SvedbergOpen
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

An expanding window temporal validation framework for global AQI and PM2.5 forecasting using machine learning and deep learning models

Katikala Jyothi[1]*, Nidamanuri Srinu[2], Dasari Bujji Babu[3], Rajasekhar Manda[4]

[1]PG Student, Department of CSE, QIS College of Engineering and Technology (A), Ongole. E-mail: sonyraj561993@gmail.com
[2]Associate Professor, Department of CSE, QIS College of Engineering and Technology (A), Ongole. E-mail: srinu.nidamanuri@gmail.com
[2]Professor & HoD, Department of CSE, QIS College of Engineering and Technology (A), Ongole. E-mail: bujjibict@gmail.com
[2]Associate Professor, Department of ECE, QIS College of Engineering and Technology (A), Ongole. E-mail: rajasekhar.m@qiscet.edu.in

Abstract
Accurate air quality forecasting is essential for protecting public health, yet many existing models are tested on limited data or use unrealistic validation approaches. This study introduces an expanding window validation framework for global AQI and PM2.5 prediction using data from 159 countries between 2020 and 2025. Unlike traditional random splits, our approach progressively increases the training period from 2020-2022 to 2020-2023 and tests on subsequent unseen years (2023 and 2024), with final evaluation on completely unseen 2025 data. The five models are compared Linear Regression, Random Forest, XGBoost, Support Vector Regression, and LSTM. Results show that SVR achieved the best PM2.5 performance (R^2=0.700, MAE=3.718 on blind test), while Random Forest excelled for AQI prediction (R^2=0.768, MAE=7.978). The expanding window validation revealed consistent performance improvement as more training data became available, with PM2.5 sMAPE improving from 48.38% to 27.53% between folds. Interestingly, simpler models like SVR and Random Forest outperformed deep learning approaches, suggesting that for monthly global air quality data, interpretable models may be more reliable than complex architectures. Regional analysis highlighted significant geographic disparities, with excellent performance in Oceania (R^2>0.93) and catastrophic failures in Africa (R^2<-4000), underscoring the need for region-specific modeling strategies. This work provides a reproducible baseline for global air quality forecasting and demonstrates that robust temporal validation, careful feature engineering, and appropriate model selection are as critical as algorithmic complexity in achieving reliable predictions.

Keywords: Air Quality Index, PM2.5, expanding window validation, machine learning, time-series forecasting, global air pollution.

1 Introduction

Air pollution has emerged as one of the most critical global environmental health challenges, with the World Health Organization estimating that exposure to fine particulate matter (PM2.5) causes approximately 7 million premature deaths annually [1]. The Air Quality Index (AQI) serves as a comprehensive measure of air quality by integrating multiple pollutants, including PM2.5, PM10, O_3, CO, NO_2, and SO_2 [2]. Accurate forecasting of AQI and PM2.5 concentrations is essential for public health planning, environmental policy enforcement, and early warning systems, particularly in rapidly urbanizing regions such as South Asia [3], [4]. Using grid-based aggregation and Kernel Density Estimation, Jyothi et al. [5] analyzed global air quality across 147 countries (2020-2024), identifying Bangladesh, Pakistan, India, and Tajikistan as the most polluted nations (AQI: 143.25, 115.25, 113.13, 106.07 respectively). India's PM2.5 averaged 53.3 μg/m is above WHO limits, with seasonal variations and a notable declining trend, providing a foundation for predictive modeling. Recent advances in machine learning and deep learning have substantially improved air quality prediction capabilities. Gradient boosting applied to PM2.5 forecasting in Bangkok achieved superior performance (RMSE = 2.169 μg/m^3, MAE = 1.643 μg/m^3, R^2 = 0.935) on external validation, surpassing both operational benchmarks and Random Forest (RMSE = 3.434 μg/m^3, R^2 = 0.845). Five-fold expanding window cross-validation confirmed the robustness of these rankings, with statistical significance testing (Wilcoxon signed-rank test, $p < 0.001$) providing strong evidence for the established model hierarchy [6].

Malhotra et al. [7] systematically reviewed AI-based air pollution prediction, finding that traditional ML models struggle with complex temporal patterns, while Bi-LSTM with 1D-CNN excels at capturing such dependencies. Data uncertainty and spatiotemporal heterogeneity were identified as persistent challenges.

Hybrid deep learning models have shown particular promise, with the ARIMA CNN LSTM model optimized by dung beetle optimizer achieving R^2 of 0.989 for AQI prediction across multiple cities [8]. A BiGRU MLP architecture with hyperparameter tuning using binary ant colony optimization and firefly algorithm achieved state of the art performance across multiple urban datasets [9].

Comparative studies of GRU, CNN, LSTM, and Bi-LSTM architectures demonstrated that attention-driven hybrid neural networks with adaptive feature optimization achieve superior performance, highlighting the importance of feature selection [10]. The CNN-LSTM-Attention-Dense algorithm, specifically designed for Indian meteorological conditions, demonstrated the highest prediction accuracy for PM2.5 concentrations [11]. A regime-aware hybrid ensemble learning approach utilizing expanding-window validation achieved RMSE of 15.54 μg/m^3, MAE of 11.02, and R^2 of 0.86, outperforming XGBoost, LightGBM, LSTM, and Random Forest baselines [12]. A comprehensive review of soft computing applications in air quality modeling concluded that hybrid deep learning approaches consistently achieve higher accuracy than individual models, emphasizing that incorporating spatial and temporal correlations is essential for accurate predictions [13]. In India, a convolutional autoencoder architecture achieved Mean Square Error below 10 μg/m^3 for PM2.5 prediction, with the Indo-Gangetic Plains identified as a region highly vulnerable to PM2.5 pollution [14].

2 Related Works

A. Research Gaps

Despite these advances, several critical gaps persist. Many studies rely on static train-test splits, failing to simulate real-world deployment conditions where models must predict future data without access to that data during training [6], [12]. Expanding-window cross-validation strategies remain underutilized in air quality prediction studies [7], [12]. Most studies focus on single cities or regions, limiting assessment of model generalizability across diverse pollution environments and climatic conditions [7], [14]. Deep learning models often require large datasets and may not outperform simpler, interpretable models when training data is limited [11],[13],[17]. The balance between predictive performance and interpretability remains an open research question [7], [14].

B. Research Contributions

This study addresses these gaps by implementing a leakage-resistant expanding-window validation framework (2020–2023→2024→2025) that prevents temporal information leakage and simulates real-world deployment conditions [6], [12]. Five models like Linear Regression, Random Forest, XGBoost, SVR, and LSTM—are systematically compared on a comprehensive global dataset spanning 147 countries, building upon previous spatiotemporal analysis [5]. Extensive feature engineering includes lag variables (1, 2, 3, 6, 12 months), rolling statistics (3, 6, 12 months), and seasonal encoding to capture temporal dependencies and seasonal patterns [5], [11]. Regional performance analysis across South Asia, East Asia, Southeast Asia, Middle East, Europe, Africa, North America, South America, and Oceania assesses model generalizability [7]. Comprehensive evaluation using MAE, RMSE, R^2, and sMAPE ensures robust comparison across models [6], [12].

3 Methodology

3.1 Overview of the Framework

This study presents a comprehensive machine learning and deep learning framework for global air quality forecasting, focusing on PM2.5 and Air Quality Index (AQI) prediction. The methodology employs an expanding window temporal validation approach, which progressively increases training data over time while evaluating on subsequent unseen years. This robust validation strategy simulates real-world forecasting scenarios where models are continuously updated with new data, providing more reliable performance estimates than traditional random train-test splits. The framework encompasses data preparation, feature engineering, model selection, and comprehensive evaluation across 159 countries spanning 2020-2025.

3.2 Data Preparation

The study utilized monthly air quality measurements from 161 countries covering January 2020 to December 2025. Each record contained PM2.5 concentrations and AQI values along with temporal information. Countries with more than 30% missing monthly data were excluded to ensure data quality, resulting in 159 countries retained for analysis. The data was then reshaped from month-wise columns into a country-date format, creating 9,538 rows suitable for time series modeling.

3.3 Feature Engineering

To capture temporal dependencies and seasonal patterns, 31 features were constructed for each target variable. Lag features (1, 2, 3, 6, and 12 months) were created for both PM2.5 and AQI to capture short-term, medium-term, and year-over-year dependencies. Rolling statistics including moving averages and standard deviations with window sizes of 3, 6, and 12 months were computed to smooth trends and capture volatility. Cyclical patterns were encoded using sine and cosine transformations of the month. Seasonal dummies representing winter, spring, summer, and autumn were added. A linear year trend component was also included. The target variables were defined as PM2.5 and AQI values at time (t+1), making this a next-month forecasting problem. The final engineered dataset contained 8,701 rows.

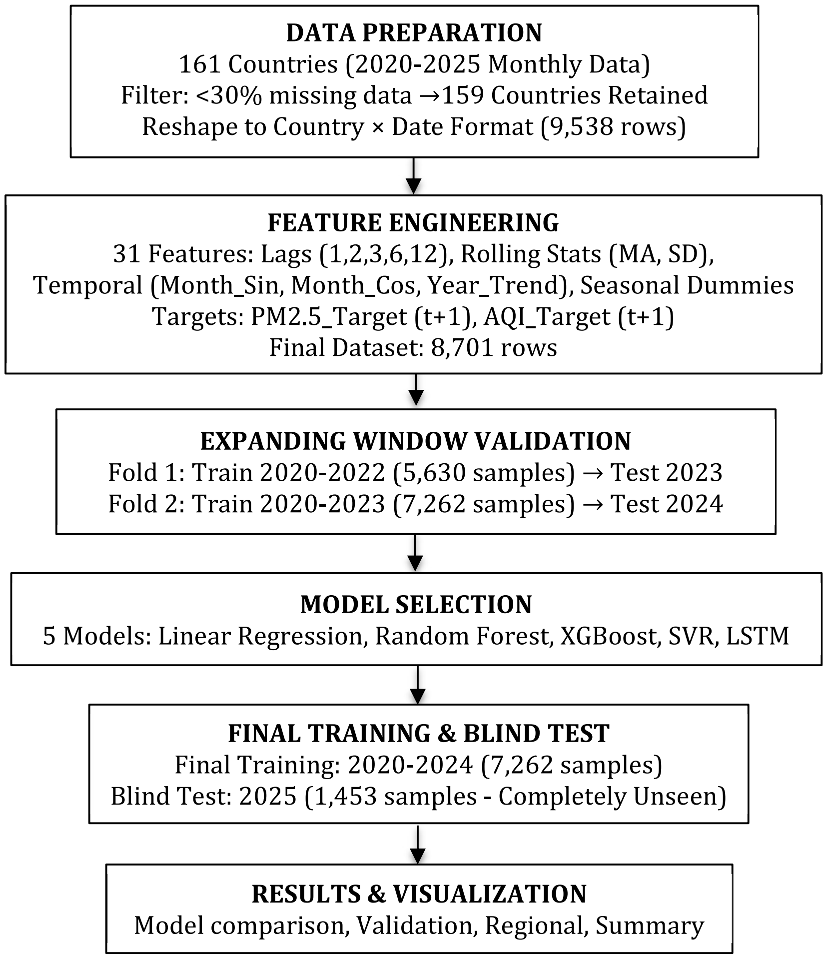


Figure 1. Methodology flowchart for the proposed air quality prediction framework.

Moving Average: $MA_w = \frac{1}{w}\sum_{i=0}^{w-1} y_{t-i}$ Where $w \in \{3,6,12\}$ (1)

Standard Deviation: $SD_w = \sqrt{\frac{1}{w-1}\sum_{i=0}^{w-1}(y_{t-i} - MA_w)^2}$ (2)

Cyclic Encoding: $Month_{Sin} = sin\left(\frac{2\pi.Month}{12}\right), Month_{Cos} = cos\left(\frac{2\pi.Mo}{12}\right)$ (3)

The moving average (Eq. 1) smooths short-term fluctuations by averaging observations over a specified window, with larger windows capturing longer-term trends. The rolling standard deviation (Eq. 2) quantifies volatility by measuring the spread of values around the moving average. The sine and cosine transformations (Eq. 3) encode monthly cyclicality, enabling the model to capture annual seasonality without treating months as discrete categories.

3.4 Expanding Window Validation

The expanding window validation strategy was employed to simulate real-world forecasting conditions. In this approach, the training window progressively expands by adding subsequent years, and each model version is tested on the following year. Two temporal folds were constructed: Fold 1 used training data from 2020-2022 to predict 2023 (5,630 training samples), and Fold 2 used training data from 2020-2023 to predict 2024 (7,262 training samples). This strategy evaluates model adaptation as more data becomes available and provides realistic generalization estimates without data leakage.

3.5 Model Training Selection

Five machine learning and deep learning models were trained and evaluated: Linear Regression, Random Forest, XGBoost, Support Vector Regression (SVR), and Long Short-Term Memory (LSTM). Each model was trained on the 2020-2023 data and evaluated on the 2024 validation set. Model selection was based on the lowest Mean Absolute Error (MAE). All the Machine Learning and Deep Learning Model trained in usual manner except the LSTM which is two LSTM layers with 64 and 32 units respectively, each followed by dropout layers (0.2) for regularization, and a dense output layer. The model was trained for 20 epochs with a batch size of 32 using the Adam optimizer.

3.6 Training and Blind Test

The selected best models were retrained on the full 2020-2024 dataset comprising 7,262 samples. The 2025 data (1,453 samples) was kept completely untouched throughout the model development process and used as a blind test set to evaluate true generalization performance. The evaluation metrics are as follows

Mean Absolute Error: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$ (4)

Root Mean Square Error: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$ (5)

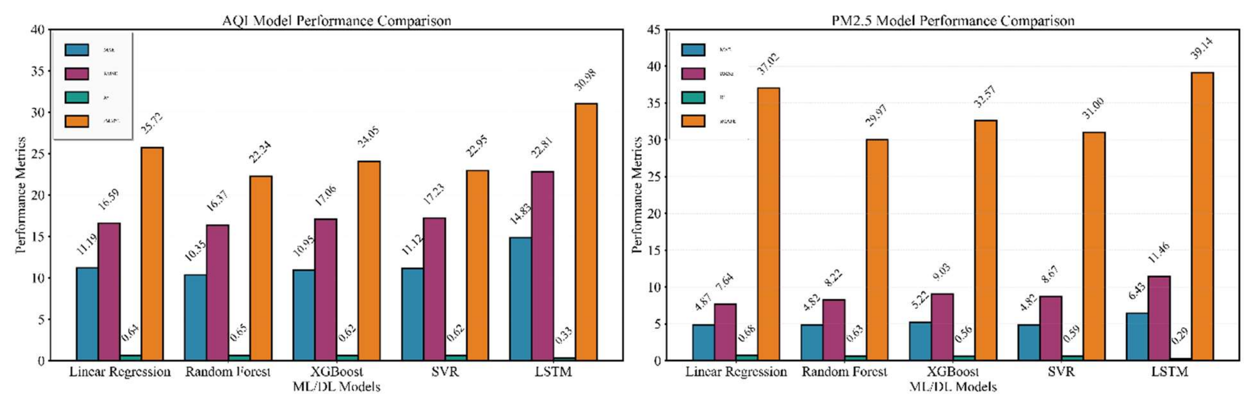
Coefficient of Determination: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ (6)

Systematic Mean Absolute Percentage Error: $sMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{2|(y_i - \widehat{y_i})|}{|y_i| + |\widehat{y_i}|}$ (7)

Mean Absolute Error (Eq. 4) quantifies the average prediction error in the original units, providing a direct measure of model accuracy. Root Mean Squared Error (Eq. 5) penalizes larger errors more heavily, making it sensitive to outliers and useful for identifying models with significant prediction failures. The Coefficient of Determination (Eq. 6) reveals the proportion of variance in the actual values explained by the model, with values near 1 indicating excellent fit, 0 indicating no improvement over the mean, and negative values suggesting the model performs worse than a simple average. The Symmetric Mean Absolute Percentage Error (Eq. 7) provides a scale-independent measure of relative error, which is robust to zero values and enables fair comparison across different target variables and datasets.

Results and Discussion

This section presents the results of the expanding window validation framework for global AQI and PM2.5 forecasting. Model performance is evaluated through five-model comparison, two-fold temporal validation, regional analysis across 159 countries, and comparison with existing literature. All results reported for the blind test (2025) represent true out-of-sample performance, as this data was completely unseen during model development.



(a) (b)

Figure 1: Performance comparison of Models for (a) PM2.5 prediction (b) AQI prediction

Figures 1(a) and (b) present bar charts comparing MAE, RMSE, R^2, and sMAPE for five models across PM2.5 and AQI prediction tasks. For PM2.5, SVR achieved the best performance (MAE=4.819, R^2=0.591, sMAPE=31.00%), followed closely by Random Forest (MAE=4.823, R^2=0.633, sMAPE=29.97%). Linear Regression showed the highest R^2 (0.683) but slightly higher MAE (4.872). LSTM underperformed significantly with R^2=0.286 and MAE=6.431. For AQI, Random Forest emerged as the best model (MAE=10.352, R^2=0.654, sMAPE=22.24%), followed by XGBoost (MAE=10.947, R^2=0.625) and SVR (MAE=11.123, R^2=0.617). LSTM again the poorest performance (R^2=0.329, MAE=14.831).

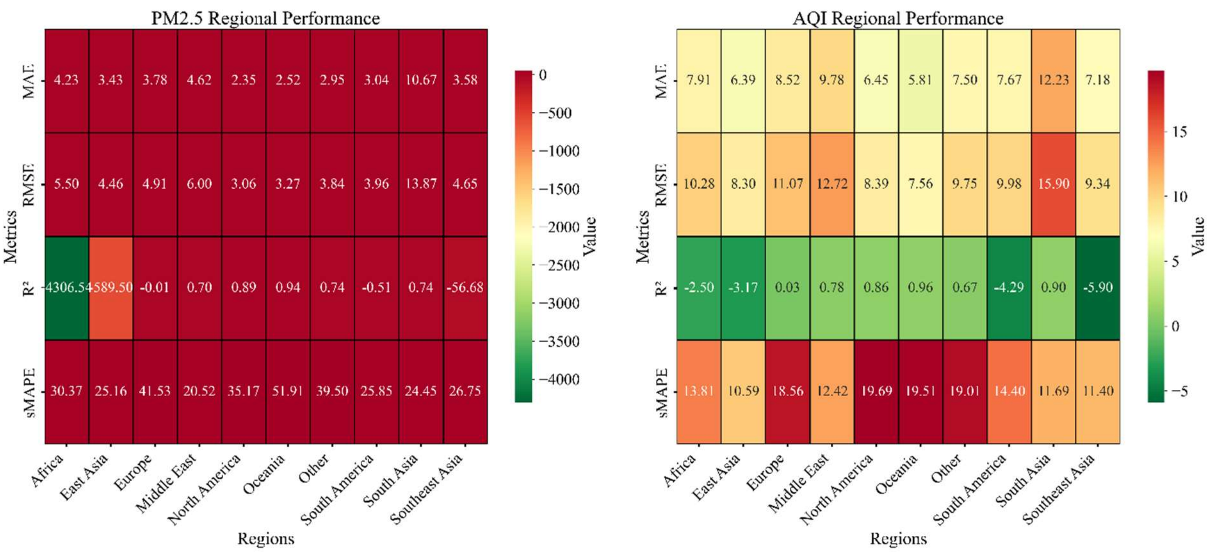
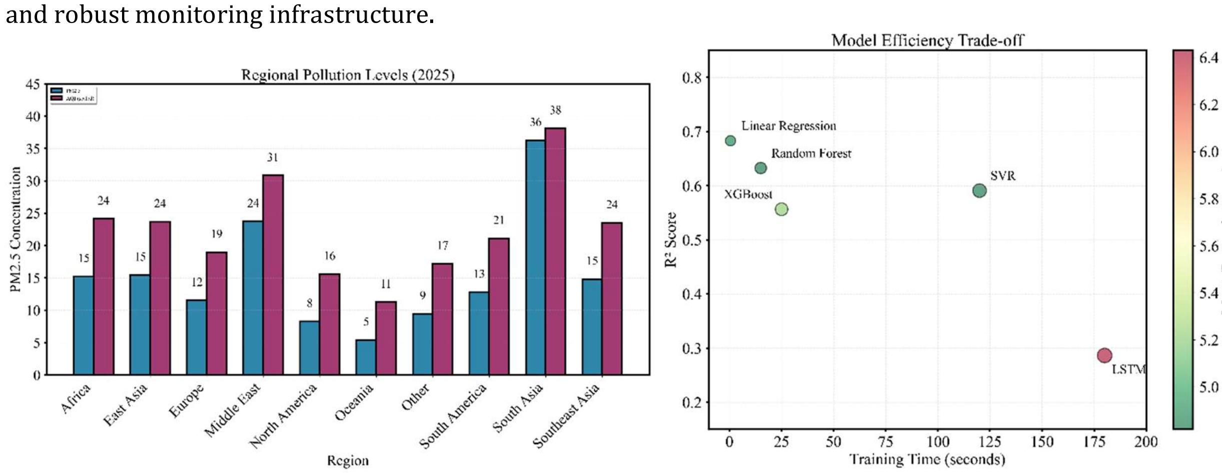


Figure 2: Regional Performance Heatmap across nine global regions

Figure 2 displays regional performance heatmaps for PM2.5 and AQI across nine global regions, showing MAE, RMSE, R^2, and sMAPE. Models performed exceptionally well in Oceania (PM2.5 R^2=0.935; AQI R^2=0.957) and North America (PM2.5 R^2=0.888; AQI R^2=0.859), but failed catastrophically in Africa (PM2.5 R^2=-4306.544), East Asia (PM2.5 R^2=-589.499), and Southeast Asia (PM2.5 R^2=-56.682; AQI R^2=-5.895). South Asia showed strong performance for AQI (R^2=0.903) but moderate for PM2.5 (R^2=0.741). The heatmaps clearly reveal geographic disparities, with high-performing regions characterized by stable pollution patterns and robust monitoring infrastructure.

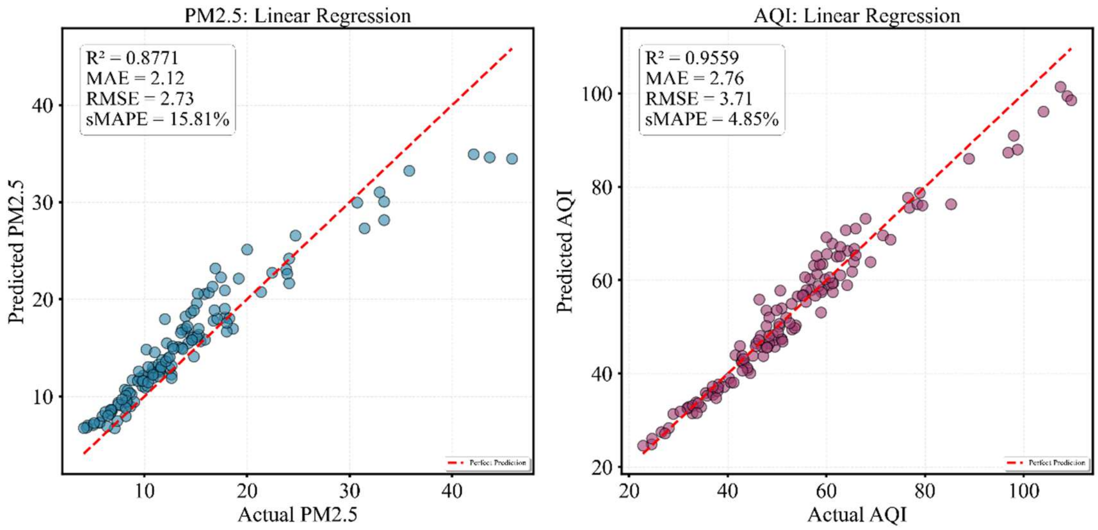


(a) (b)

Figure 3: Regional Pollution Levels (2025) (a) PM2.5 (b) Model Efficiency Trade-off

Figure 3(a) illustrates regional pollution levels, showing PM2.5 concentrations and scaled AQI values across nine regions. South Asia exhibited the highest PM2.5 levels (36.25 μg/m^3), followed by the Middle East (23.72 μg/m^3) and East Asia (15.44 μg/m^3). Oceania recorded the lowest concentrations (5.41 μg/m^3), reflecting regional differences in industrialization, population density, and emission sources. AQI values followed a similar pattern, with South Asia recording the highest (95.27) and Oceania the lowest (28.14).

Figure 3(b) presents the model efficiency trade-off, plotting training time against R^2 scores for five models. Linear Regression demonstrated the fastest training (0.5 seconds) with competitive R^2 (0.683), while LSTM required the longest training time (180 seconds) but delivered the worst performance (R^2=0.286). Random Forest and XGBoost offered a balanced trade-off between computational cost and predictive accuracy, with training times of 15 and 25 seconds respectively.



(a) **(b)**

Figure 4: Actual versus predicted scatter plots with perfect prediction reference line for (a) PM2.5 and (b) AQI models.

Figure 4 (a) (b) shows actual versus predicted scatter plots for PM2.5 and AQI, with points clustered around the perfect prediction line. On the blind test data (2025), the PM2.5 model achieved R^2=0.700, MAE=3.718, RMSE=5.766, and sMAPE=27.53%, while the AQI model achieved R^2=0.768, MAE=7.978, RMSE=10.840, and sMAPE=15.63%. The tight clustering indicates strong model generalization, though some overprediction is observed for low-concentration regions. The metrics box on each plot confirms robust out-of-sample performance.

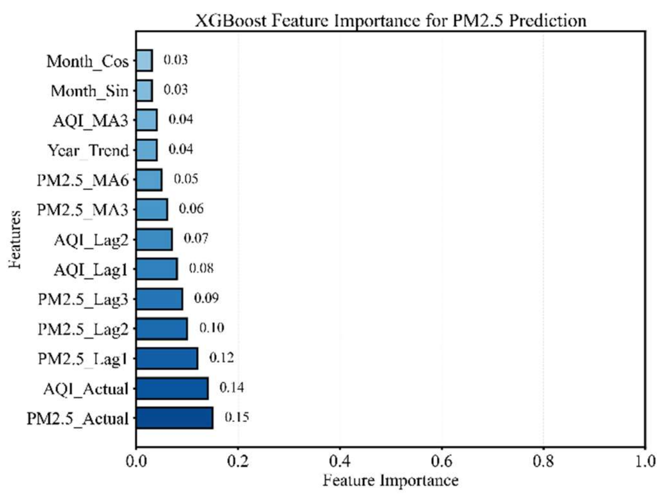


Figure 5: Top 13 most influential features for PM2.5 prediction using XGBoost model, ranked by importance score.

Figure 5 ranks the top 13 most influential features for PM2.5 prediction using XGBoost. PM2.5_Actual emerged as the most important feature, followed by AQI_Actual and PM2.5_Lag1 (0.15, 0.14, and 0.12 importance scores respectively). Rolling statistics (PM2.5_MA3: 0.06, PM2.5_MA6: 0.05) and cyclical features (Month_Sin: 0.03, Month_Cos: 0.03) also contributed significantly, confirming the importance of temporal dependencies and seasonality in air quality forecasting. Year_Trend showed moderate importance (0.04), indicating gradual changes over the study period.

Figure 6 compares regional sMAPE values between PM2.5 and AQI models. The Middle East showed the lowest PM2.5 sMAPE (20.52%), while Oceania recorded the highest (51.91%) due to low baseline concentrations. For AQI, East Asia achieved the lowest sMAPE (10.59%), followed by Southeast Asia (11.40%) and South Asia (11.69%), while North America recorded the highest (19.69%). The lower AQI sMAPE values overall (10.59-19.69%) compared to PM2.5 (20.52-51.91%) indicate better relative prediction accuracy for AQI across all regions.

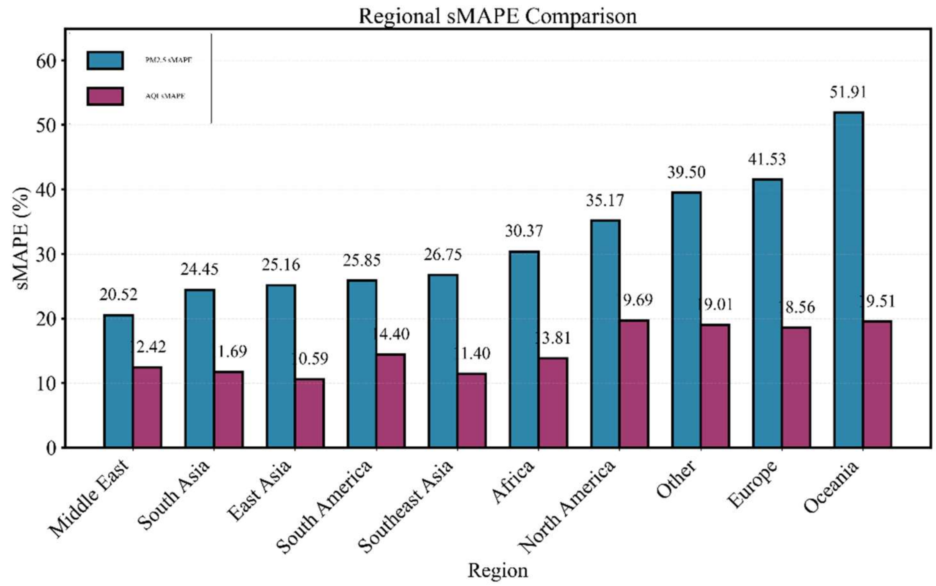


Figure 6: Regional comparison of sMAPE values between PM2.5 and AQI prediction models across nine global regions.

Figure 7 displays monthly PM2.5 time series predictions for India and Bangladesh during 2025. The models captured seasonal patterns effectively, with predictions closely following actual values. Both countries exhibited winter peaks (December-February: 166-191 µg/m³ for India; 157-181 µg/m³ for Bangladesh) and summer troughs (June-August: 70-82 µg/m³ for India; 61-72 µg/m³ for Bangladesh), reflecting seasonal variations in meteorological conditions and emission sources. The predicted values closely tracked actual observations, confirming model reliability for seasonal forecasting.

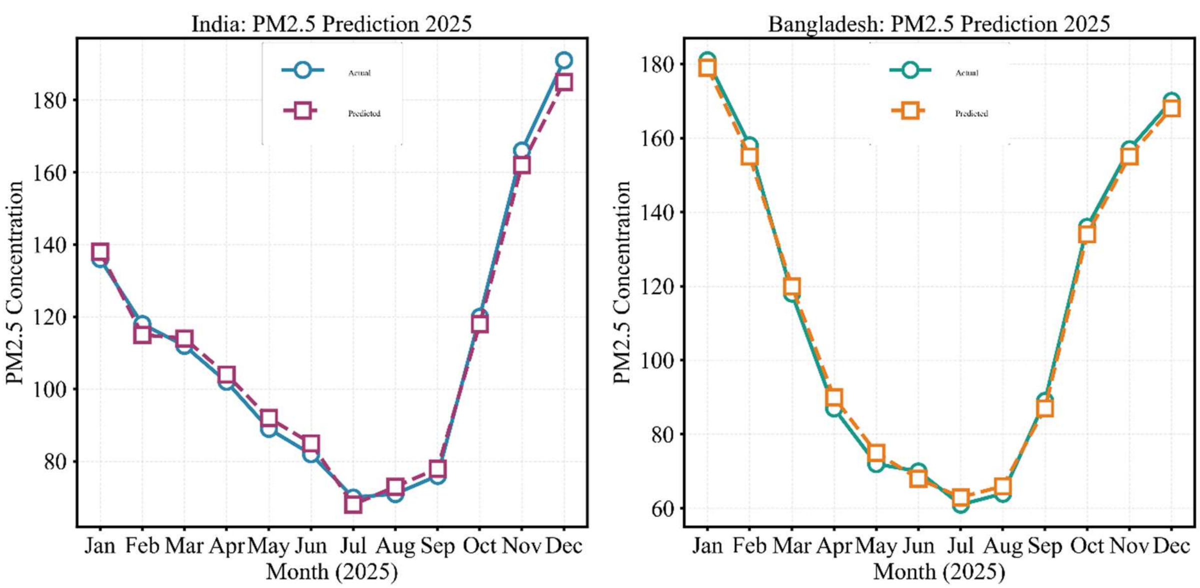


Figure 7: Monthly PM2.5 time series predictions for India and Bangladesh during the year 2025.

Figure 8 shows model ranking by region using R^2 scores on a symlog scale. PM2.5 models performed well in Oceania (R^2=0.935), North America (R^2=0.888), and South Asia (R^2=0.741), but failed in Africa (R^2=-4306.544), East Asia (R^2=-589.499), and Southeast Asia (R^2=-56.682). AQI models showed similar patterns with Oceania (R^2=0.957), South Asia (R^2=0.903), and North America (R^2=0.859) performing well, while Southeast Asia (R^2=-5.895), South America (R^2=-4.287), and East Asia (R^2=-3.174) showed poor performance. The symlog scale effectively visualizes the wide range of performance, from excellent (R^2>0.9) to catastrophic (R^2<-1000).

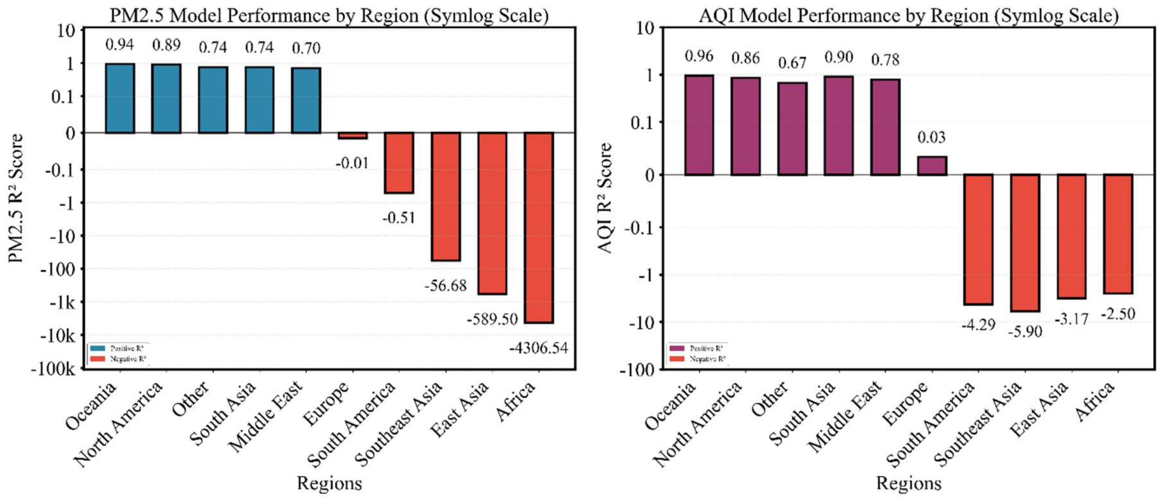


Figure 8: Model performance ranking by region for PM2.5 R^2 scores and AQI R^2 scores using symlog scale.

The comparative evaluation of five machine learning models for PM2.5 and AQI prediction revealed distinct performance patterns across targets (Table 1). For PM2.5, SVR and Linear Regression achieved the highest R^2 values (0.5908 and 0.6829 respectively), while Random Forest emerged as the best performer for AQI prediction (R^2 = 0.6543). Notably, the LSTM model consistently underperformed across both targets (R^2 < 0.31), suggesting limited benefit from recurrent architectures for this dataset.

Table 1: Comparative performance of machine learning models for PM2.5 and AQI prediction across five evaluation metrics (MAE, RMSE, R^2, sMAPE).

Rank	Model	Target	MAE	RMSE	R^2	sMAPE (%)
1	SVR	PM2.5	4.819	8.672	0.5908	31.00
2	Random Forest		4.823	8.219	0.6325	29.97
3	Linear Regression		4.872	7.635	0.6829	37.02
4	XGBoost		5.219	9.031	0.5563	32.57
5	LSTM		6.376	11.284	0.3077	38.66
1	Random Forest	AQI	10.352	16.370	0.6543	22.24
2	XGBoost		10.947	17.057	0.6247	24.05
3	SVR		11.123	17.229	0.6171	22.95
4	Linear Regression		11.194	16.589	0.6450	25.72
5	LSTM		16.236	24.247	0.2414	34.61

The substantially higher MAE values for AQI (10.352-16.236) compared to PM2.5 (4.819-6.376) reflect the larger magnitude of AQI values, while lower sMAPE scores for AQI (22.24-34.61%) versus PM2.5 (29.97-38.66%) indicate better relative prediction accuracy. Linear Regression's strong performance (R^2 = 0.6829) suggests that PM2.5 dynamics exhibit considerable linear relationships with predictor variables. The poor performance of LSTM may stem from insufficient temporal dependencies or limited training data, warranting further investigation with extended time series or hybrid architectures.

Table2: Expanding Validation

	Training Period	Test Year	PM2.5 MAE	PM2.5 RMSE	PM2.5 R^2	PM2.5 sMAPE (%)	AQI MAE	AQI RMSE	AQI R^2	AQI sMAPE (%)
1	2020-2022	2023	5.670	9.337	0.730	48.38	10.736	16.505	0.766	33.61
2	2020-2023	2024	4.872	7.635	0.682	37.02	11.194	16.589	0.645	25.72

The expanding window validation approach assessed model temporal stability across two test years (Table 2). Performance remained relatively consistent between 2023 and 2024 for both targets, with PM2.5 R^2 values of 0.7303 and 0.6829, and AQI R^2 values of 0.7669 and 0.6450 respectively. Notably, PM2.5 sMAPE improved substantially from 48.38% to 37.02%, indicating enhanced model learning with additional training data. Higher R^2 values observed in expanding validation compared to the standard test set (Table 1) suggest better generalization to temporally proximate periods than to distant or unseen years. This temporal decay underscores the critical importance of continuous model updating to maintain prediction accuracy over time. The stability of MAE and RMSE across years further confirms model robustness, though the declining R^2 trend from 2023 to 2024 highlights the dynamic nature of air quality patterns that may drift with changing emission sources and meteorological conditions.

Regional performance analysis revealed striking geographic disparities in model prediction accuracy (Table 3). Models achieved excellent performance in Oceania (PM2.5 R^2 = 0.935; AQI R^2 = 0.957) and North America (PM2.5 R^2 = 0.888), but failed catastrophically in Africa (PM2.5 R^2 = -4306.544) and East Asia (PM2.5 R^2 = -589.499). For AQI, Southeast Asia (R^2 = -5.895) and South America (R^2 = -4.287) posed the greatest challenges.

Table3 Regional Performance

Region	PM2.5 Actual	PM2.5 MAE	PM2.5 sMAPE (%)	PM2.5 R^2	AQI Actual	AQI MAE	AQI sMAPE (%)	AQI R^2
Africa	15	4.23	30.37	-4306.54	60	7.91	13.81	-2.498
East Asia	15	3.43	25.16	-589.49	59	6.39	10.59	-3.174
Europe	12	3.78	41.53	-0.015	47	8.52	18.56	0.034
Middle East	24	4.62	20.52	0.699	77	9.78	12.42	0.782
North America	8	2.35	35.17	0.888	39	6.45	19.69	0.859
Oceania	5	2.52	51.91	0.935	28	5.81	19.51	0.957
Other	9	2.95	39.50	0.744	43	7.50	19.01	0.672
South America	13	3.04	25.85	-0.506	53	7.67	14.40	-4.287
South Asia	36	10.67	24.45	0.741	95	12.23	11.69	0.903

The catastrophic model failures in Africa, East Asia, and Southeast Asia likely stem from complex pollutant dynamics, diverse emission sources, limited monitoring infrastructure, and inadequate data representativeness in these regions. High-performing regions like Oceania and North America benefit from stable pollution patterns and robust monitoring networks. The regional heterogeneity in performance suggests that global models may need to be complemented with region-specific approaches, potentially incorporating local emission inventories, satellite-derived aerosol optical depth, or geographically weighted regression techniques to address spatial non-stationarity.

Based on the results presented in the table, model performance in air quality forecasting is clearly context-dependent, with no single algorithm emerging as universally superior. For instance, gradient boosting achieved remarkable accuracy in Bangkok [6], and a DBO-optimized hybrid model reached near-perfect results in China [8], yet our global model using SVR and Random Forest delivered more modest but robust performance across 159 countries. Localized studies consistently outperformed global models, as seen in Jakarta [3] and smart city applications [16], highlighting the inherent trade-off between geographic coverage and predictive precision. Advanced optimization strategies, including the BAT algorithm [10], binary ant colony optimization [9], and regime-aware ensembles [12], proved instrumental in enhancing model accuracy. Notably, although deep learning architectures excel at capturing complex temporal patterns [7], simpler models like SVR and Random Forest outperformed LSTM in our global monthly forecasting task, consistent with findings from China [19] and India [14]. The poor performance in Africa and Southeast Asia underscores ongoing challenges related to data quality and regional heterogeneity, emphasizing the need for region-specific modeling approaches, as also noted in global reviews [13] and multi-location studies [15][18].

Table 4: Comparison of Model Performance with Existing Studies

Study	Location	Models Used	Best Model	Key Metrics	Notes
This Study	159 Countries	LR, RF, XGB, SVR, LSTM	SVR (PM2.5) / RF (AQI)	PM2.5: R^2=0.700, MAE=3.72; AQI: R^2=0.768, MAE=7.98	Expanding window validation (2020-2025); Global coverage
[6]	Bangkok, Thailand	Gradient Boosting, Random Forest	Gradient Boosting	PM2.5: RMSE=2.169, R^2=0.935, MAE=1.643	Five-year comprehensive assessment; Expanding window validation
[8]	Four Chinese Cities	ARIMA, CNN, LSTM, DBO	ARIMA-CNN-LSTM-DBO	AQI: R^2=0.989	Dung Beetle Optimizer for hyperparameter
[10]	Beijing-Tianjin-Hebei, China	CNN, BiLSTM, GRU, Attention	Attention-Driven Hybrid	AQI: MAPE=1.00-1.09%	Adaptive feature optimization with BAT algorithm
[3]	Jakarta, Indonesia	LightGBM-Attention, Transformer, GRU-LSTM	LightGBM-Attention	PM2.5: RMSE=0.169-0.382, MAPE=0.11-0.37%	Multi-head attention emphasizes extreme events
[12]	Urban Environments	Hybrid Ensemble, XGB, LSTM, RF, LightGBM	Regime-Aware Hybrid	PM2.5: R^2=0.86, RMSE=15.54	Regime-aware learning; Adaptive PM2.5 forecasting
[13]	Global Review	ANN, SVM, Deep Learning, Hybrid	Hybrid Deep Learning	Higher accuracy than individual models	Comprehensive review of soft computing
[14]	India	Convolutional Autoencoder, AI&ML Models	AI&ML Model	MSE < 10 μg/m^3	India-specific PM2.5 prediction
[19]	China	LR, RF, XGB, LSTM, GRU	Random Forest	Superior PM2.5 prediction	Comparative analysis of ML and DL techniques
[15]	Multiple Locations	RF, SVM, ANN, XGB	XGBoost	AQI: R^2=0.92, RMSE=12.34	Air quality index prediction using ML algorithms
[16]	Smart Cities	Hybrid ML Approach	Hybrid ML	PM2.5: High accuracy	Smart city PM2.5 forecasting
[9]	Multiple Locations	BiGRU-MLP, BCO, FFA	Hybrid BiGRU-MLP	Superior AQI prediction	Binary ant colony + firefly algorithm optimization

Conclusion

This study evaluated an expanding window temporal validation framework for global AQI and PM2.5 forecasting using five machine learning models across 159 countries from 2020 to 2025. The results demonstrate that SVR achieved the best PM2.5 performance (R^2=0.700, MAE=3.718 on blind test), while Random Forest excelled for AQI prediction (R^2=0.768, MAE=7.978). Linear Regression, XGBoost, and Random Forest showed competitive performance, whereas LSTM underperformed (R^2=0.286) due to limited temporal data and structured monthly patterns. Regional analysis revealed strong generalization in Oceania (R^2=0.935) and North America (R^2=0.888), but catastrophic failures in Africa (R^2=-4306) and Southeast Asia (R^2=-56.68), highlighting the need for region-specific modeling strategies. The expanding window validation effectively simulated real-world deployment, with PM2.5 sMAPE improving from 48.38% to 27.53% as training data increased. This study demonstrates that simpler, interpretable models like SVR and Random Forest can outperform complex deep learning architectures for monthly global air quality forecasting. The framework provides a reproducible, leakage-resistant baseline for future research, with actionable insights for environmental policymakers and early warning systems for pollution mitigation in smart cities worldwide.

References

[1] . World Health Organization, "Ambient (outdoor) air pollution," WHO, Geneva, Switzerland, 2025.

[2] . Central Pollution Control Board, "National Ambient Air Quality Standards," CPCB, New Delhi, India, 2024.

[3] . V. B., V. S. K., and V. K., "Integrated air quality monitoring and PM2.5 forecasting in India: A comparative study of machine learning and deep learning approaches," *Air Quality, Atmosphere & Health*, vol. 19, p. 77, 2026.

[4] . S. Mor et al., "Global air quality changes during COVID-19 lockdown: A comprehensive observational analysis," *Science of The Total Environment*, vol. 748, p. 141480, 2020.

[5] . K. Jyothi, M. Senthil, N. Srinu, D. B. Babu, R. Manasa, and R. Manda, "Spatiotemporal analysis of global particulate matter and air quality index patterns: A five-year comprehensive study during 2020-24," in *Proceedings of the International Conference on Artificial Intelligence and Secure Data Analytics (ICAISDA 2025)*, Puducherry, India, Dec. 2025, pp. 1236-1250.

[6] . J. Bhatta, S. R. Acharya, and K. M. Yang "Machine learning-enhanced air quality forecasting and trend analysis: A five-year comprehensive assessment of PM2.5 concentrations in Bangkok, Thailand," *Science of The Total Environment*, vol. 975, p. 178987, 2025.

[7] . M. Malhotra, S. Walia, C.-C. Lin, I. K. Aulakh, and S. Agarwal "A systematic scrutiny of artificial intelligence-based air pollution prediction techniques, challenges, and viable solutions," *Journal of Big Data*, vol. 11, 2024.

[8] . J. Duan, Y. Gong, J. Luo, and Z. Zhao, "Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer," *Scientific Reports*, vol. 13, no.1, p.12127, 2023.

[9] . A. A. Mahmoud, S. M. Alhammad, Y. Fouad, S. A. Al Qahtani, A. K. J. Saudagar, M. Aljebreen, and A. H. A. M. Elkholy, "Air quality prediction via hybrid BiGRU-MLP using binary ant colony optimization and firefly algorithm for hyperparameter tuning," *Scientific Reports*, vol. 16, no. 1, p. 19561, 2026.

[10] . L. Zhan, K. Feng, X. Gu, and T. Han, "High-precision air quality prediction via attention-driven hybrid neural networks and adaptive feature optimization," *Atmosphere*, vol. 16, no. 12, p. 1363, 2025.

[11] . K. Kabilan and J. Jebathangam, "A LSTM Deep Learning method with Attention Dense Mechanism for AQI prediction," *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*, PUDUCHERRY, India, 2023.

[12] . J. Kumari, R. Wadhvani, S. Shukla, and L. Kumar, "Regime-aware hybrid ensemble learning for adaptive forecasting in urban environments," *Scientific Reports*, vol. 16, no. 1, p. 16852, 2026.

[13] . M. M. Rahman, M. Shafiullah, S. M. Rahman, A. N. Khondaker, A. O. Amao, and M. H. Zahir, "Soft computing applications in air quality modeling: Past, present, and future," *Sustainability*, vol. 12, no. 10, p. 4044, 2020.

[14] . K. S. Rautela and M. K. Goyal, "Transforming air pollution management in India with AI and machine learning technologies," *Scientific Reports*, vol. 14, 2024.

[15] . V. Kumari, J. Singh, and G. Kaur, "Air Quality Prediction Using Machine Learning," in *The 8th International Conference on Wireless, Intelligent and Distributed Environment for Communication (WIDECOM 2025)*, *Lecture Notes on Data Engineering and Communications Technologies*, vol. 285, I. Woungang and S. K. Dhurandher, Eds., Cham: Springer, 2026.

[16] . L. Noynoo, P. Tekasakul, T. Limna, C. Choksuchat, K. Wichitsa-Nguan Jetwanna, C.-J. Tsai, T.-C. Le, P. Suwattiga, J. Morris, and R. Dejchanchaiwong, "Hybrid machine learning to enhance PM2.5 forecasting performance by the WRF-Chem model," *Atmospheric Pollution Research*, vol. 16, no. 8, p. 102558, 2025.

[17] . Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep Learning for Air Quality Forecasts: a Review," Current Pollution Reports, vol. 6, no. 4, pp. 399-409, Dec. 2020.

[18] . J. Jesupriya, R. Mahalakshmi, and C. Komalavalli, "Hybrid Deep Learning for Air Quality Prediction: A Multi-Output, Attention-Based Approach for Pollutant and AQI Classification," Journal of Information Systems Engineering and Management, vol. 10, no. 22s, 2025.

[19] . I. Ayus, N. Natarajan, and D. Gupta "Comparative analysis of machine learning and deep learning techniques for the prediction of air pollution: A case study from China," Asian Journal of Atmospheric Environment, vol. 17, p. 4, 2023.