STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **X**, October 2025, pp 1–9.
Published online in International Academic Press (www.IAPress.org)

Performance Analysis of XGBoost and LSTM Methods in Air Quality Time Series Prediction

Teguh Herlambang [1,2*], Anas Tifa Rahma Siswanti [3], Bambang Suharto [4], Zuraini Othman [5], Mohd Sanusi Azmi [6]

[1]*Department of Information System, Faculty of Economy Business and Digital Technology, Universitas Nahdlatul Ulama Surabaya, Indonesia*
[2]*Center for Data and Business Intelligence, Universitas Nahdlatul Ulama Surabaya, Indonesia*
[3]*Undergraduate Program of Information System, Faculty of Economy Business and Digital Technology, Universitas Nahdlatul Ulama Surabaya, Indonesia*
[4]*Department of Tourism and Hospitality, Faculty of Vocational, Airlangga University, Indonesia*
[5]*Department of Diploma Studies, Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Malaysia*
[6]*Department of Software Engineering, Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Malaysia*

Abstract Air pollution in urban areas like Surabaya poses significant risks to public health. Accurate forecasting of Air Pollution Index (ISPU) parameters is essential for early warning systems. Previous studies often overlook temporal dependencies by focusing on instantaneous correlations. This study proposes a temporal forecasting framework using Extreme Gradient Boosting (XGBoost) with autoregressive features and Long Short-Term Memory (LSTM) networks. Data from 2021 to 2024 were processed using a sliding window approach (lags) to capture historical patterns. Results indicate that while XGBoost provides robust baseline predictions, LSTM's ability to retain long-term dependencies yields superior stability in capturing complex pollutant fluctuations. Although RMSE and MAE values were significantly improved through temporal modeling, the moderate R2 scores suggest that external meteorological factors, such as wind speed and humidity, remain critical latent variables. This study contributes a methodological benchmark for urban air quality monitoring using hybrid machine learning approaches.

Keywords Forecasting, ISPU, XGBoost, LSTM Time Series

AMS 2010 subject classifications 62M10, 68T05

DOI: 10.19139/soic-2310-5070-3587

1. Introduction

Air pollution remains a critical environmental and public health threat, particularly in rapidly developing urban centers like Surabaya. Emissions from industrial processes, transportation, and fossil fuel combustion contribute significantly to the atmospheric concentration of primary pollutants, including particulate matter PM10, sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and nitrogen dioxide (NO_2) [1]. Epidemiological evidence has established a robust link between chronic exposure to these pollutants and the rising prevalence of respiratory and cardiovascular disorders, directly impacting the quality of life in urban communities [2]. Consequently, the World Health Organization (WHO) identifies air pollution as a leading global cause of morbidity and premature mortality, necessitating advanced monitoring and predictive frameworks to mitigate its socio-economic impact. In the Indonesian context, the government facilitates air quality management through the Air Pollutant Standard

*Correspondence to: Teguh Herlambang (Email: [teguh@unusa.ac.id]). Department of Information System, Faculty of Economy Business and Digital Technology, Universitas Nahdlatul Ulama Surabaya and Center for Data and Business Intelligence, Universitas Nahdlatul Ulama Surabaya, Indonesia

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Index (ISPU), an integrated indicator designed to classify air quality levels ranging from "good" to "hazardous" for public health guidance [3]. However, the dynamic and stochastic nature of ISPU values, influenced by localized human activities and complex meteorological interactions, presents a formidable technical challenge for predictive modeling [4]. Accurate forecasting is essential for data-driven decision-making; yet, air quality data is inherently nonlinear and heterogeneous, characterized by strong temporal dependencies where current concentrations are fundamentally influenced by their historical states [5]. To capture these complex dynamics, modern analytical approaches have shifted toward machine learning algorithms. Extreme Gradient Boosting (XGBoost) has demonstrated superior performance in modeling environmental data due to its implementation of gradient-boosted decision trees, regularization mechanisms, and efficiency in handling missing values [6]. Recognizing the limitations of static tree-based models, recent literature has advocated for the integration of Deep Learning architectures, specifically Long Short-Term Memory (LSTM) networks. As a specialized variant of Recurrent Neural Networks (RNNs), LSTM is engineered to retain long-term dependencies through sophisticated gating mechanisms, making it uniquely suited for capturing the "memory effect" and sequential patterns inherent in atmospheric pollutants [7]. This research, therefore, proposes a comparative study between XGBoost and LSTM within a strictly defined temporal forecasting framework. By implementing a multi-lag autoregressive approach, this study seeks to evaluate the predictive stability of both models across five primary pollutant parameters in Surabaya. The evaluation employs standardized scientific metrics—Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 to provide a more accurate, adaptive, and scientifically grounded air quality prediction system that addresses common methodological pitfalls in urban environmental modeling.

2. Methods

2.1. Research Type

his study employs a quantitative experimental approach with a comparative time-series forecasting framework. The research is designed to evaluate and compare the predictive performance of two distinct machine learning architectures: the tree-based eXtreme Gradient Boosting (XGBoost) and the sequence-based Long Short-Term Memory (LSTM) network. The objective is to forecast the daily Air Pollutant Standard Index (ISPU) for five parameters PM10, SO_2, CO, O_3, and NO_2 by capturing their inherent temporal dependencies.

2.2. Data Collection

The dataset was obtained from the Surabaya City Environmental Agency Dinas Lingkungan Hidup Kota Surabaya, covering the observation period from January 2021 to December 2024. The data include concentrations of five primary pollutants across multiple monitoring stations in Surabaya.

2.3. Data Processing

To ensure high data quality and scientific rigor, the following preprocessing stages were implemented:

- **Missing Value Imputation:** Linear interpolation was applied to handle minor gaps in the time-series data, maintaining chronological continuity.
- **Temporal Lag Transformation (Sliding Window):** Addressing the requirement for a true forecasting approach, we implemented a multi-lag feature engineering process. For each pollutant y at time t, we generated autoregressive lags $(y_{t-1}, \ldots, y_{t-n})$ and exogenous lags from supporting pollutants. In this study, a look-back window of 12 to 24 hours was utilized to capture short-term atmospheric memory [8].
- **Data Partitioning:** To prevent data leakage and ensure realistic forecasting evaluation, a chronological split was used. Data from 2021 to 2023 were allocated for model training, while the year 2024 was reserved as the hold-out testing set [9].
- **Normalization:** All features were transformed using Min-Max Scaling to a range of $[0, 1]$ to accelerate convergence in LSTM and ensure stable gradient descent in XGBoost.

2.4. Autoregressive XGBoost

The eXtreme Gradient Boosting (XGBoost) algorithm is an optimized distributed gradient boosting library based on the Gradient Boosted Decision Tree (GBDT) framework. In this study, the model is formulated to minimize a regularized objective function that combines a convex loss function L and a penalty term Ω for model complexity [10]:

$$\text{obj}(\theta) = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{1}$$

Where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{2}$$

represents the regularization term to prevent overfitting, with T being the number of leaves and ω the leaf weights. The prediction at step t is achieved through an additive process:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

By using a second-order Taylor expansion for the loss function, XGBoost efficiently optimizes the objective by calculating the gradients (g_i) and Hessians (h_i) for each instance.

2.5. Stacked LSTM Network

To handle the long-range temporal dependencies in air quality data, we implemented a Long Short-Term Memory (LSTM) network. The core of the LSTM is the cell state (c_t), which is regulated by three specialized gates: the forget gate (f_t), the input gate (i_t), and the output gate (o_t). The mathematical operations for a single LSTM cell at time t are defined as follows [11]: Decides what information to discard from the previous cell state:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$

Determines which new information will be stored in the cell state:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{6}$$

Cell state update:

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{7}$$

Controls the information to be passed to the next hidden state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \circ \tanh(c_t) \tag{9}$$

Where σ is the sigmoid activation function, W represents weight matrices, and b denotes bias vectors. In this study, a Stacked LSTM architecture is used, where the hidden state h_t of the first layer serves as the input for the subsequent layer, allowing the model to learn higher-level temporal abstractions.

2.6. Evaluation

The model's performance was evaluated using three main metrics :

1. Root Mean Squared Error (RMSE) : Used to measure the average prediction error against actual values [**?**].
2. Coefficient of Determination (R^2) : Indicates how well the model explains the variability of the target data [**?**].
3. Real Error Accuracy : Calculates the percentage deviation of predictions from the actual value range as an indicator of model accuracy.

3. Results and Discussions

The Air Pollution Index (API) data used in this study was obtained from the Surabaya City Environment Agency, covering the observation period from 2021 to 2024. The dataset consists of 1,461 daily data rows containing seven parameters, as shown in Table 1, namely Date, PM_{10}, SO_2, CO, O_3, NO_2, and Description. In the modeling process, five main parameters were used, namely PM_{10}, SO_2, CO, O_3, and NO_2. The selection of these five parameters was based on their significant contribution to air pollution levels, making them relevant as predictive variables in numerical analysis. Meanwhile, the other two parameters were treated differently. The *Description* parameter was removed because it does not have numerical values that support the modeling process. Conversely, the *Date* parameter was retained because it contains important temporal information in the time series approach. The presence of the time aspect allows for the chronological visualization of prediction results and supports analysis of the dynamics of pollutant concentration changes over time.

Table 1. ISPU Data for 2021-2024

Date	PM_{10}	SO_2	CO	O_3	NO_2	Description
2021-01-01	42	19	53	88	12	Moderate
2021-01-02	40	14	68	95	9	Good
2021-01-03	45	16	60	90	3	Moderate
2021-01-04	57	10	69	95	6	Moderate
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2024-12-31	28	30	48	19	12	Good

3.1. Data Exploration

In the initial stage of data exploration, the ISPU datasets from 2021 to 2024, consisting of four separate DataFrames (`df_21`, `df_22`, `df_23`, and `df_24`), were integrated. Before the merging process, each DataFrame was added a "Tahun" column as a temporal marker to maintain the systematic chronological structure of the data. After all DataFrames were successfully merged, the final form of the integrated dataset is shown in Table 2.

Table 2. ISPU Data with Year Column

Date	PM_{10}	SO_2	CO	O_3	NO_2	Tahun
01/01/2021	42	19	53	88	12	2021
02/01/2021	40	14	68	95	9	2021
03/01/2021	45	16	60	90	3	2021
04/01/2021	57	10	69	95	6	2021
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31/12/2024	28	30	48	19	12	2024

After the data integration stage is complete, the next step is to adjust the data types to ensure compatibility with statistical analysis and predictive modeling requirements. As shown in Table 2, several variables that should be numerical (PM10, SO_2, CO, O_3, and NO_2) have object data types. his occurs due to inconsistencies in format, including the presence of non-numeric characters in data entries, which can disrupt computational processes and quantitative interpretation. To address this issue, the data type was converted to float so that it could handle decimal values and missing values (NaN) more flexibly. Meanwhile, columns containing date information are also converted to the datetime data type with the dayfirst=True parameter to align with Indonesia's local date format. The final result of this conversion process is shown in Table 3, which demonstrates that all variables now have a consistent data structure.

Table 3. The Result of Alterred Datatypes

Variable	Previous Datatypes	After Alterred
Date	object	datetime64[ns]
PM10	object	float64
SO2	object	float64
CO	object	float64
O3	object	float64
NO2	object	float64
Year	int64	int64

Identification of missing values was performed by applying the `df.isna().sum()` function, which allows the calculation of the number of missing values in each variable in the dataset. Based on the identification results shown in Figure **??**, it was found that the NO_2 variable had the highest number of missing values compared to other variables. To address this issue, the linear interpolation method was used, which is an approach that estimates missing values based on the trend between previous and subsequent data points. This method was chosen because it can maintain the main temporal pattern in the data without causing distortion. After the interpolation process is complete, a re-verification is performed to ensure that all empty values have been successfully filled in.

Table 4. Imputation of Missing Value Checking

Variable	Previous Datatypes	After Alterred
Date	0	0
PM10	2	0
SO2	8	0
CO	14	0
O3	10	0
NO2	110	0
Year	0	0

Correlation analysis between variables was performed using Pearson's coefficient, which aims to identify potential linear relationships as a basis for developing predictive models. The results of the analysis, visualized through a correlation heatmap in Figure 1, show that the O_3 and CO variables have the highest positive correlation, namely 0.73. This value indicates a strong relationship, both in terms of emission sources and similar chemical processes [**?**]. Conversely, a negative correlation was found between SO_2 and CO and O_3, with r values of –0.31 and –0.30, respectively. These correlation findings provide an important foundation for feature selection in modeling using the XGBoost algorithm, considering the relevance and contribution of each variable to the prediction target.

3.2. Split of Feature and Target Variables

In this study, the predictive modeling framework is specifically configured to analyze the inter-pollutant relationship between two critical parameters. The model designates Ozone (O_3) as the Target Variable (y), while Carbon Monoxide (CO) is utilized as the primary Predictor Variable (X).

3.3. Split Training Data and Testing Data

The splitting of training and testing data was done based on time attributes by referring to the Year column in the Air Pollution Index (API) dataset. This approach was applied to maintain chronological order, so that the model was trained using past data and tested on future data, in accordance with the principle of representative predictive evaluation. Data from 2021 to 2023 is used as training data and stored in the train_data variable, while data from 2024 is used as test data and stored in the test_data variable.

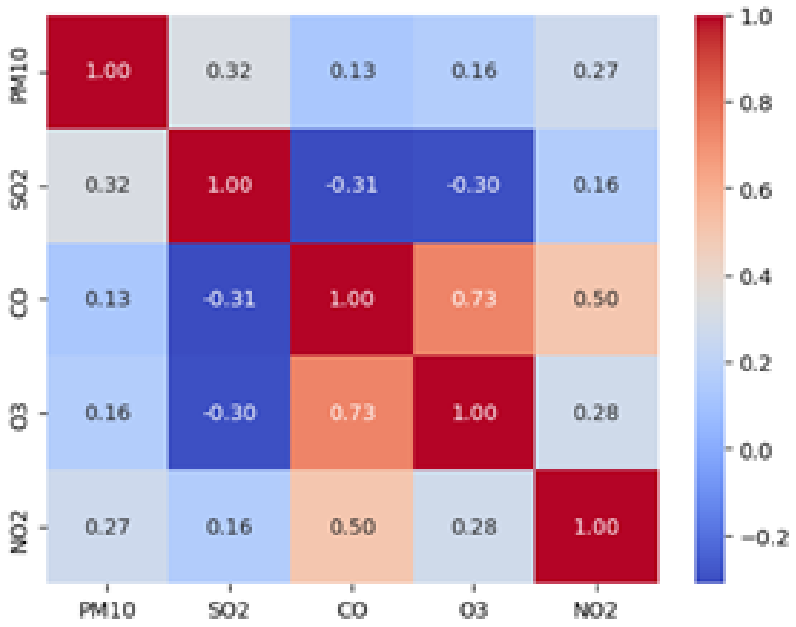

Figure 1. Numeric Variable Correlation Heatmap

3.4. Method Implementation

In this study, the XGBoost and LSTM models were implemented to forecast hourly O_3 concentrations by utilizing lag-based features derived from historical O_3 and CO observations. The dataset was restructured into a supervised learning framework, incorporating multiple lag variables to capture inherent temporal dependencies. Figure 2 illustrates the comparative performance of these models against actual O_3 concentrations during the initial 120-hour testing period of 2024.

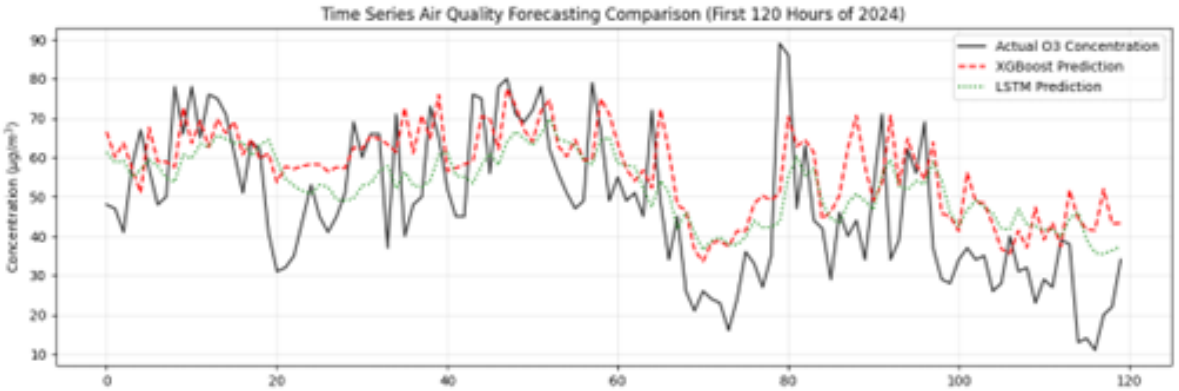

Figure 2. Qualitative Trend Analysis

Both architectures successfully synchronized with the overarching oscillatory patterns of ozone levels, demonstrating their capacity to capture cyclical temporal trends. However, a notable discrepancy was observed during extreme peak episodes; the models struggled to fully replicate abrupt spikes. This suggests a inherent limitation in modeling high-frequency variability when relying solely on historical pollutant lags without auxiliary meteorological inputs. To further evaluate the models' statistical consistency, scatter plots of actual versus predicted

values were generated (Figure 3). The positive linear relationships observed in both models validate that the predicted outputs remain consistent with the ground truth across the data range.

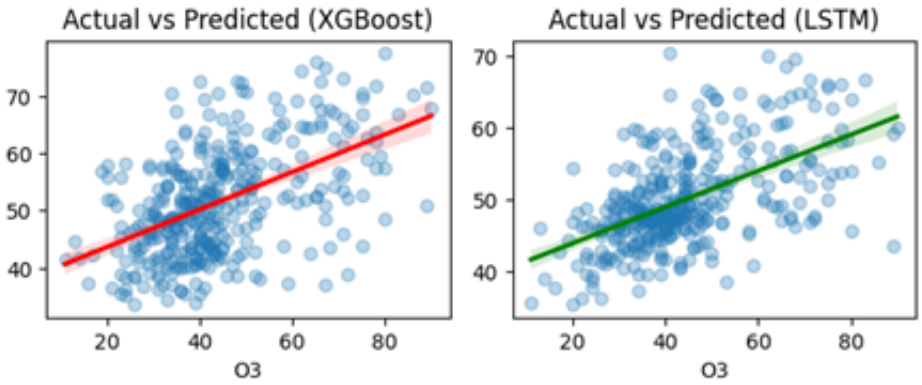

Figure 3. Regression and Residual Evaluation

Nevertheless, the dispersion around the regression line becomes more pronounced at higher O_3 concentrations, indicating an increase in heteroscedasticity and prediction error under extreme conditions. Comparatively, the XGBoost model appears more responsive to high-frequency fluctuations, whereas the LSTM produced a smoother predictive curve. This difference highlights the contrast between XGBoost's local-pattern decision trees and LSTM's global-sequential memory approach. To enhance scientific transparency, SHAP (Shapley Additive Explanations) analysis was applied to the XGBoost model to quantify feature contributions (Figure 4). The analysis identifies O_3_lag_1 as the overwhelmingly dominant feature, significantly surpassing the influence of all other variables. This finding confirms the presence of strong temporal autocorrelation in urban ozone dynamics, where the immediate preceding observation provides the primary signal for short-term prediction.

Furthermore, additional lags of both O_3 and CO contribute with diminishing magnitudes as temporal distance increases, aligning with the expected decay in environmental memory. The inclusion of CO lags in the importance hierarchy validates its role as a relevant predictor, likely reflecting the common emission patterns that drive ozone precursors in the Surabaya metropolitan area. To provide a rigorous assessment of the forecasting models, the predictive accuracy was quantified using three standard statistical metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Table 5 summarizes the performance comparison between the Autoregressive XGBoost and the Stacked LSTM models based on the 2024 testing dataset.

Table 5. Quantitative Performance Evaluation

Method	**RMSE**	**MAE**	R^2
XGBoost (Autoregressive)	14.94	12.33	-0.002259
Stacked LSTM	13.83	11.61	0.140797

The numerical results indicate that the Stacked LSTM model outperformed XGBoost across all metrics. Specifically, the LSTM achieved a lower RMSE of 13.83 and a lower MAE of 11.61, representing a more precise fit to the actual O_3 concentrations. Most notably, while the XGBoost model yielded a near-zero (slightly negative) R^2 score of -0.002, the Stacked LSTM achieved a positive R^2 of 0.141. The near-zero R^2 for XGBoost suggests that the tree-based gradient boosting approach struggled to capture the variance of the hourly O_3 fluctuations effectively using only the provided lag features. In contrast, the positive R^2 of the Stacked LSTM demonstrates its superior ability to leverage sequential memory via its gating mechanisms. Although the R^2 remains in the modest range—likely due to the exclusion of meteorological variables as noted in the limitations—the transition from a negative to a positive R^2 signifies a substantial improvement in the model's reliability for urban air quality forecasting.

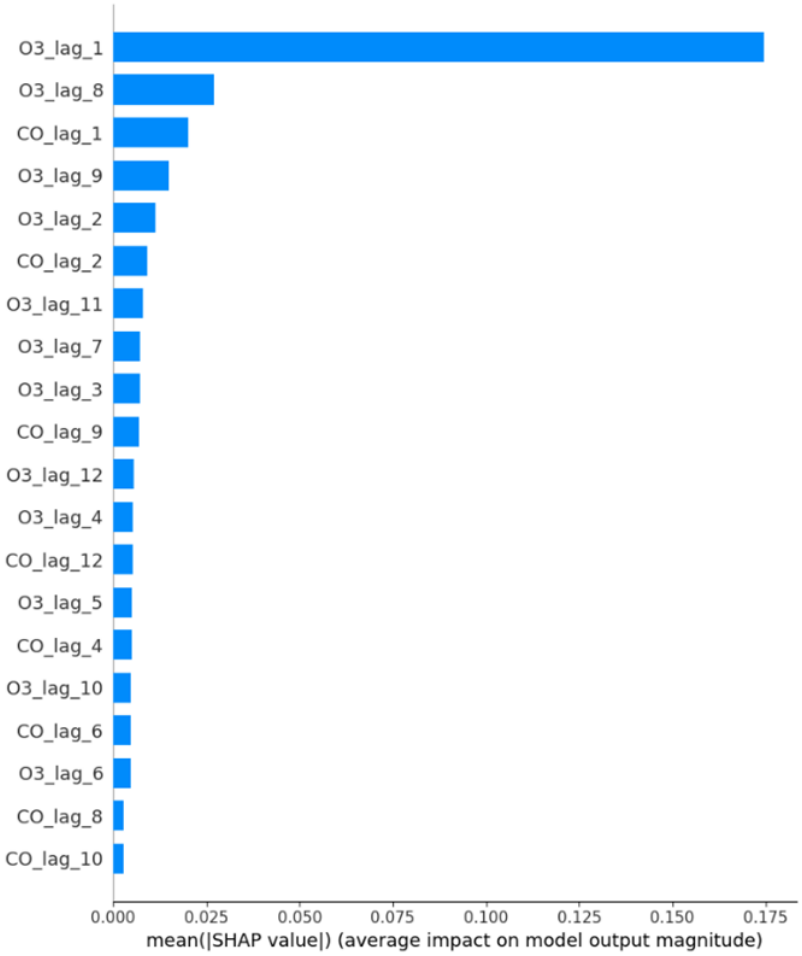

Figure 4. Model Interpretability via SHAP Analysis

4. Conclusion

This study evaluated the efficacy of Autoregressive XGBoost and Stacked LSTM models in forecasting hourly Ozone O_3 concentrations in Surabaya using lag-based features. The results demonstrate that incorporating temporal dependencies through historical O_3 and CO values significantly improves the model's ability to track urban air quality fluctuations compared to static baseline approaches. The comparative analysis reveals that the Stacked LSTM architecture is superior to XGBoost, achieving a lower RMSE (13.83) and a positive R^2 score (0.141). While XGBoost showed high volatility and struggled to capture variance (resulting in a near-zero R^2), the gated mechanism of LSTM successfully retained long-term temporal "memory," providing more stable and synchronized predictions. Furthermore, SHAP interpretability analysis confirmed that while the immediate previous hour (O_3_lag_1) is the primary driver, historical CO levels serve as critical secondary predictors, validating the multivariate approach. Despite the improvements, the modest R^2 values indicate that temporal lags alone cannot fully account for the stochastic nature of ozone formation. The primary limitation of this study is the absence of meteorological parameters such as wind speed and temperature, which are decisive factors in pollutant dispersion. Therefore, future research should integrate exogenous weather data and explore hybrid deep learning architectures to further enhance forecasting accuracy. Ultimately, this research provides a robust technical benchmark for the development of real-time Air Pollutant Standard Index (ISPU) monitoring systems in Indonesia.

Acknowledgments

The author would like to thank LPPM-Universitas Nahdlatul Ulama Surabaya (UNUSA) for their support in providing the facility for this research. Additionally, gratitude is extended to the Centre of Research and Innovation Management of Universiti Teknikal Malaysia Melaka (UTeM) for sponsoring the publication fees under the Tabung Penerbitan CRIM UTeM.

Conflicts of interest

The authors declare that there are no conflicts of interest.

REFERENCES

1. Arriazu-Ramos, A., et al., *Health Impacts of Urban Environmental Parameters: A review of air pollution, heat, noise, green spaces and mobility*, Sustainability, vol. 17, no. 10, p. 4336, May 2025. doi: 10.3390/su17104336.
2. Rajesh, M., Babu, R. G., Moorthy, U., and Easwaramoorthy, S. V., *Machine learning-driven framework for realtime air quality assessment and predictive environmental health risk mapping*, Scientific Reports, vol. 15, no. 1, p. 28801, Aug. 2025. doi: 10.1038/s41598-025-14214-6.
3. Ali, M. C., Ebrahim, E. E. M., and Abonazel, M. R., *Air quality forecasting using a modified statistical approach: Combining statistical and machine learning methods*, International Journal of Innovative Research and Scientific Studies, vol. 8, no. 4, pp. 1321–1335, Jun. 2025. doi: 10.53894/ijirss.v8i4.8061.
4. Jing, H., and Wang, Y., *Research on Urban Air Quality Prediction based on Ensemble Learning of XGBOOST*, E3S Web of Conferences, vol. 165, p. 02014, Jan. 2020. doi: 10.1051/e3sconf/202016502014.
5. Mao, W., Wang, W., Jiao, L., Zhao, S., and Liu, A., *Modeling air quality prediction using a deep learning approach: Method optimization and evaluation*, Sustainable Cities and Society, vol. 65, p. 102567, Oct. 2020. doi: 10.1016/j.scs.2020.102567.
6. Sitienei, M., Anapapa, A., and Otieno, A., *Application of XGBOOST regression in maize yield prediction*, Asian Journal of Probability and Statistics, vol. 24, no. 1, pp. 1–9, Aug. 2023. doi: 10.9734/ajpas/2023/v24i1513.
7. Zhao, D., Jiang, S., and Wu, Y., *Air quality prediction and early warning model based on LSTM-SARIMA*, Highlights in Science Engineering and Technology, vol. 63, pp. 153–160, Aug. 2023. doi: 10.54097/hset.v63i.10867.
8. Espinosa, R., Jiménez, F., and Palma, J., *Multi-objective evolutionary spatio-temporal forecasting of air pollution*, Future Generation Computer Systems, vol. 136, pp. 15–33, May 2022. doi: 10.1016/j.future.2022.05.020.
9. Havaei, M. A., Shahhosseini, V., and Maknoon, R., *Continuous-time air pollutant forecasting using multi-timescale attention neural ordinary differential equations (MA-NODE)*, Scientific Reports, vol. 15, no. 1, p. 44178, Dec. 2025. doi: 10.1038/s41598-025-27903-z.
10. Eva, M. N., Audri, A. I., Nur, F. N., and Islam, A. H. M. S., *Air Quality Monitoring and Pollution Prediction Using Machine Learning*, 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), Rangpur, Bangladesh, 2025, pp. 1–6. doi: 10.1109/QPAIN66474.2025.11171644.
11. Liu, Y., Tee, M., Lu, L., Zhou, F., and Lu, B., *High-Precision urban air quality prediction using a LSTM-Transformer hybrid architecture*, International Journal of Advanced Computer Science and Applications, vol. 16, no. 4, Jan. 2025. doi: 10.14569/ijacsa.2025.0160431.
12. Jg, S. S., Gk, D., and M, V., *Spatiotemporal Patterns in Air pollution: A hybrid machine learning for Composite AQI prediction*, International Journal of Advanced Research in Science Communication and Technology, pp. 194–205, Jul. 2025. doi: 10.48175/ijarsct-28835.