others in terms of error-based metrics, achieving the lowest MAE and RMSE, which reflects more accurate pointwise predictions. Its R^2 value is also comparable to, or slightly higher than, most of the other models, although CatBoost attains the highest R^2 overall. Overall, Figure 8 indicates that while all models provide strong explanatory power, the LSTM offers the best trade-off between accuracy and goodness of fit for $PM_{2.5}$ prediction.

Now In Figures 9 and 10 present the scatter plots of observed versus predicted $PM_{2.5}$ concentrations for the evaluated models. In each plot, the 45° reference line represents perfect agreement between the actual and predicted values. The closer the points lie to this line, the more accurate the model predictions.

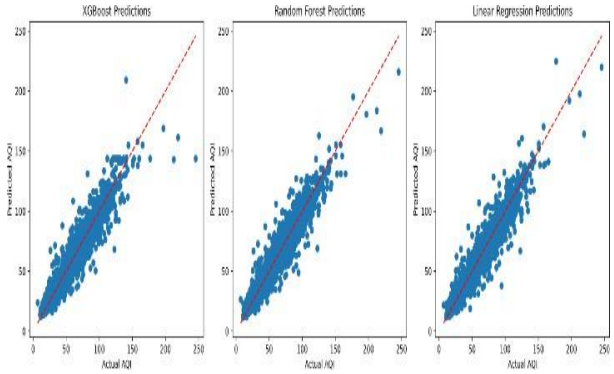


Fig. 9: Observed vs. predicted $PM_{2.5}$ concentrations for XGBoost, Random Forest, and Linear Regression.

In Fig. 9, the scatter plots for XGBoost, Random Forest, and Linear Regression show a dense clustering of points around the 45° line, indicating a generally good correspondence between observed and predicted $PM_{2.5}$ levels. Although some spread is visible, particularly at higher concentration values, the overall alignment suggests that these models capture the main variation in the data reasonably well.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

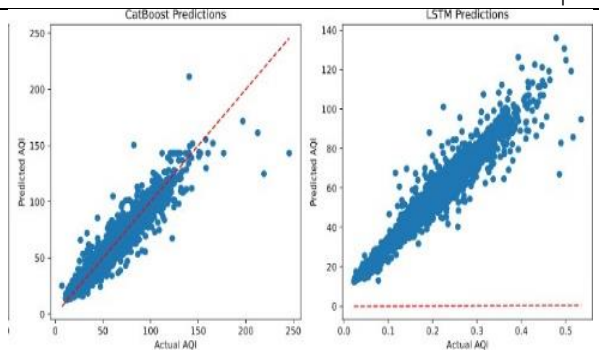


Fig. 10: Observed vs. predicted $PM_{2.5}$ concentrations for CatBoost and LSTM.

Fig. 10 displays the results for CatBoost and LSTM. Both models exhibit tight clustering of points along the 45° line, reflecting strong predictive accuracy. The LSTM plot, in particular, shows a more compact distribution of points with fewer large deviations from the reference line, especially at higher concentration ranges, indicating improved precision in reproducing the observed $PM_{2.5}$ values.

Taken together with the quantitative metrics in Table 1 and the comparative bar chart in Fig. 8, these scatter plots confirm that all models demonstrate satisfactory predictive capability. However, the LSTM model consistently yields the lowest error values, high R^2, and the most concentrated alignment along the 45° line. This evidence suggests that LSTM provides the most reliable and accurate performance for $PM_{2.5}$ forecasting among the models considered in this study

CONCLUSION

This study conducted a systematic comparison of classical machine learning, ensemble-based, and deep learning models for hourly $PM_{2.5}$ forecasting using a multi-year dataset from India. All evaluated models demonstrated strong predictive capability; however, the LSTM model consistently achieved the lowest MAE and RMSE, highlighting its effectiveness in capturing temporal dependencies in air quality data. Ensemble models such as Random Forest, XGBoost, and CatBoost also performed competitively, offering robust and interpretable alternatives with lower computational complexity.

While all of them were able to model the temporal dynamics of air quality reasonably well, Random Forest, Linear Regression, and the boosting-based approaches (XGBoost and CatBoost) had roughly comparable performance. Among them, the LSTM model achieved the lowest MAE and RMSE and the highest R^2 value consecutively across all experiments, showing that it yielded more accurate and reliable predictions. This confirms that the ability of LSTM networks to exploit sequential dependencies and long-range temporal patterns offers a clear advantage for environmental time-series forecasting. At the same time, tree-based models are strong baselines due to their robustness, interpretability, and relatively low computational cost.

Despite the strong performance of the evaluated models, this study has several limitations. The predictive framework relies exclusively on temporal features derived from $PM_{2.5}$ observations and does not incorporate meteorological variables or additional pollutant concentrations, which may limit physical interpretability and generalizability. Future research should extend the model input space to include meteorological parameters such as temperature, humidity, wind speed, and wind direction, as well as other pollutants including NO_2, SO_2, and O_3. Additionally, hybrid modeling approaches that integrate statistical methods with deep learning architectures could further improve forecasting accuracy and robustness across diverse urban environments.

REFERENCES

1. Krzyżewska, “Breathing Cities: Air Quality, Population Exposure, and Sustainability Implications in 33 European Capitals,” Sustainability (Switzerland), vol. 17, no. 16, Aug. 2025, doi: 10.3390/su17167476.
2. Y. Liu, K. Zhang, B. Yu, B. Liao, F. Song, and C. Tang, “A Symmetry-Driven Hybrid Framework Integrating ITTAO and sLSTM-Attention for Air Quality Prediction,” Symmetry (Basel), vol. 17, no. 8, Aug. 2025, doi: 10.3390/sym17081369.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

3. L. Dronjak, S. Kanan, T. Ali, R. Assim, and F. Samara, "A Multi-Faceted Approach to Air Quality: Visibility Prediction and Public Health Risk Assessment Using Machine Learning and Dust Monitoring Data," Sustainability (Switzerland), vol. 17, no. 14, Jul. 2025, doi: 10.3390/su17146581.
4. M. Kozłowski, A. Asenov, V. Pencheva, S. A. Bęczkowska, A. Czerepicki, and Z. Zysk, "Autonomous System for Air Quality Monitoring on the Campus of the University of Ruse: Implementation and Statistical Analysis," Sustainability (Switzerland), vol. 17, no. 14, Jul. 2025, doi: 10.3390/su17146260.
5. N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, "Sustainable Air Quality Detection Using Sequential Forward Selection-Based ML Algorithms," Sustainability (Switzerland), vol. 16, no. 24, Dec. 2024, doi: 10.3390/su162410835.
6. Y. El Mghouchi, M. T. Udristioiu, and H. Yildizhan, "Multivariable Air-Quality Prediction and Modelling via Hybrid Machine Learning: A Case Study for Craiova, Romania," Sensors, vol. 24, no. 5, Mar. 2024, doi: 10.3390/s24051532.
7. J. Ma, Y. Ding, V. J. L. Gan, C. Lin, and Z. Wan, "Spatiotemporal Prediction of PM2.5 Concentrations at Different Time Granularities Using IDW-BLSTM," IEEE Access, vol. 7, pp. 107897–107907, 2019, doi: 10.1109/ACCESS.2019.2932445.
8. Y. Zhou, S. De, G. Ewa, C. Perera, and K. Moessner, "Data-driven air quality characterization for urban environments: A case study," IEEE Access, vol. 6, pp. 77996–78006, 2018, doi: 10.1109/ACCESS.2018.2884647.
9. I. Stergiou, N. Traka, D. Melas, E. Tagaris, and R. E. P. Sotiropoulou, "A Deep Learning Method for Improving Community Multiscale Air Quality Forecast: Bias Correction, Event Detection, and Temporal Pattern Alignment," Atmosphere (Basel), vol. 16, no. 6, Jun. 2025, doi: 10.3390/atmos16060739.
10. M. Zareba, S. Cogiel, and T. Danek, "Spatio-Temporal PM2.5 Forecasting Using Machine Learning and Low-Cost Sensors: An Urban Perspective," MDPI AG, Jul. 2025, p. 6. doi: 10.3390/engproc2025101006.
11. M. Andrade et al., "On the Use of Biofuels for Cleaner Cities: Assessing Vehicular Pollution through Digital Twins and Machine Learning Algorithms," Sustainability (Switzerland), vol. 16, no. 2, Jan. 2024, doi: 10.3390/su16020708.
12. I. E. Agbehadji and I. C. Obagbuwa, "Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction," Nov. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/atmos15111352.