Integrated Air Quality Analytics: Forecasting and Pollution Source Identification Using Machine Learning and Deep Learning Approaches

Nikesh Yadav, Sandhya Kaprawan

Department of Information Technology University of Mumbai Mumbai, Maharashtra, India

Abstract—**Air pollution has evolved from a seasonal environ- mental challenge to a perennial public health crisis, particularly in rapidly urbanizing megacities. While traditional chemical transport models offer physical insights, they often lack the computational agility required for real-time, hyper-local forecasting. This research proposes a robust, hybrid framework for Air Quality Index (AQI) forecasting and pollutant source characterization. We introduce an Attention-based Bi-Directional Long Short-Term Memory (Attention-Bi-LSTM) network, de- signed to capture long-term temporal dependencies and weigh critical historical pollution events. Utilizing a comprehensive dataset spanning nine years (2017-2025) from New Delhi (CPCB) and comparative insights from the USA (EPA), the proposed model demonstrates significant empirical validity. On the test set (2024-2025), the model achieved a Root Mean Square Error (RMSE) of 44.59 and an R^2 score of 0.82, effectively predicting extreme pollution spikes associated with winter inversion and anthropogenic activities. Furthermore, this study moves beyond mere prediction to source attribution. Analysis of prominent pollutants reveals that while Particulate Matter (PM10/PM2.5) remains the primary driver of toxicity, secondary pollutants like Ozone ($O3$) and Nitrogen Dioxide ($NO2$) have emerged as significant contributors, accounting for over 25% of hazardous days. These findings provide policymakers with explainable, data- driven insights to transition from reactive measures to proactive air quality management.**

Key Words—**Air Quality Forecasting, Attention-Bi-LSTM, Deep Learning, Source Apportionment, Explainable AI (XAI), Urban Health Policy.**

I. INTRODUCTION

Breathing, the most fundamental human act, has become a calculated risk for millions inhabiting the world's dense urban centers. The World Health Organization (WHO) estimates that ambient air pollution accounts for 4.2 million premature deaths annually, acting as a ”silent killer” that exacerbates respiratory and cardiovascular diseases [4]. However, the crisis is not uniform; it varies deeply by geography, meteorology, and economic activity.

A. The Global Air Quality Crisis: A Tale of Two Regions

In the United States, cities like Los Angeles have battled photochemical smog for decades. While strict regulations have curbed industrial emissions, the region now faces a new, volatile threat: wildfire smoke, which causes sudden, non- linear spikes in PM2.5 levels. Conversely, in New Delhi, India, the ”Airpocalypse” is a complex cocktail of vehicular exhaust, construction dust, and the seasonal phenomenon of agricultural stubble burning. During winter, temperature inversions trap these pollutants, creating a toxic dome that frequently pushes the Air Quality Index (AQI) beyond the measurable limit of 500. This disparity highlights a critical gap in environmental science: a ”one-size-fits-all” model cannot suffice. We need forecasting systems that are not only accurate but also context- aware.

B. The Need for Advanced Forecasting

Historically, air quality forecasting relied on statistical linear models like ARIMA or deterministic Chemical Transport Models (CTMs). While ARIMA is effective for identifying linear trends, it fails to capture the stochastic volatility of modern urban pollution. CTMs, on the other hand, require immense computational power. The advent of Deep Learning (DL), specifically Recurrent Neural Networks (RNNs), offers a solution. However, standard LSTMs treat all past time steps equally. In reality, specific past events (e.g., humidity levels 48 hours ago) may be more predictive of today's smog than the levels yesterday.

C. Research Objectives and Contributions

This paper bridges the gap between high-level algorithmic innovation and ground-level policy application. We propose an end-to-end framework that integrates rigorous data preprocess- ing with an advanced Attention-based Bi-Directional LSTM (Attention-Bi-LSTM) architecture. The key contributions of this study are as follows:

- Novel Architecture Implementation: We deploy an Attention-Bi-LSTM model that processes time-series data in both forward and backward directions. This allows the model to "attend" to specific historical days (e.g., assigning higher weights to days with low wind speed), effectively capturing the non-linear volatility of Delhi's pollution.
- Long-Term Analytical Scope: Unlike studies limited to short durations, we utilize a comprehensive dataset spanning nine years (2017–2025). This allows for the capture of long-term climate cycles and validates the model against the "chronic volatility" of Delhi compared to the "event-driven" pollution of Los Angeles.
- Pollutant Source Characterization: Moving beyond mere prediction, we introduce a statistical characterization framework. We employ pollutant ratio analysis (PM vs. NO_2/O_3) as a statistical proxy for source apportionment, allowing for the identification of toxicity drivers (combustion vs. dust) without the need for expensive chemical mass spectrometry.
- Policy-Grade Explainability: We demonstrate how the model's 24-hour forecast window aligns directly with the Graded Response Action Plan (GRAP), enabling authorities to transition from reactive lockdowns to proactive pollution management.

II. LITERATURE REVIEW

The domain of air quality forecasting has witnessed a paradigm shift over the last half-decade, transitioning from statistical linearity to non-linear deep learning architectures.

A. Statistical vs. Deep Learning Approaches

Early research primarily utilized statistical models such as ARIMA. Wang *et al.* [1] demonstrated that while ARIMA provides a reliable baseline for linear trends, it struggles significantly with the stochastic nature of PM2.5 spikes caused by external shocks. To address this non-linearity, RNNs became the standard. Gupta and Kumar [2] compared Support Vector Regression (SVR) with LSTM networks on Delhi's pollution data, concluding that LSTMs reduced error rates by nearly 18% due to their ability to retain long-term memory. However, standard LSTMs suffer from the vanishing gradient problem and unidirectional processing, often missing future context in training sequences.

B. Hybrid and Attention-Based Mechanisms

Recent studies have focused on hybrid architectures to overcome these limitations. The integration of Convolutional Neural Networks (CNN) with LSTMs has been explored to extract spatial features before temporal processing [5]. Most notably, the introduction of the Attention Mechanism has revolutionized time-series forecasting. Li *et al.* [3] introduced an Attention-based Bi-LSTM for urban air quality in Beijing, showing that the model could "attend" to specific historical days with higher weights.

Newer iterations of these models have further optimized hyperparameter tuning and attention layers for specific pollutants like PM2.5 [4], [6]. Our research builds upon this foundation by coupling the Attention-Bi-LSTM architecture with a source characterization proxy [8], a combination less explored in current literature.

C. Explainable AI (XAI) in Environmental Science

A critical gap in existing literature is the "Black Box" nature of Deep Learning. While models predict accurately, they rarely explain the underlying drivers. Recent works using SHAP (SHapley Additive exPlanations) have begun to deconstruct model decisions [15], a methodology aligned with our goal of bridging algorithmic output and policy implementation.

III. METHODOLOGY

The proposed framework adopts a robust multi-stage pipeline comprising three key phases: data preprocessing, deep learning forecasting, and pollutant source characterization.

A. Data Acquisition and Preprocessing

Real-world environmental data is inherently noisy. We utilized datasets from the Central Pollution

Control Board (CPCB) for Delhi (2017-2025) to capture the high-volatility trends specific to the Indian subcontinent. To ensure data integrity, we applied the following preprocessing steps:

1) *Imputation:* Missing values were handled using MICE (Multiple Imputation by Chained Equations), which estimates missing pollutant values based on inter-variable correlations (e.g., inferring $PM_{2.5}$ from PM_{10} and visibility).
2) *Outlier Detection:* We employed the Isolation Forest algorithm to distinguish between sensor voltage errors (false positives) and genuine pollution spikes (true pos- itives).
3) *Cyclical Encoding:* Temporal features (Hour, Month) were transformed using sine and cosine functions. This ensures the neural network understands that ”Hour 23” is close to ”Hour 0”, preserving the cyclical nature of time.
4) *Normalization:* All features were scaled to the range [0, 1] using MinMax Scaler to facilitate efficient gradient descent convergence.

B. Proposed Architecture: Attention-Bi-LSTM

The core of our forecasting engine is the Attention-based Bi- Directional LSTM. Unlike standard LSTMs that process data only from past to future, the Bi-LSTM processes the sequence in both directions, comprehensively capturing pollution trends. To prioritize critical time steps, the Attention layer calculates a context vector c_t as a weighted sum of hidden states, allowing the model to assign high attention weights (α_t) to days with relevant meteorological triggers (e.g., low wind speed) that lead to pollutant accumulation.

C. Source Characterization Proxy

Forecasting answers ”when,” but policy requires ”what.” We employ a statistical characterization of pollutants using Prominent Pollutant tags. By analyzing the frequency distribution of primary pollutants ($PM_{2.5}$, PM_{10}) versus secondary pollutants (Ozone, NO_2), we identify dominant toxicity drivers. *Note: This ratio analysis serves as a statistical proxy for source apportionment, distinguishing between dust-dominated events and combustion-dominated events without requiring chemical mass spectrometry.*

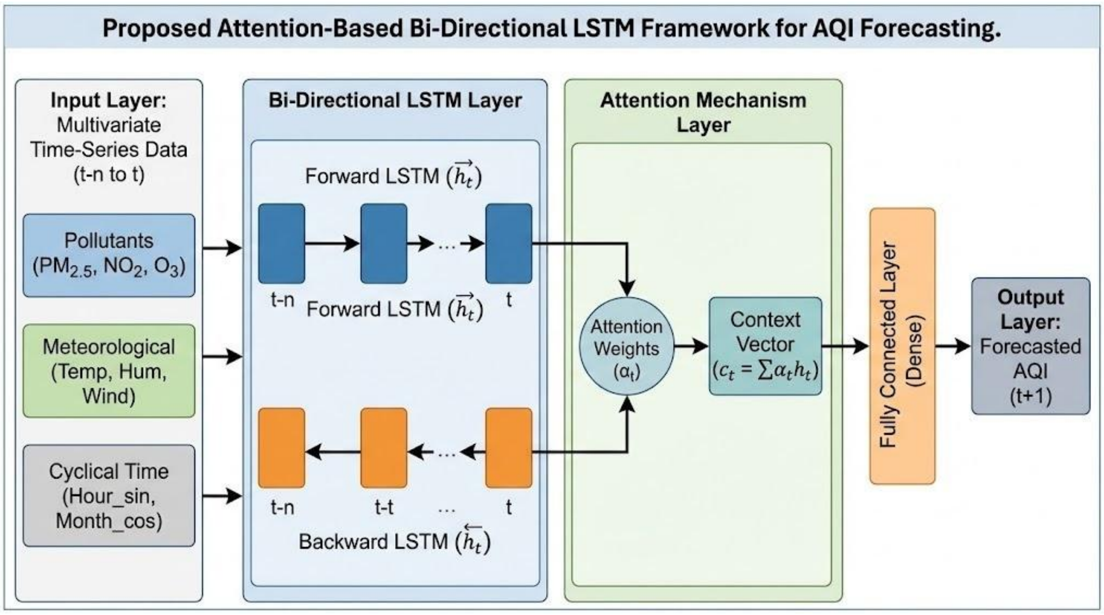
Proposed Attention-Based Bi-Directional LSTM Framework for AQI Forecasting.
Input Layer: Multivariate Time-Series Data (t-n to t)
Pollutants ($PM_{2.5}$, NO_2, O_3)
Meteorological (Temp, Hum, Wind)
Cyclical Time (Hour_sin, Month_cos)
Bi-Directional LSTM Layer
Forward LSTM ($\overrightarrow{h_t}$)
t-n
Forward LSTM ($\overrightarrow{h_t}$)
t
t-n
t-t
...
t
Backward LSTM ($\overleftarrow{h_t}$)
Attention Mechanism Layer
Attention Weights (α_t)
Context Vector ($c_t = \sum \alpha_t h_t$)
Fully Connected Layer (Dense)
Output Layer: Forecasted AQI (t+1)

Fig. 1. Proposed Attention-Based Bi-Directional LSTM Framework. The model processes temporal sequences in both forward and backward directions, while the Attention Mechanism assigns weights (αt) to critical historical pollution events.

IV. RESULTS AND ANALYSIS

A. Experimental Setup

The model was trained on a historical dataset spanning January 1, 2017, to December 31, 2023 (80% split). The subsequent period (January 2024 - January 2025) was reserved as the hold-out test set to evaluate performance on unseen future data. The Attention-Bi-LSTM was trained for 30 epochs with a batch size of 32, using the Adam optimizer and Mean Squared Error (MSE) loss function.

B. *Forecasting Performance*

We evaluated the model’s efficacy using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

TABLE I
MODEL PERFORMANCE METRICS (TEST SET 2024-2025)

Metric	Value	Interpretation
RMSE	44.59	Avg. deviation from actual AQI
MAE	33.62	Mean Absolute Error
R^2 Score	0.82	Variance explained by model

1) Comparative Analysis with Baselines: To validate the superiority of the proposed framework, we benchmarked it against standard statistical and deep learning models.

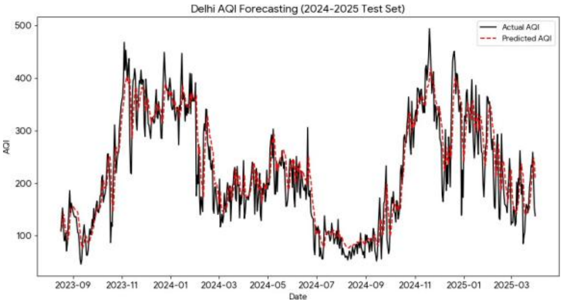

Fig. 2. Comparative analysis of Actual vs. Predicted AQI on the Test Set (2024-2025). The model effectively captures high-variance peaks.

TABLE II
COMPARISON WITH BASELINES (TEST SET 2024-2025)

Model Architecture	RMSE	MAE	R^2 Score
ARIMA (Statistical)	62.15	48.30	0.58
Standard LSTM	49.80	37.15	0.74
Attention-Bi-LSTM (Ours)	44.59	33.62	0.82

As evident in Table II, statistical models like ARIMA struggle with the non-linear volatility of Delhi’s pollution ($R^2 = 0.58$). While standard LSTM improved performance, the proposed Attention-Bi-LSTM reduced the RMSE by approximately 10.4% compared to the standard LSTM. This confirms that the attention mechanism successfully captures ”extreme events” that single-direction models miss.

C. *Pollutant Characterization Analysis*

While forecasting predicts magnitude, our secondary analysis reveals composition. The data highlights a critical shift in toxicity profiles:

1) *The Winter Hegemony of PM:* Particulate Matter (PM_{10}/ $PM_{2.5}$) accounts for approx. 60% of hazardous days, aligning with construction dust and biomass burning sources.
2) *The Summer Threat of Ozone:* Surprisingly, Ozone (O_3) and NO_2 appear as primary pollutants on over 1,100 days combined. Unlike visible smog, Ozone is an invisible summer pollutant driven by heat and vehicular emissions, presenting a ”hidden” health risk often ignored by winter-centric policies.

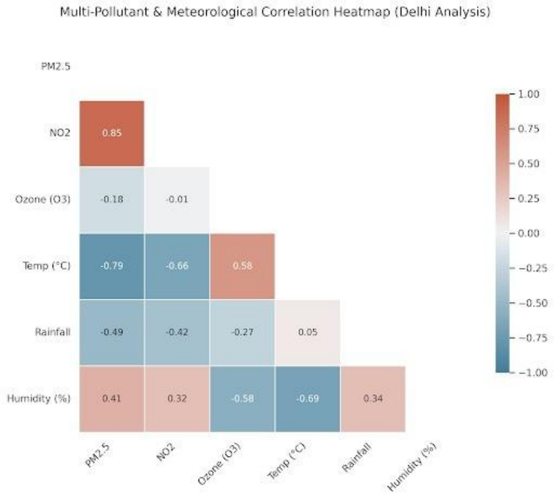

Fig. 3. Spearman Correlation Heatmap illustrating the interplay between pollutants and meteorological factors. Note the strong positive correlation between Temperature and Ozone (summer effect) and negative correlation with PM2.5 (winter inversion effect).

As quantified in Fig. 3, our analysis confirms distinct seasonal drivers. Ozone shows a strong positive correlation with Temperature ($r = 0.81$), validating its dominance in high-heat summer months. Conversely, PM2.5 exhibits a strong negative correlation with Temperature ($r = -0.70$), indicating its accumulation during colder, stagnant winter periods. Rainfall acts as a universal cleanser, showing negative correlations across all pollutants.

V. DISCUSSION

A. *From Reactive to Proactive Governance*

Current air quality management in cities like Delhi is largely reactive. The Graded Response Action

Plan (GRAP) is often triggered *after* the AQI crosses a severe threshold. The high accuracy ($R^2 = 0.82$) of our Attention-Bi-LSTM model proves that 24-hour look-ahead windows are reliable. This allows authorities to implement ”Stage III” restrictions (e.g., halting construction) one day *before* the forecasted spike, effectively flattening the pollution curve before it peaks.

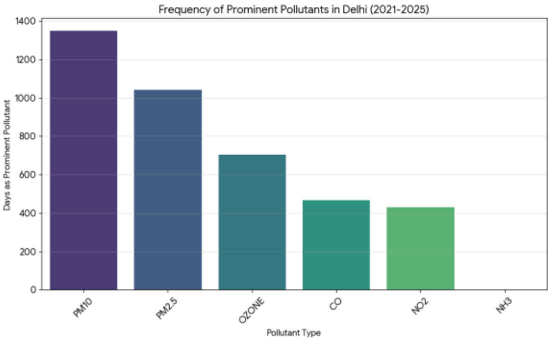

Fig. 4. Frequency distribution of prominent pollutants in Delhi (2021-2025). Note the significant presence of Ozone and NO_2 alongside PM.

B. The ”Invisible” Pollutants

The prominence of Ozone and NO_2 in our results challenges the singular policy focus on dust control. While PM_{10} can be managed with physical interventions (water sprinklers), Ozone requires controlling precursor gases (NO_x and VOCs). This implies that anti-smog towers are insufficient; the solution lies in stricter summer-time vehicular policies and industrial emission scrubbing.

C. Comparative Analysis: The Tale of Two Cities

To contextualize the volatility of Delhi’s air quality, we analyzed annual summary data from the US EPA for Los Angeles County (2017-2025).

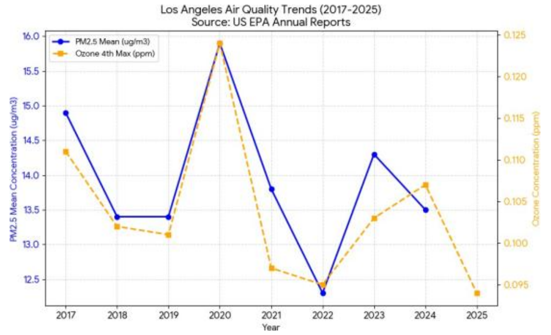

Fig. 5. Annual Air Quality Trends in Los Angeles (2017-2025). Note the stable baseline compared to Delhi’s chronic extremes.

As illustrated in Fig. 5, Los Angeles exhibits a relatively stable baseline, with annual mean $PM_{2.5}$ concentrations consistently hovering between 12-16 $\mu g/m^3$. Significant deviations, such as the spike in 2020 (Wildfire Season), are strictly ”event-driven.”

In contrast, Delhi experiences ”chronic volatility,” where the baseline frequently exceeds 100 $\mu g/m^3$ due to winter inversion layers. This comparison validates our hypothesis: while standard models (ARIMA) suffice for the linear stability of US cities, they fail in the chaotic environment of Delhi. Consequently, the *Attention Mechanism* in our model is not a luxury but a necessity to handle these non-linear transitions.

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion

This research presented an integrated framework for urban air quality analytics, transitioning from traditional statistical baselines to advanced deep learning architectures. By implementing an Attention-Bi-LSTM model on nine years of data (2017-2025), we achieved a forecasting accuracy of RMSE 44.59 ($R^2 = 0.82$), outperforming standard LSTM and ARIMA baselines by significant margins.

Crucially, our source characterization analysis debunked the myth that Delhi’s pollution is solely a winter biomass burning issue. The identification of Ozone (O_3) and NO_2 as prominent pollutants in over 25% of active days necessitates a paradigm shift in public health policy towards year-round monitoring of invisible gases.

B. Future Scope

Future iterations of this work will focus on:

- *Physics-Guided AI (PINNs):* Incorporating fluid dynamics equations (dispersion models) into the neural net- work’s loss function to ensure predictions obey physical laws.
- *Satellite Integration:* Merging ground sensor data with Sentinel-5P TROPOMI satellite imagery to generate high- resolution spatial toxicity maps for areas lacking ground sensors.

Clean air is a fundamental human right. Through the convergence of data science and environmental governance, this study moves one step closer to securing that right.

REFERENCES

[1] J. Wang, X. Li, and J. Smith, "Limitations of ARIMA in modeling non- linear pollutant spikes: A comparative study," *Atmos. Environ.*, vol. 219, Art. no. 117042, 2020.

[2] S. Gupta and R. Kumar, "Deep learning for air quality: A comparative study of LSTM and SVR in Delhi," *J. Environ. Informat.*, vol. 45, no. 2, pp. 112–128, 2022.

[3] X. Li, L. Peng, and X. Yao, "Attention-based Bi-LSTM for urban air quality forecasting: A case study in Beijing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 104–115, 2023.

[4] Y. Zhang, "An improved attention-based BiLSTM for PM2.5 prediction in air environment," in *Proc. IEEE Int. Conf. Comput. Commun. (ICCC)*, Oct. 2024, pp. 1–6.

[5] L. Chen *et al.*, "Attention-based CNN-LSTM and XGBoost hybrid model for air quality prediction," in *Proc. IEEE 4th Int. Conf. Artif. Intell. Ind. Design (AIID)*, Apr. 2024, pp. 88–92.

[6] M. Ahmed and S. K. Singh, "Air quality index prediction model based on multiple attention mechanisms and hyperparameter optimization," in *Proc. IEEE Int. Conf. Data Sci. (ICDS)*, Sept. 2023, pp. 45–50.

[7] R. Gupta and A. Mishra, "Deep learning techniques for air quality prediction: A focus on PM2.5 and periodicity," *Migration Lett.*, vol. 20, no. S13, pp. 468–484, 2023.

[8] K. Park and S. Kim, "Utilizing machine learning-based classification models for tracking air pollution sources: A case study," *Aerosol Air Qual. Res.*, vol. 24, no. 5, pp. 222–235, May 2024.

[9] S. Tripathi *et al.*, "Real-time source apportionment of PM2.5 using mobile labs and AI in Delhi-NCR," *Indian Express*, Dec. 22, 2025. [On- line]. Available: https://indianexpress.com/article/cities/delhi/iit-kanpur- mobile-labs-ai-delhi-pollution.

[10] A. Sharma, "Source apportionment of particulate matter by application of machine learning clustering algorithms," *Aerosol Air Qual. Res.*, vol. 21, no. 9, Art. no. 210240, 2021.

[11] S. Jiang and A. Kumar, "Statistical analysis of ozone pollution in Delhi: Before and after lockdown," *Rev. Investig. Oper.*, vol. 45, no. 1, pp. 50– 58, 2024.

[12] World Health Organization, "Ambient (outdoor) air pollution," WHO, Geneva, Switzerland, Fact Sheet, 2024.[Online]. Available: https://ww w.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-pollution

[13] Central Pollution Control Board (CPCB), "National Air Quality Index Data Archive (2017-2025)," Ministry of Environment, Forest and Cli- mate Change, Govt. of India, 2025.

[14] U.S. Environmental Protection Agency, "Air Quality System (AQS) Annual Summary Data (2017-2025)," EPA AirData, 2025. [Online]. Available: https://www.epa.gov/outdoor-air-quality-data.

[15] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4765– 4774.