National Conference on Behavioural Economics and Intelligent Decision Systems for Climate Change and Sustainable Development
Journal of Advanced Research in Alternative Energy, Environment and Ecology
Volume 13, Issue 1&2 - 2026

Research Article

# Timeseries Forecasting of Delhi's Air Quality Index Using Statistical and Neural Network Models

Dushyant Sharma[1], Prashant Kumar[2]

[1]Mathematics and Computing Department of Applied Sciences National Institute of Technology Delhi, India
[2]Department of Applied Sciences National Institute of Technology Delhi, India
**DOI:** https://doi.org/10.24321/2455.3093.202606

I N F O

**E-mail Id:**
242432005@nitdelhi.ac.in
**How to cite this article:**
Sharma D, Kumar P. Timeseries Forecasting of Delhi's Air Quality Index Using Statistical and Neural Network Models. *J Adv Res Alt Energ Env Eco* 2026; 13(1&2): 108-114.

Date of Submission: 2025-10-04
Date of Acceptance: 2025-11-01

A B S T R A C T

Delhi consistently experiences severe air quality episodes, with particulate loads frequently breaching recom mended limits. Anticipating the Air Quality Index (AQI) with adequate lead time is central to timely advisories, exposure man agement, and responsive control actions. This work assembles a comparative forecasting framework spanning statistical baselines, machine learning, and deep sequence models to ... predict AQI along with $PM_{2.5}$ and $PM_{10}$ (both in $\mu g\ m^{-3}$). The study evaluates classical time-series tools (ARIMA, Prophet), non-linear regres sors (Support Vector Regression, Random Forest), and recurrent neural networks (Long Short-Term Memory). Using five years of hourly observations, we adopt a uniform pipeline for cleaning, scaling, and strictly forward-in-time validation. Empirical results show LSTM and Random Forest deliver consistent gains over the statistical baselines, capturing rapid fluctuations and seasonal shifts more faithfully. Overall, the analysis underscores the value of hybrid, data-driven approaches for reliable urban air quality forecasting and supports targeted mitigation in highly polluted settings.

**Keywords:** Air Quality Index, Time-Series Forecasting, Machine Learning, Deep Learning, ARIMA, Prophet, Support Vector Regression, LSTM, Random Forest, Delhi

## Introduction

Air pollution in large metropolitan regions has emerged as a persistent environmental and public health challenge, with Delhi frequently ranking among the worst-affected cities. Con centrations of particulate matter ($PM_{2.5}$ and $PM_{10}$), nitrogen oxides, and carbon monoxide often exceed national and inter national standards, elevating risks to human health and urban resilience.[1] The Air Quality Index (AQI) condenses multi- pollutant conditions into a single indicator and is widely used for communication, advisories, and action planning. Accurate and timely AQI prediction is therefore vital for safeguarding communities and guiding short-term interventions.

Forecasting remains difficult because pollution arises from interacting sources and meteorology, exhibiting strong nonlin- earity and pronounced seasonality. Traditional au- toregressive methods have long been used for time-series prediction, yet their linear assumptions can limit fidelity under complex dynamics. Recent advances in machine learning and deep learning provide mechanisms to capture


*Copyright (c) 2026: Author(s). Published by Advanced Research Publications*

nonlinear behaviour and temporal dependencies directly from historical records.

This study develops and benchmarks a set of models for Delhi's AQI forecasting, with emphasis on $PM_{2.5}$ and $PM_{10}$ as dominant drivers. We compare statistical approaches (ARIMA, Prophet), non-linear machine learning methods (Support Vec- tor Regression, Random Forest), and sequential deep networks (Long Short-Term Memory). The results offer a clear view of the comparative strengths of these families and inform the design of robust forecasting pipelines for operational air quality management in a highly polluted megacity.

## Literature Survey

Air quality prediction has been explored using both traditional statistical tools and modern machine learning ap- proaches.[2] Earlier research predominantly relied on Autoregressive Integrated Moving Average (ARIMA) models, which capture persistence, trends, and short-term seasonal behaviour.[6] While suitable for relatively stable dynamics, ARIMA suffers from limitations when confronted with the nonlinear interactions typical of urban pollution, such as rapid chemical transformations, sudden emission spikes, or abrupt meteorological changes.

The emergence of machine learning brought new opportu- nities for handling these complexities. Models such as Support Vector Regression (SVR) introduced the ability to capture nonlinear pollutant–meteorology relationships through kernel functions, while ensemble techniques like Random Forest gained popularity due to their robustness to noise and their capacity to represent intricate feature interactions without extensive manual feature engineering. These methods demon- strated better predictive accuracy compared to classical linear frameworks, particularly under conditions of irregular emis- sions and heterogeneous pollutant sources.[2] Recent studies have explored increasingly complex architectures, including hybrid deep learning models,[1] meteorology-integrated hy- brid frameworks,[9,10] and temporal graph networks,[3] demonstrating the field's rapid evolution.

Motivated by these developments, this study provides a comparative analysis of statistical methods (ARIMA, Prophet), non-linear machine learning algorithms (SVR, Random For- est), and deep sequential models (LSTM). The goal is to evaluate their relative strengths within a unified experimental setup that reflects the high emission intensity, meteorological variability, and severe seasonal pollution cycles characteristic of Delhi.

## Dataset and Preprocessing

The analysis draws on the Air Quality Data in India (2015–2020),[8] which contains hourly records from national monitoring stations managed by the Central Pollution Control Board (CPCB).[7] For this study, only Delhi-specific entries are retained. Variables include AQI, AQI category, and pollu- tant concentrations ($PM_{2.5}$, $PM_{10}$, NO, $NO_2$, $NO_x$, $NH_3$, CO, $SO_2$, $O_3$, Benzene, Toluene, and Xylene). These indicators support both AQI forecasting and pollutant-specific modelling.

**Key dataset attributes**

- **Datetime:** Timestamp of measurement in hourly resolutioin
- **Pollutants:** Major regulated pollutants relevant to AQI.
- **AQI:** Computed overall index value.
- **AQI Bucket:** Category labels (e.g., Good, Moderate, Poor).

Figures 1–3 summarise the underlying data patterns, includ- ing distributional skewness, long-term cycles, and pollutant interdependence.

LSTM forecasting architecture with look-back window L and a dense readout layer. Hidden (h) and cell (c) states propagate across time; the final hidden state feeds a dense head to produce the one-step-ahead prediction $\hat{y}_{t+1}$.

## Experiments and Methodology

### Experimental Setup

To ensure a fair and reproducible comparison across models, we designed a consistent evaluation framework. The available dataset was divided chronologically, with 80% of the records used exclusively for training and the remaining 20% reserved for testing. This forward-in-time split was chosen to mimic realistic forecasting conditions and to minimise the risk of information leakage between past and future observations.

All algorithms were implemented in the Python ecosystem using established libraries. The statistical baselines were developed with the statsmodels implementation of ARIMA and the prophet package for trend-seasonality decomposition. For machine learning baselines, we employed scikit-learn to train Support Vector Regression (SVR) and Random Forest models. Sequential deep learning mod- els, specifically Long Short-Term Memory (LSTM) networks, were implemented using TensorFlow/Keras.

Model hyperparameters were tuned through a validation subset carved out from the training split, allowing robust opti- misation without contaminating the test data. This ensures that performance metrics reported in the results section correspond strictly to unseen observations, thereby providing a reliable assessment of each model's generalisation ability.

### Evaluation Metrics

To quantitatively assess model performance, we employed a set of widely accepted regression metrics:

ISSN: 2455-3093
DOI: https://doi.org/10.24321/2455.3093.202606

**Mean Absolute Error (MAE)**

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

$$RMSE = \sqrt{MSE} \quad (3)$$

**LSTM Model Architecture**

The Long Short-Term Memory (LSTM) network is designed to capture temporal dependencies in sequential data. Its archi- tecture is based on memory cells with input, forget, and output gates. The update equations for each time step t are:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}t \quad (8)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \odot \tanh(C_t) \quad (10)$$

Here, $x_t$ is the input, $h_t$ the hidden state, $C_t$ the cell state, $\sigma$ the sigmoid function, and $\odot$ elementwise multiplication.

## Results

The performance comparison of the different forecasting models on the test dataset is summarised in Table I. The findings reveal a clear ranking in predictive capability across the model families. Traditional linear approaches such as ARIMA and Prophet provide a useful point of reference but fall short when confronted with the nonlinear variability and abrupt fluctuations that characterise Delhi's air quality. Their reliance on stationary assumptions and linear trend decomposition limits their adaptability to sudden pollution surges.

**Table I.Comparative Performance of Forecasting Models for pm2.5 and pm10 . Bold Values Indicate The Best Scores Across Metrics**

| | | | | | |
|---|---|---|---|---|---|
| ARIMA | PM2.5 | 32.5 | 1850.2 | 43.0 | 0.75 |
| | $PM_{10}$ | 45.1 | 3100.5 | 55.7 | 0.72 |
| Prophet | PM2.5 | 28.9 | 1540.7 | 39.2 | 0.80 |
| | $PM_{10}$ | 40.2 | 2650.1 | 51.5 | 0.78 |
| SVR | PM2.5 | 25.1 | 1210.3 | 34.8 | 0.85 |
| | $PM_{10}$ | 35.6 | 2100.8 | 45.8 | 0.83 |
| LSTM | PM2.5 | 18.2 | 850.6 | 29.1 | 0.91 |
| | $PM_{10}$ | 25.8 | 1450.2 | 38.1 | 0.89 |
| Random Forest | PM2.5 | 19.5 | 920.4 | 30.3 | 0.89 |
| | $PM_{10}$ | 27.1 | 1580.9 | 39.8 | 0.88 |

In contrast, machine learning and deep learning frameworks show a marked enhancement in predictive accuracy. The Random Forest model demonstrates notable skill, largely due to its ensemble mechanism which aggregates multiple decision trees. This structure enables it to capture complex pollutant interactions, account for outliers, and generalise well across diverse conditions. Support Vector Regression also surpasses the linear baselines, though its kernel-based learning still strug- gles to fully adapt under highly volatile pollution episodes.

The Long Short-Term Memory (LSTM) architecture consistently achieves the most reliable forecasts, with the lowest error scores (MAE, RMSE) and the highest $R^2$ values. Its recurrent structure and gating mechanisms allow it to en- code temporal dependencies effectively, learning both long- term seasonal cycles and short-lived pollution spikes. The close alignment between LSTM predictions and observed AQI trends further underscores its robustness. The superior perfor- mance of the LSTM network can be attributed to its recurrent architecture and gating mechanism, which are specifically designed to capture long-range temporal dependencies in time- series data. Similarly, the Random Forest's strong performance stems from its ensemble nature, which reduces variance by averaging multiple decision trees, making it robust to noise and effective at modeling complex interactions without overfitting.

Overall, the comparative results emphasize that hybrid, data- driven techniques—particularly those leveraging sequential deep learning architectures—offer the greatest promise for delivering accurate and stable AQI forecasts in urban environ- ments marked by rapid variability and multifactor influences.

Figure 7 highlights the distribution of prediction errors across models, showing that the LSTM network achieves the narrowest spread of residuals and the highest $R^2$ values for both $PM_{2.5}$ and $PM_{10}$. The Random Forest model also performs competitively, owing to its ensemble structure which mitigates overfitting. A qualitative inspection of LSTM fore- casts, presented in Fig. 6, shows close alignment with observed time series, including during periods of elevated pollution.

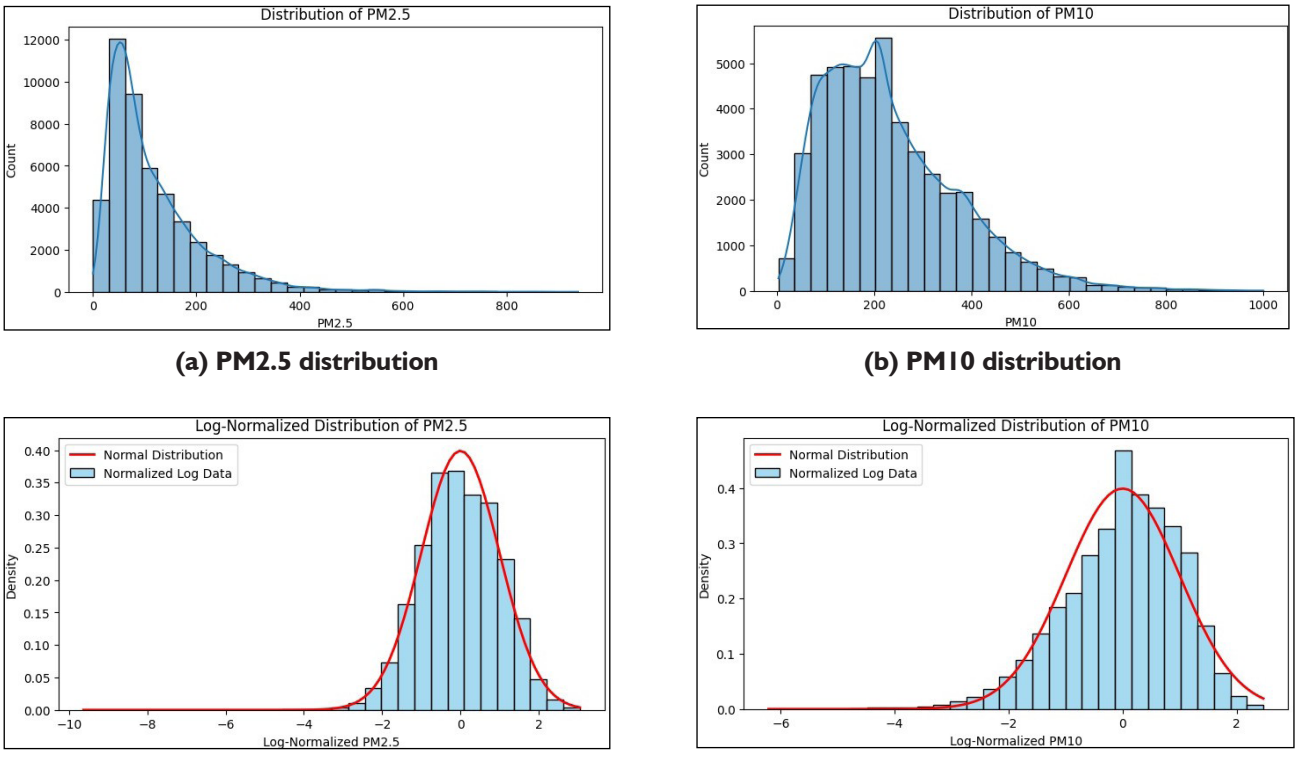
## Model Uncertainty and Limitations

Despite the promising results, several factors introduce uncertainty into the presented forecasting framework. First, the air quality data used in this study are subject to instrumental and sampling errors from ground-based monitoring stations, which can propagate through model training and evaluation. Additionally, spatial coverage across Delhi is uneven—some monitoring sites may not fully capture localised emission events, leading to potential under representation of micro-scale variability.

**ISSN: 2455-3093**
**DOI: https://doi.org/10.24321/2455.3093.202606**

A second source of uncertainty arises from the temporal resolution and missing data. Hourly records are occasionally missing or interpolated, which may affect the model's ability to learn transient pollution spikes. Although preprocessing techniques such as normalisation and outlier handling were applied, residual data inconsistencies remain a source of model uncertainty.

From the modelling perspective, algorithmic bias and hyperparameter sensitivity contribute to performance vari- ability. While Random Forest and LSTM models exhibit strong generalisation, their forecasts depend heavily on architecture depth, look-back window size, and feature scaling strategy. The absence of meteorological covariates such as wind speed, humidity, and temperature further limits interpretability under changing weather regimes. Future extensions should include these physical drivers and perform ensemble calibration to better quantify prediction confidence intervals.




**(a) PM2.5 distribution**

**(b) PM10 distribution**

**(c) Log-normalized PM2.5 distribution**

**(d) Log-normalized PM10 distribution**

**Figure 1.Exploratory Data Analysis (EDA) of Delhi's AQI dataset. The plots illustrate raw and log-normalized pollutant distributions, highlighting skewed patterns and heavy-tailed behavior typical of urban particulate data (units in $\mu g$ m−3)**

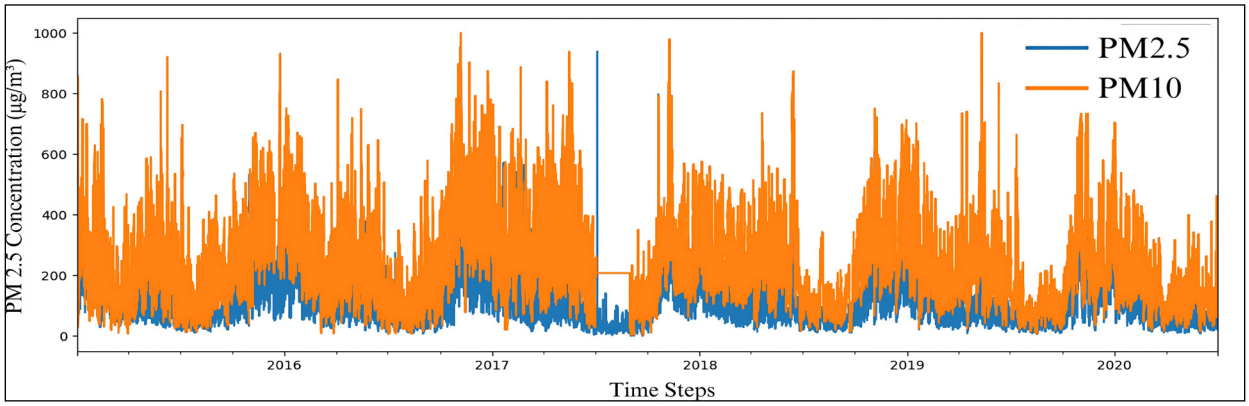



**Figure 2.Time series of PM2.5 and PM10 concentrations in Delhi (2015–2020), showing seasonal cycles and extreme pollution episodes (units in $\mu g$ m−3)**

ISSN: 2455-3093
DOI: https://doi.org/10.24321/2455.3093.202606

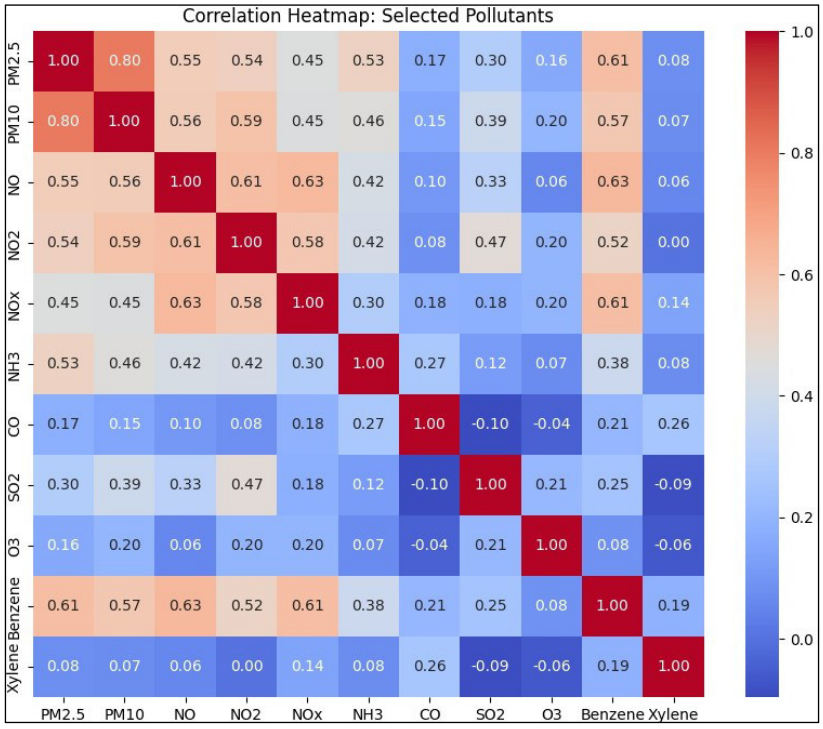



**Figure 3.Correlation structure between pollutants and AQI in Delhi, indicating strong positive association of AQI with particulate matter (PM2.5, PM10) and moderate links with gaseous pollutants.**

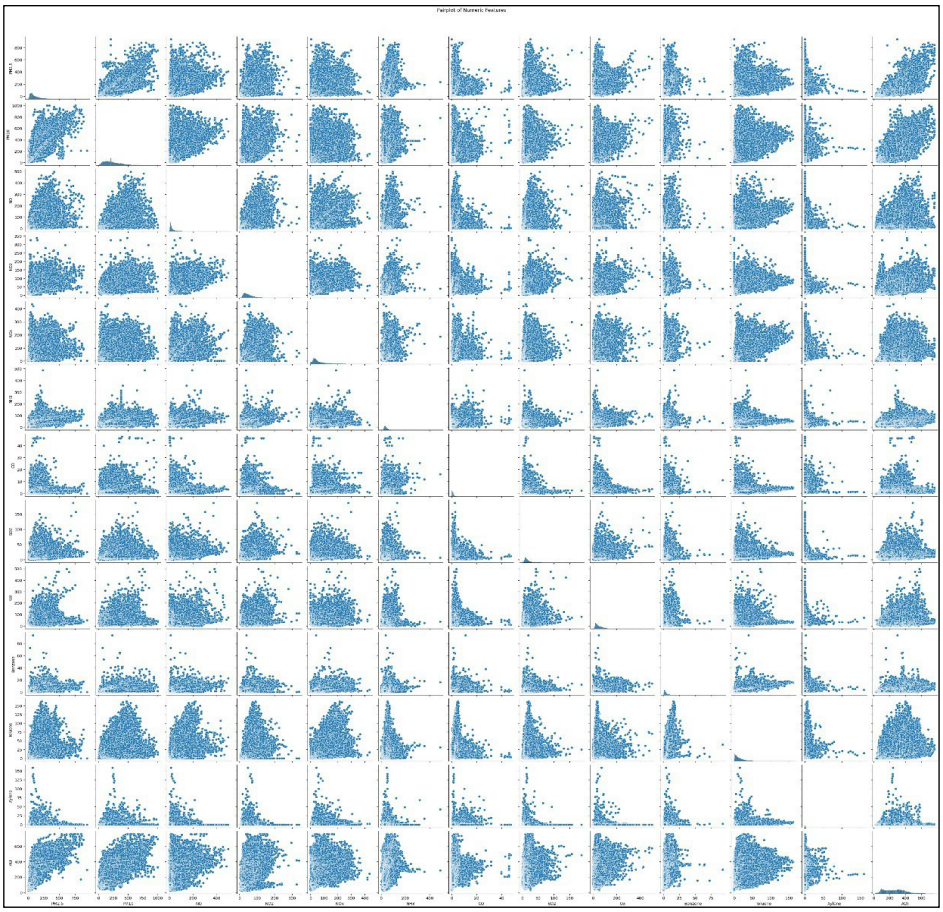


**Figure 4.Pairwise relationships between key pollutants. This visualization helps in identifying multicollinearity and understanding feature interactions and distributions across all considered variables**

**ISSN: 2455-3093**
**DOI: https://doi.org/10.24321/2455.3093.202606**

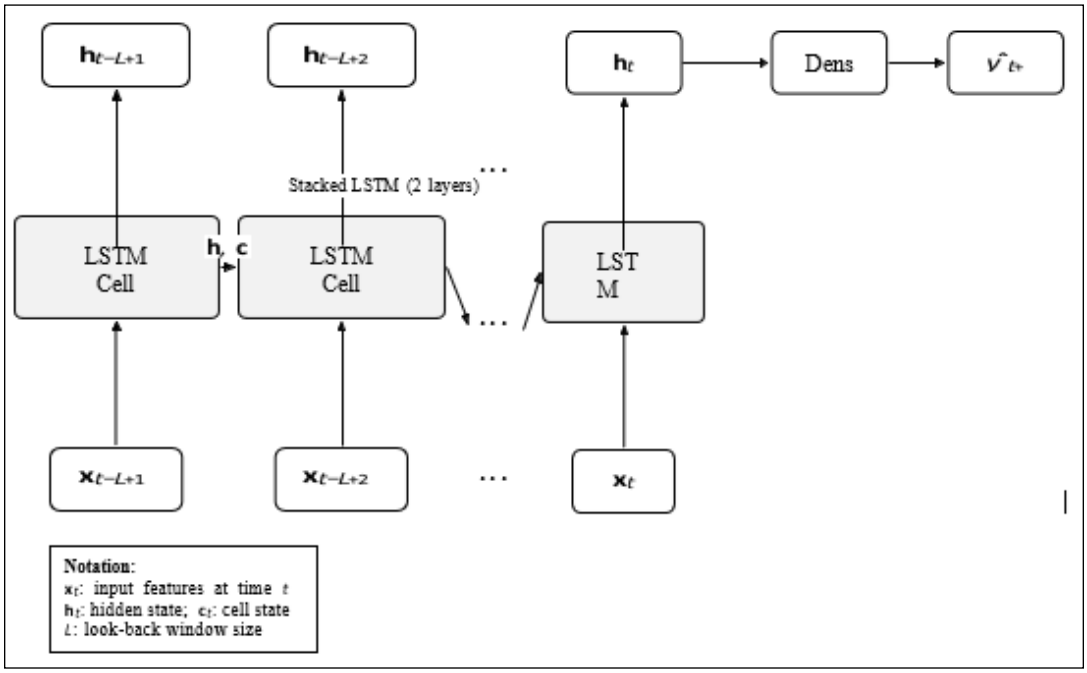



**Figure 5.LSTM forecasting architecture with look-back window L and a dense readout layer. Hidden (h) and cell (c) states propagate across time; the final hidden state feeds a dense head to produce the one-step-ahead prediction y^t+1**

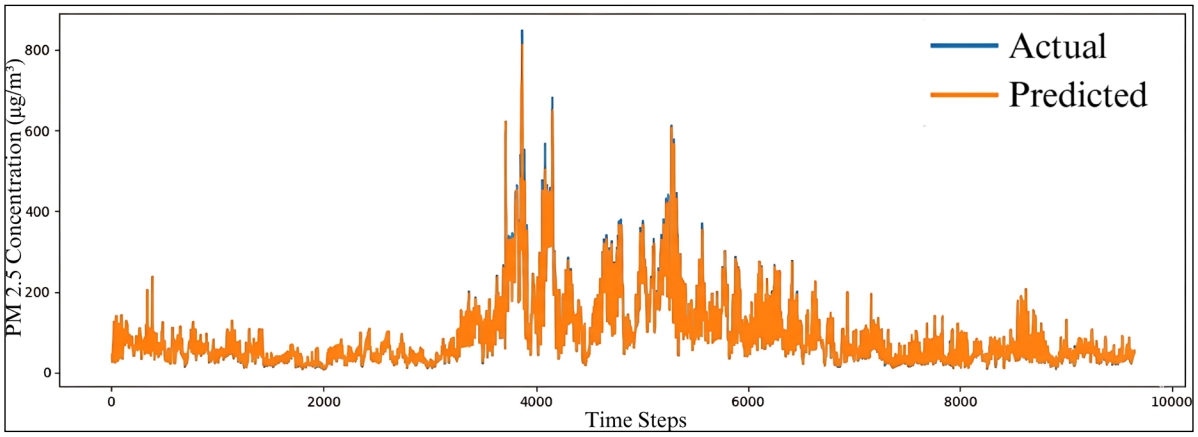



**Figure 6.Actual vs. Predicted values for the LSTM model on the PM2.5 time series.**

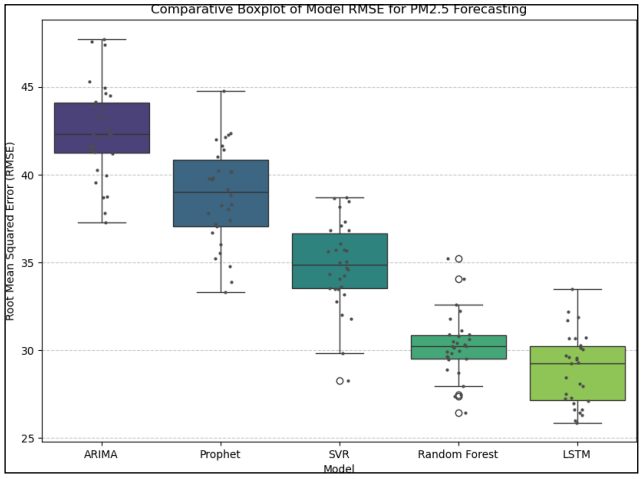



**Figure 7.Comparative boxplot of RMSE distributions for each model**

ISSN: 2455-3093
DOI: https://doi.org/10.24321/2455.3093.202606

## Conclusion

This study carried out a comprehensive comparison of mul- tiple forecasting strategies for Delhi's air quality, examining statistical baselines, machine learning regressors, and deep learning sequence models. The evaluation results demonstrated a clear performance hierarchy. As shown in Table I, traditional methods such as ARIMA and Prophet achieved $R^2$ values between 0.72 and 0.80, with relatively high error magnitudes (MAE 30–45). While these approaches provided useful baselines, they were unable to capture abrupt pollution spikes or complex nonlinear interactions.

Machine learning models offered substantial improvements. Support Vector Regression reduced MAE to 25.1 for $PM_{2.5}$ and

35.6 for $PM_{10}$, while Random Forest achieved even lower error scores (MAE 19.5 for $PM_{2.5}$ and 27.1 for $PM_{10}$), reflecting the benefits of ensemble learning in handling heterogeneity and noise. The best results, however, were obtained from the Long Short-Term Memory (LSTM) network, which reached MAE of 18.2 ($PM_{2.5}$) and 25.8 ($PM_{10}$), with $R^2$ values above 0.89. These findings confirm the advantage of recurrent architectures in modeling both gradual seasonal cycles and sudden high- intensity events.

The exploratory data analysis phase proved instrumental in shaping the modelling pipeline. Identifying skewed pollu- tant distributions, seasonal variability, and strong correlations between AQI and particulate matter justified the use of nor- malisation, variance-stabilising transformations, and temporal feature engineering. These steps ensured consistent model performance and avoided bias from heavy-tailed distributions.

Looking ahead, several extensions could enhance the predictive framework. Incorporating meteorological variables such as temperature, wind patterns, and humidity may improve early detection of sudden regime shifts.[4] Integrating multi-station data across Delhi would allow for spatially resolved forecast- ing, highlighting local hotspots. Moreover, recent advances in attention-based recurrent networks, graph neural networks, and transformer-based models offer promising directions for longer-horizon forecasts and enhanced interpretability.[5]

In summary, the study highlights that while statistical mod- els provide useful benchmarks, machine learning and espe- cially deep recurrent networks deliver superior accuracy and robustness. These findings reinforce the potential of hybrid, data-driven forecasting systems to support evidence-based policy-making and timely public advisories in cities facing persistent air quality challenges.

From a policy perspective, the proposed framework can serve as a foundational tool for municipal and regional au- thorities. By integrating such predictive models into air-quality monitoring networks, agencies like the CPCB and Delhi Pol- lution Control Committee (DPCC) could issue more accurate early warnings, dynamically adjust traffic management or industrial operations, and design localised emission-control measures. Furthermore, coupling model outputs with low-cost IoT sensor networks and citizen-science data could enhance spatial coverage, improve transparency, and support data- driven environmental governance. In the long term, such op- erational forecasting systems would facilitate proactive rather than reactive responses to pollution episodes, improving both public health outcomes and regulatory efficiency.

## References

1. H. Kumar, M. P. Singh, and R. Gupta, "A Hybrid CNN-LSTM Model for Spatiotemporal PM2.5 Concentration Forecasting in Delhi," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–12, 2022.
2. A. Sharma, S. Goyal, and A. Singh, "Machine Learning for Air Quality Forecasting: A Comprehensive Review," Journal of Cleaner Production, vol. 358, p. 132035, 2022.
3. R. Verma and P. Nagpal, "An Attention-Based Spatio-temporal Graph Convolutional Network for Air Quality Prediction in Urban India," Atmospheric Pollution Research, vol. 14, no. 5, p. 101734, 2023.
4. S. Jain, A. Patel, and S. Kumar, "Impact of Meteorological Parameters on Air Quality Index Prediction Using Deep Learning," in Proc. IEEE Int. Conf. on Big Data, Sorrento, Italy, 2023, pp. 2451–2460.
5. P. Li, Y. Wang, and Z. Zhang, "Transformer-Based Models for Time Series Forecasting: A Survey and Taxonomy," ACM Computing Surveys, vol. 55, no. 11, pp. 1–37, 2023.
6. D. Gupta and S. Singh, "Forecasting PM10 Concentrations in India: A Comparative Analysis of ARIMA, Prophet, and XGBoost," Environmen- tal Science and Pollution Research, vol. 29, pp. 78910–78925, 2022.
7. Central Pollution Control Board (CPCB), India. [Online]. Available: https://cpcb.nic.in/
8. "Air Quality Data in India (2015–2020)," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india
9. Y. Zhang, J. Chen, and L. Liu, "Hybrid LSTM–CNN Model for Air Quality Prediction Integrating Meteorological Parameters," Atmospheric Pollution Research, vol. 14, no. 5, p. 101255, 2023.
10. J. Li and P. Wang, "Meteorology-Driven Deep Learning for PM2.5 Forecasting: A Hybrid Approach," Environmental Modelling & Software, vol. 157, p. 105473, 2022.

**ISSN: 2455-3093**
**DOI: https://doi.org/10.24321/2455.3093.202606**