Analysis of PM2.5 pollutant sources in Jakarta using deep learning models and back trajectory approach

Hendro Pratama Saragih, Imas Sukaesih Sitanggang, Hendra Rahmawan
Department of Computer Science, School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia

Article Info

Article history:

Received Dec 23, 2025
Revised Jun 16, 2026
Accepted Jun 27, 2026

Keywords:

Backward trajectory
HYSPLIT
LSTM
PM2.5
Regional transport

ABSTRACT

PM2.5 concentrations in Jakarta frequently exceed World Health Organization (WHO) air quality guidelines, indicating the need for an integrated approach for pollution prediction and source assessment. This study develops a spatiotemporal prediction framework using a long short-term memory (LSTM) model integrated with the hybrid single particle Lagrangian integrated trajectory (HYSPLIT) model for backward trajectory analysis. Daily PM2.5 data from five monitoring stations were combined with meteorological variables from ERA5, Visualcrossing, and the global data assimilation system, with spatial context evaluated using Sentinel-2 land cover maps. After hyperparameter tuning, the optimized model demonstrated robust predictive capabilities, achieving a peak coefficient of determination (R^2) of 75.87% on the test data. The framework exhibited exceptional relative accuracy, particularly at the Jagakarsa and Kebun Jeruk stations, which recorded mean absolute percentage error (MAPE) values of 13.34% and 17.80%, respectively. Backward trajectory analysis during selected pollution episodes indicates two dominant regional transport pathways that may influence PM2.5 levels in Jakarta. These pathways are associated with air mass transport over industrial and built-up areas in eastern and northern regions surrounding Jakarta. Land cover analysis shows limited vegetation along these pathways. Overall, elevated PM2.5 events are associated with combined local emissions, regional transport, and meteorological conditions that limit pollutant dispersion near the surface.

Corresponding Author:

Imas Sukaesih Sitanggang
Department of Computer Science, School of Data Science, Mathematics and Informatics, IPB University
16680 Dramaga, Bogor, Indonesia
Email: imas.sitanggang@apps.ipb.ac.id

1. INTRODUCTION

Jakarta faces a serious challenge in air quality management, with data [1] showing that the annual average PM2.5 concentration reaches 108.6 μg/m³, far exceeding the World Health Organization (WHO) interim target of 35 μg/m³ [2]. This high level of pollution is triggered by massive anthropogenic activities, as a result of various transportation activities, both domestic and industrial, as well as construction activities aimed at improving the city's infrastructure. Air quality in Jakarta is influenced by relatively stagnant meteorological conditions, with average surface wind speeds ranging from 1.5 to 2.0 m/s. The dominant eastward wind pattern carries accumulated pollutants from the buffer zone [2]. The health and environmental impacts of this situation are significant. A study by Dinas Lingkungan Hidup Provinsi DKI Jakarta [3] found a positive correlation between an increase in PM2.5 and a 12% surge in cases of acute respiratory infections (ARI).

Despite its high urgency, the current air quality management system has fundamental limitations. The research highlights that traditional emission source analysis based on static inventories fails to account for real-time atmospheric dynamics. Furthermore, conventional prediction methods often neglect the integration of crucial atmospheric parameters, as criticized [4]. While most prior studies focus on air quality prediction or source analysis as isolated tasks, this study bridges the gap by proposing a novel, integrated system architecture. Specifically, this is the first integrated long short-term memory (LSTM) and hybrid single particle Lagrangian integrated trajectory (HYSPLIT) framework applied to Jakarta that combines local meteorological monitoring with satellite validation (Sentinel-2) for precise source attribution. This end-to-end framework not only forecasts temporal PM2.5 concentrations but also dynamically triggers spatial backward trajectory analysis to highlight transboundary emission corridors, providing a crucial, actionable tool for regional air quality policy and engineering early-warning systems. The research [5] has demonstrated that the LSTM model is capable of predicting PM2.5 with high accuracy (R^2 >0.9), which in this study will be combined with the HYSPLIT model for cross-border emission source tracking, validated for effectiveness by [6], to produce an early warning system and mitigation that is more precise.

2. METHOD

This study comprises several stages: data preparation, data preprocessing, model training, hyperparameter tuning, model testing, HYSPLIT simulation, and PM2.5 source analysis. Details of the research stages are presented in Figure 1.

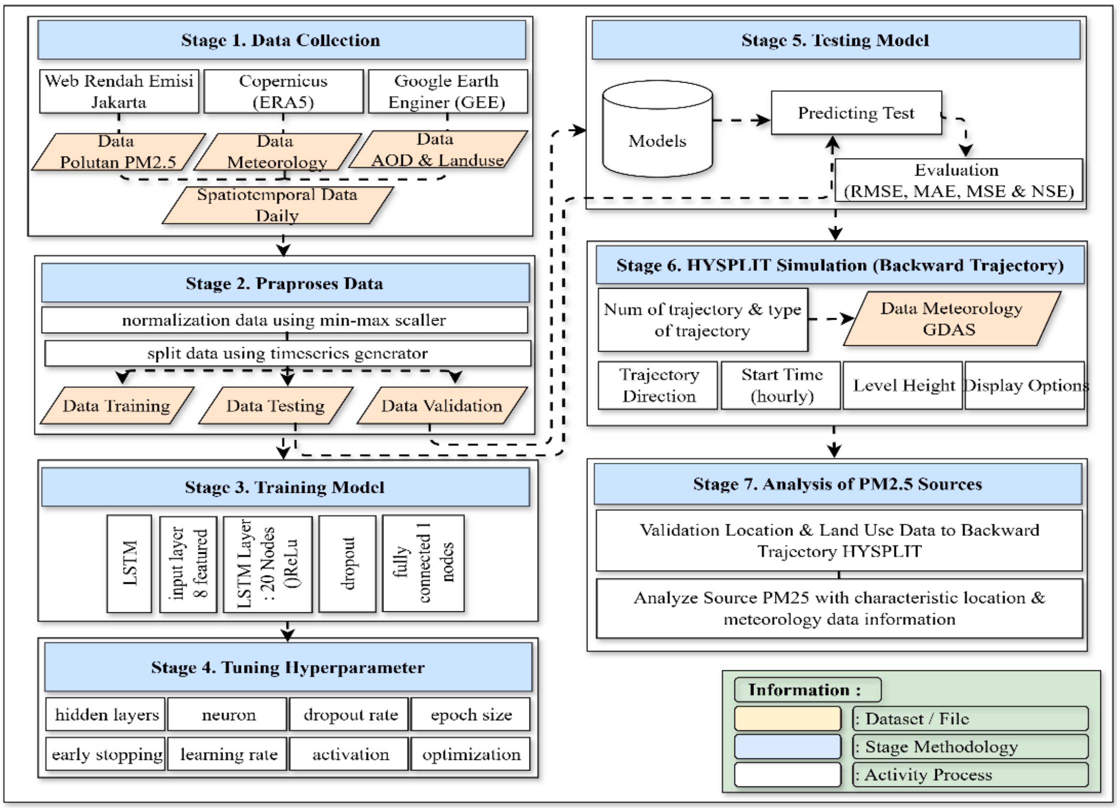

Figure 1. Research steps

2.1. Data collection

The dataset used various sources [7] relevant to the research focus, namely PM2.5 concentration predictions and source tracking, covering the period from January 2022 to December 2024. PM2.5 data were obtained via web scraping (with permission from the owner, Jakarta Environmental Service) from Rendah Emisi Jakarta Web. The data provides daily air quality information from several monitoring stations. Meanwhile, meteorological data were obtained from two sources: ERA5 Copernicus (European Centre for Medium-Range Weather Forecasts: ECMWF) [8], [9] and Visualcrossing [10] for model-prediction input features [11], [12], and global data assimilation system (GDAS) of the National Oceanic and Atmospheric Administration (NOAA) as input for HYSPLIT simulations. Data collection was carried out by integrating datasets from various sources to support comprehensive spatiotemporal analysis, including daily PM2.5

concentration data obtained through web scraping techniques from the Jakarta Rendah Emisi web portal at five monitoring stations (Jan 2022 – Dec 2024), global meteorological data (temperature, humidity, wind, rainfall, and dew point) from ERA5 Copernicus, Visualcrossing, and GDAS, as well as spatial land cover data and aerosol optical depth (AOD) extracted from Sentinel-2 and MODIS satellite imagery using Google Earth Engine (GEE) [13].

2.2. Data preprocessing

The data preprocessing stage is applied to ensure the quality of model input, starting with the handling of missing values using linear interpolation to maintain the continuity of the time trend, followed by the normalization of all variables to the range 0 until 1 using Min-Max Scaler for training stability, and ending with chronological data splitting into 70% training data, 15% validation data, and 15% test data [14] to evaluate the model's generalization ability on future data.

2.3. Training the LSTM model for PM2.5 prediction

LSTM model training was conducted using a multivariate time-series dataset [15] that had been normalized with a Min-Max Scaler and arranged into a sequence of 30 timesteps to predict daily PM2.5 concentrations one step ahead. The model architecture consisted of two stacked LSTM layers followed by a dense fully connected layer as the output, with the learning process optimized using the Adam algorithm, sigmoid, and rmsprop [16], and an early stopping mechanism to mitigate the risk of overfitting [17].

Figure 2 illustrates the architecture of the deep learning model employed in this study, which processes six meteorological input features via a hidden layer comprising stacked LSTM units with batch normalization and dropout, to predict PM2.5 concentration. The network architecture utilizes mean squared error (MSE) as the loss function, which is highly suitable for regression tasks in continuous pollutant concentration forecasting. The selected temporal window of 30 timesteps was explicitly chosen as it optimally captures the monthly cyclical patterns of local meteorology and anthropogenic emissions in Jakarta without introducing excessive historical noise.

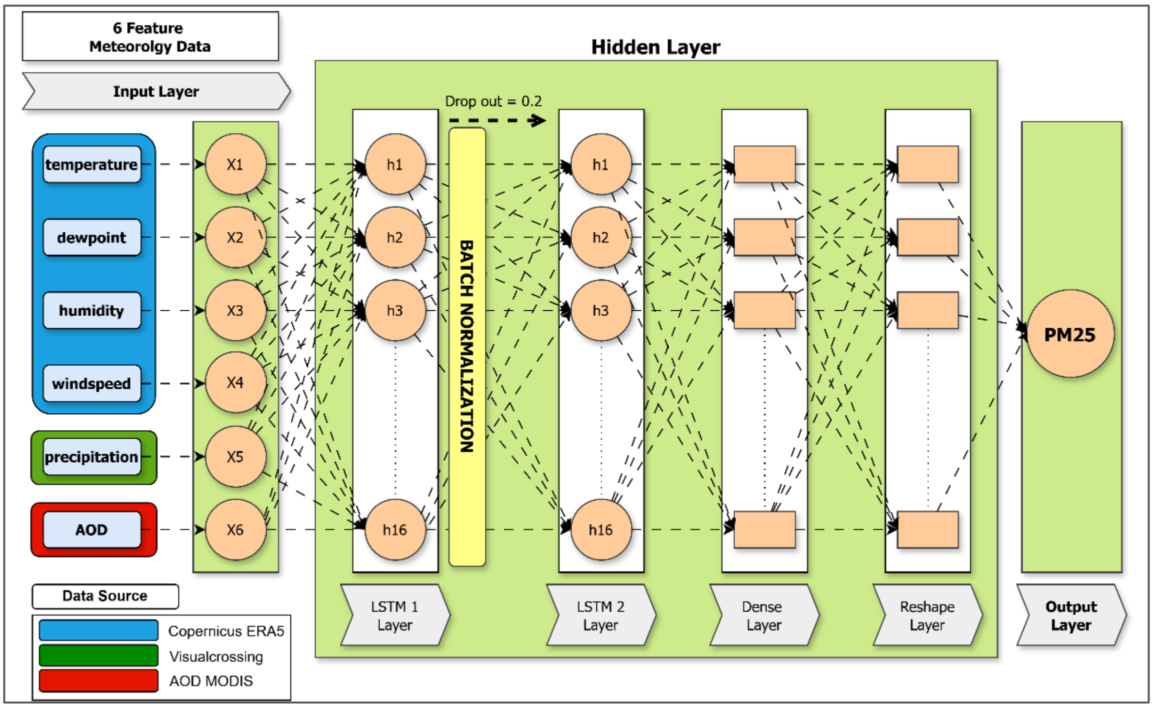

Figure 2. Multivariate LSTM models architecture

2.4. Hyperparameter tuning on LSTM models

Hyperparameter tuning is performed to optimize the LSTM architecture. The main parameters adjusted include the number of hidden layers, the number of neurons per layer, the dropout rate, the epoch size, and the activation function [18]. In addition, the early stopping technique is used to prevent overfitting, and various learning rate values and optimization algorithms are tested. These adjustments are made iteratively based on validation evaluation results to select the model configuration that provides the best performance. The hyperparameter tuning process uses a grid search [19], [20] to test various parameter combinations. The grid parameters used are shown in Table 1.

Hyperparameter tuning was performed to find the optimal configuration for the LSTM model. Through a grid search approach, various combinations of parameters such as the number of hidden layers and neurons, dropout rate, epoch size, and activation function were tested iteratively.

Table 1. Grid search parameter for multivariate LSTM model

No	Parameter	Value
1	Optimizer	adam, nadam, rmsprop, sgd
2	Learning Rate	0.0001, 0.001, 0.01, 0.1
3	LSTM units	16, 32, 64
4	Dropout rate	0.2
5	Epoch	1000 with early stopping
6	Batch size	32

2.5. LSTM model for PM2.5 prediction

The LSTM model was tested by applying the trained model to a test set comprising separate period data that the model had never seen before to evaluate its generalization ability and prediction accuracy in real-world scenarios. Model performance was comprehensively validated by comparing the predicted daily PM2.5 concentration values with actual observation data using evaluation metrics, namely root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2), to assess the accuracy of the data distribution pattern.

1. RMSE measures how far the prediction results are from the actual values [21]. The RMSE formula is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(yi - \hat{yi})^2} \tag{1}$$

2. MAE to measure the average absolute error [21]. The following is the mathematical of MAE formula:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|yi - \hat{yi}\right| \tag{2}$$

3. R² Score as an indicator of goodness-of-fit [21]. The mathematical equation for the R² formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(yi-\hat{yi})^2}{\sum_{i=1}^{N}(yi-\bar{yi})^2} \tag{3}$$

4. MAPE to calculate the absolute percentage error [22]. The following is the mathematical equation for the MAPE formula:

$$\mathrm{MAPE} = \sum_{i=1}^{N}\left|\frac{yi-\hat{yi}}{\hat{yi}}\right| \tag{4}$$

where:
yi : actual value i
$\hat{yi}$: the i-th predicted value
$\bar{yi}$: the average of all actual values (yi)
N : total number of observations

This verification step serves as a quality gate to ensure the reliability of predictions before they are further integrated with the HYSPLIT trajectory simulation.

2.6. HYSPLIT simulation

From an engineering system perspective, the HYSPLIT backward trajectory serves as a dynamically triggered atmospheric verification module. It is autonomously activated only when the LSTM predicts PM2.5 concentrations exceeding 75 µg/m³. This specific threshold is strongly justified as it aligns with the 'Unhealthy' category under Indonesia's Air Pollution Standard Index (ISPU), ensuring that the high computational cost of spatial trajectory simulations is efficiently allocated only during extreme, high-impact pollution episodes [23]. In this scenario, the HYSPLIT model runs a 72-hour backward trajectory analysis at an altitude of 500 meters above ground level using wind data from GDAS, to track air mass trajectories and identify the geographical origins of regional potential pollutant sources [24] contributing to pollution episodes in Jakarta.

2.7. Analysis of PM2.5 sources

Validation with satellite data, such as Sentinel-2, plays a key role in this integration. Land use data from Sentinel-2 helps classify potential pollutant sources, such as distinguishing between emissions and vegetation [25]. Additionally, satellite AOD data is used to correct surface pollutant concentrations. A case study in Beijing shows that integration of Sentinel-2 data with HYSPLIT can reduce uncertainty in source identification [26]. Some frameworks [27], [28] provide algorithms for trajectory clustering, connecting pollution peaks to dominant sources, and visualizing results in GIS. An example of its successful application is the Mekong Delta, where this framework helped identify PM2.5 sources from land fires [6]. This approach not only improves analysis accuracy but also facilitates data visualization and interpretation.

3. RESULTS AND DISCUSSION

3.1. Data preprocessing

The preprocessing stage is a crucial phase for ensuring optimal data quality and structure for the LSTM architecture. Before data processing, it is necessary to analyze the missing values from each station, as shown in Table 2. Then, compare the statistical summary values from each station for the PM2.5 variable before preprocessing, as shown in Table 3, and the statistical summary values after preprocessing, as shown in Table 4.

Table 2. Number of missing value in PM2.5 Data

PM2.5 station	Number of data	Missing values
DKI1 Bundaran HI	895	201
DKI2 Kelapa Gading	994	102
DKI3 Jagakarsa	895	201
DKI4 Lubang Buaya	895	201
DKI5 Kebun Jeruk	895	201

Table 3. Summary of PM2.5 data statistics before data preprocessing

PM2.5 station	Number of data	Average	Standard deviation	Min	Median	Max
DKI1 Bundaran HI	895	70.16	21.52	0	73.50	125.16
DKI2 Kelapa Gading	994	56.49	19.61	0	69.52	109.50
DKI3 Jagakarsa	895	71.61	19.45	0	72.61	137.16
DKI4 Lubang Buaya	895	72.47	20.70	0	75.55	131.72
DKI5 Kebun Jeruk	895	56.49	37.94	0	68.80	153.77

Table 4. Summary statistics of PM2.5 data after data preprocessing

PM2.5 Station	Number of data	Average	Standard deviation	Min	Median	Max
DKI1 Bundaran HI	1096	70.21	19.73	3.33	73.03	125.17
DKI2 Kelapa Gading	1096	66.40	20.23	0.72	68.13	109.55
DKI3 Jagakarsa	1096	71.16	17.89	2.55	71.84	137.16
DKI4 Lubang Buaya	1096	71.47	20.39	7.39	75.08	131.72
DKI5 Kebun Jeruk	1096	70.03	20.92	4.22	68.91	153.77

3.2. LSTM model

The prediction model was developed using a multivariate LSTM architecture optimized through grid search to determine the best hyperparameters (number of neurons, learning rate, optimizer) at each station. Table 5 shows the predictive performance of the reliable model, as indicated by the highest R^2 value of 75.87% at the DKI1 Bundaran HI Station and the lowest MAPE of 13.34% at the DKI3 Jagakarsa Station, demonstrating the model's effectiveness in capturing daily PM2.5 fluctuation patterns.

The variance in predictive performance across stations highlights the influence of local urban micro-environments. The lowest MAPE achieved at Jagakarsa (13.34%) can be attributed to its suburban topography and denser vegetation cover, which buffers sudden emission spikes and results in more stable meteorological patterns. Conversely, Bundaran HI achieved the highest $\mathbf{R^2}$ (75.87%) due to its highly rhythmic and predictable anthropogenic emissions, primarily driven by dense, daily vehicular traffic in the central business district. Furthermore, while the overall $\mathbf{R^2}$ values (65.55% - 75.87%) demonstrate moderate to high predictive capability, their limitations must be acknowledged.

The model occasionally underestimates sudden, anomalous peak events. This limitation stems from unpredictable local emission sources (sudden waste burning or abrupt construction activities) that are not captured by standard meteorological or AOD input features. LSTM significantly excels in capturing non-linear temporal dependencies, though future iterations could benefit from integrating real-time traffic density APIs to further push the $\mathbf{R^2}$ boundary.

Table 5. LSTM modeling accuracy results for each PM2.5 monitoring station

PM2.5 Station	R^2	RMSE	MAE	MAPE
DKI1 Bundaran HI	75.87%	11.4671	8.5046	29.92%
DKI2 Kelapa Gading	70.28%	11.4967	8.6267	47.38%
DKI3 Jagakarsa	65.55%	9.6808	7.3604	13.34%
DKI4 Lubang Buaya	70.01%	12.6069	9.3328	16.35%
DKI5 Kebun Jeruk	74.55%	14.2493	10.1972	17.80%

3.3. HYSPLIT simulation

The HYSPLIT backward trajectory simulation was conducted in response to an extreme pollution episode (>75 μg/m³) and concentrations that exceeded the thresholds set by national regulations. Based on the Indonesian Government Regulation No. 22 of 2021, which sets the ambient air quality standard (NAB) at 55 μg/m³ per day. In accordance with DKI Jakarta Governor Regulation Number 77 of 2020 on air pollution control, this extreme threshold exceedance triggered an investigation into the source of potential pollutants to track the origin of the air mass during the previous 24-72 hours, which successfully revealed two main corridors of regional pollutant transport from the east-southeast and north-northwest, and identified the meteorological phenomenon of subsidence (the descent of air masses from an altitude of >1000 m to the surface) which plays a crucial role in trapping pollutants in the human respiratory tract. Table 6 describes the extreme pollutant episodes that occurred during the period from January 2022 to December 2024. Based on the data distribution in Figure 3, backward trajectories were generated for each PM2.5 monitoring station during pollutant episodes, as shown in Table 6. Figure 3(a) presents the HYSPLIT backward trajectory for the DKI1 Bundaran HI Station, Figure 3(b) for the DKI2 Kelapa Gading Station, Figure 3(c) for the DKI3 Jagakarsa Station, Figure 3(d) for the DKI4 Lubang Buaya Station, and Figure 3(e) for the DKI5 Kebun Jeruk Station. These trajectories reveal different regional transport pathways of air masses that may contribute to elevated PM2.5 concentrations at each monitoring station.

Table 6. Pollutant episodes with the highest PM2.5 concentrations

PM2.5 station	Longitude	Latitude	Episode pollutant	PM2.5 concentration	AOD
DKI1 Bundaran HI	106.82350	-6.19466	2 Oct 2023	125.20	1.25
DKI2 Kelapa Gading	106.91088	-6.15357	17 Nov 2024	112.80	1.36
DKI3 Jagakarsa	106.80367	-6.35693	18 Jun 2024	137.20	1.88
DKI4 Lubang Buaya	106.90919	-6.28889	28 Sep 2023	123.60	2.71
DKI5 Kebun Jeruk	106.75318	-6.20734	17 Nov 2024	153.80	1.89

To map the origins of air masses carrying pollutants during extreme pollution episodes, HYSPLIT backward trajectory simulations were run at five selected monitoring stations in Jakarta when PM2.5 concentrations exceeded 75 μg/m³. This summary of the simulation results presents key parameters, including the timing of occurrence, the dominant direction of air-mass trajectories, and the identified vertical atmospheric dynamics.

3.4. Analysis of PM2.5 sources

Pollutant source analysis was conducted through spatial integration (overlay) between the HYSPLIT trajectories. The Sentinel-2 land cover map, which validated that the air mass carrying the pollutants was consistently crossing areas with high emission densities such as the manufacturing industrial corridor in

Karawang-Bekasi, heavy industrial areas in Banten, and coal-fired power plants on the north coast, confirms that extreme pollution in Jakarta is the result of a superposition of local urban emissions and transboundary pollution.

Land cover analysis is explained in Table 7 and illustrated for all monitoring stations in Figure 4. The overlay results between the Esri Sentinel-2 land cover map and the HYSPLIT backward trajectories are presented in Figure 4. Figure 4(a) illustrates the DKI1 Bundaran HI Station, Figure 4(b) shows the DKI2 Kelapa Gading Station, Figure 4(c) presents the DKI3 Jagakarsa Station, Figure 4(d) depicts the DKI4 Lubang Buaya Station, and Figure 4(e) illustrates the DKI5 Kebun Jeruk Station. The overlay demonstrates that the air-mass pathways intersect different land-cover types, including built-up areas, industrial zones, agricultural land, and coastal regions, providing spatial evidence for identifying potential PM2.5 emission sources.

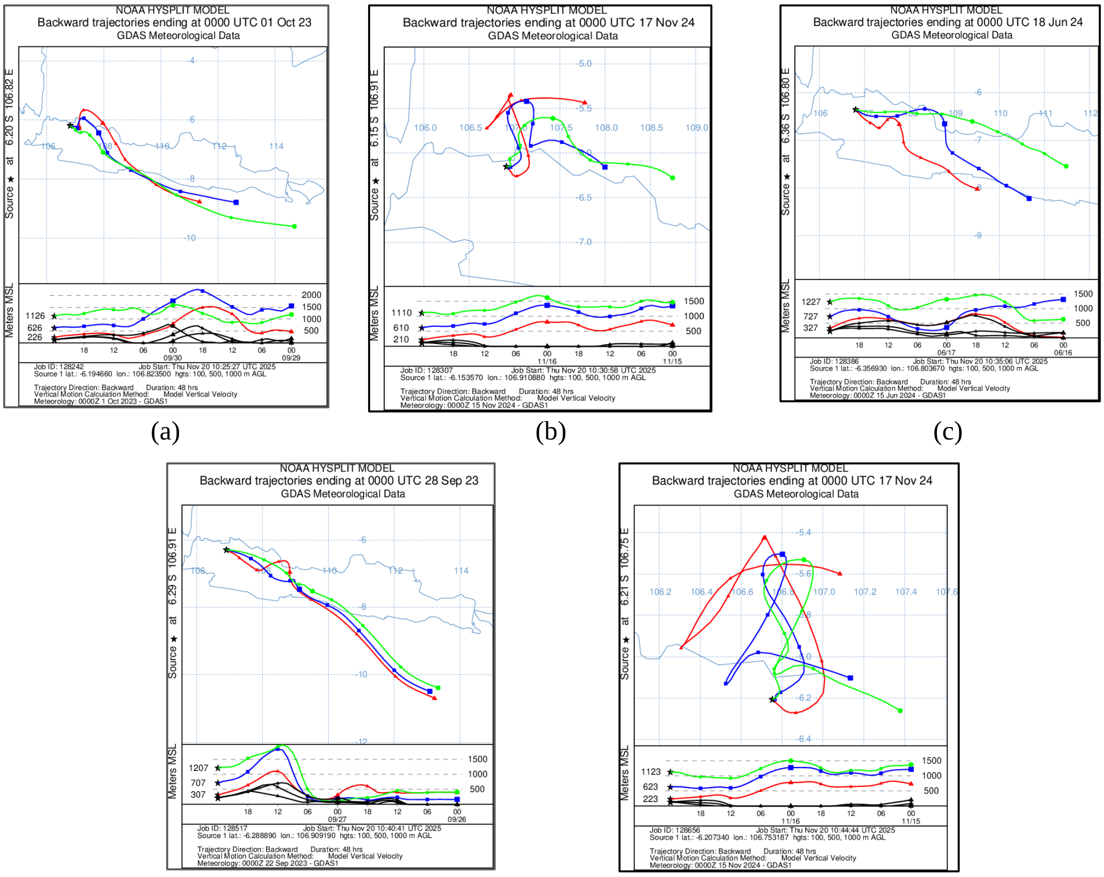

Figure 3. HYSPLIT backward trajectory for each PM2.5 station in Jakarta: (a) DKI1 Bundaran HI Station, (b) DKI2 Kelapa Gading Station, (c) DKI3 Jagakarsa Station, (d) DKI4 Lubang Buaya Station, (e) DKI5 Kebun Jeruk Station

Table 7. Result of overlay land cover ESRI Sentinel2 and HYSPLIT trajectory for each station PM2.5

Monitoring station	Dominant land cover classification	Specific areas passed	Identification of potential emission sources
DKI1 Bundaran HI	Built-up areas & crops	Karawang - Bekasi Industry	Manufacturing industry & land burning in Pantura.
DKI2 Kelapa Gading	Built-up area & water	Priok & Marunda Ports	Emissions from ships, ports, and coastal industries.
DKI3 Jagakarsa	Built-up area	Cikarang - Cibinong Industry	Accumulation of eastern and local domestic industrial emissions.
DKI4 Lubang Buaya	Built-up area (dense)	Jababeka & MM2100	Direct advection of factory emissions & heavy logistics.
DKI5 Kebun Jeruk	Built-up area & crops	Cilegon & Serang Industries	Cross-border transportation for Banten power plants & steel industry.

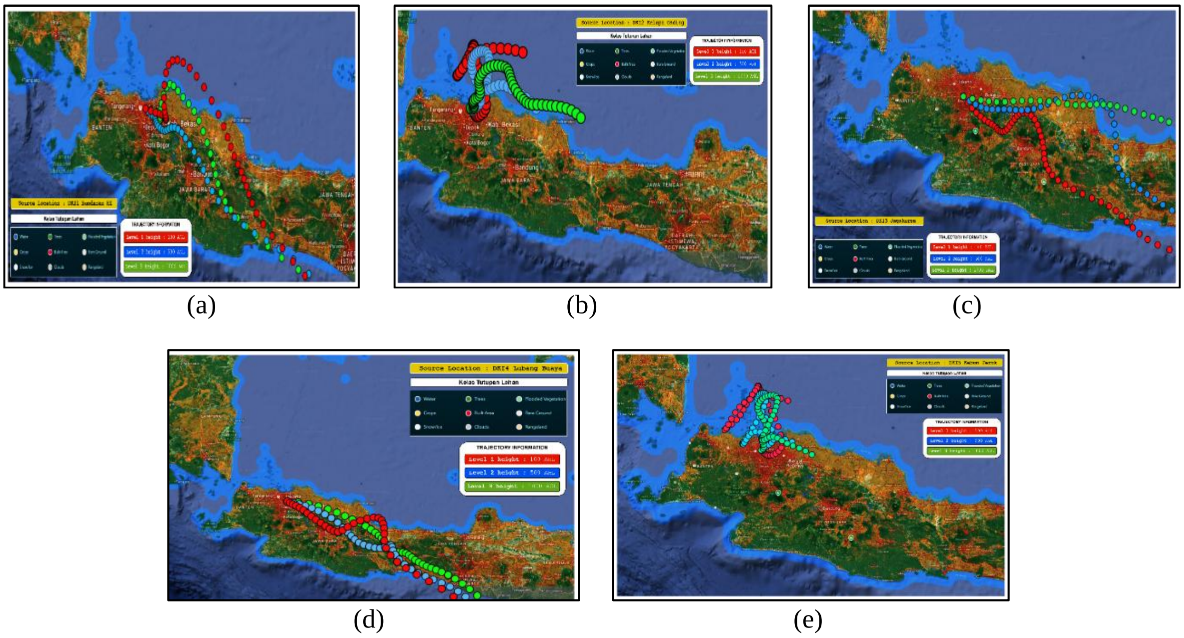

Figure 4. Overlay land cover Esri Sentinel-2P with HYSPLIT backward trajectory: (a) DKI1 Bundaran HI Station (b) DKI2 Kelapa Gading Station (c) DKI3 Jagakarsa Station, (d) DKI4 Lubang Buaya Station, and (e) DKI5 Kebun Jeruk Station

4. CONCLUSION

This study successfully developed a hybrid framework that integrates deep learning predictions and spatial source analysis to model PM2.5 pollution in Jakarta. The multivariate LSTM model built showed the best predictive performance at the DKI1 Bundaran HI Station (R^2 value of 75.87%) and the highest relative accuracy at the DKI3 Jagakarsa Station (MAPE value of 13.34%). Source analysis using HYSPLIT backward trajectories triggered when PM2.5 concentrations exceeded the threshold 75 $\mu g/m^3$ revealed specific pollutant transport patterns during extreme pollution episodes from January 2022 to December 2024.

Specifically, trajectory analysis identified two main corridors for regional pollutant transport. During pollution episodes at Jagakarsa Station (June 18, 2024) and Lubang Buaya Station (September 28, 2023). Spatial validation using Esri Sentinel-2 land cover maps confirms these findings with surface evidence. For the Lubang Buaya Station, the air-mass transport pathway intersects large built-up regions of the Cikarang–Bekasi industrial corridor, which may represent a potential regional emission influence during the analyzed episode. Meanwhile, for the Kebun Jeruk Station, the trajectory crosses a mosaic of Built Area in Cilegon-Serang and crops (agricultural land) on the North Coast of Banten. The lack of vegetation cover and trees along the advection path and in the Jakarta receptor area indicates low natural absorption capacity (sinks), a condition that is further influenced by subsidence-related meteorological processes (air mass descent) from altitudes exceeding 1000 meters, which promote the accumulation of both regional and local pollutants within the surface layer.

ACKNOWLEDGMENTS

The authors thank the Jakarta Environmental Service for providing pollutant data.

FUNDING INFORMATION

The authors thank the School of Data Science, Mathematics, and Informatics at IPB University for partially supporting the publication costs.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hendro Pratama Saragih	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Imas Sukaesih Sitanggang	✓	✓			✓	✓				✓		✓	✓	✓
Hendra Rahmawan	✓	✓			✓	✓				✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo: **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The datasets, pre-processing scripts, and LSTM configuration files generated and analyzed during the current study are available in the public repository to ensure full reproducibility: https://github.com/hendrosaragih/IPBU_PM25_AnalysisSourcePollution_Jakarta.

REFERENCES

[1] Badan Pusat Statistik, "Statistik lingkungan hidup Indonesia 2024 (Environment statistics of Indonesia 2024)," Jakarta, 2024.

[2] World Health Organization, "WHO global air quality guidelines," 2021.

[3] Dinas Lingkungan Hidup Provinsi DKI Jakarta, "Laporan pemantauan kualitas udara 2023 (Air quality monitoring report 2023)," 2023.

[4] T. Istiana *et al.*, "Causality analysis of air quality and meteorological parameters for PM2.5 characteristics determination: evidence from Jakarta," *Aerosol and Air Quality Research*, vol. 23, no. 9, p. 230014, 2023, doi: 10.4209/aaqr.230014.

[5] Z. Zhang, S. Zhang, C. Chen, and J. Yuan, "A systematic survey of air quality prediction based on deep learning," *Alexandria Engineering Journal*, vol. 93, pp. 128–141, Apr. 2024, doi: 10.1016/j.aej.2024.03.031.

[6] Y. Kim, S.-B. Park, S. Lee, and Y.-K. Park, "Comparison of PM2.5 prediction performance of the three deep learning models: A case study of Seoul, Daejeon, and Busan," *Journal of Industrial and Engineering Chemistry*, vol. 120, pp. 159–169, Apr. 2023, doi: 10.1016/j.jiec.2022.12.022.

[7] S. Pongpiachan *et al.*, "Combined use of principal component analysis/multiple linear regression analysis and artificial neural network to assess the impact of meteorological parameters on fluctuation of selected PM2.5-bound elements," *PLOS ONE*, vol. 19, no. 3, p. e0287187, Mar. 2024, doi: 10.1371/journal.pone.0287187.

[8] B. Liu *et al.*, "Evaluating urban and nonurban PM 2.5 variability under clean air actions in China during 2010–2022 based on a new high-quality dataset," *International Journal of Digital Earth*, vol. 17, no. 1, Dec. 2024, doi: 10.1080/17538947.2024.2310734.

[9] Y. Yang *et al.*, "Estimation of PM2.5 concentration across China Based on multi-source remote sensing data and machine learning methods," *Remote Sensing*, vol. 16, no. 3, p. 467, Jan. 2024, doi: 10.3390/rs16030467.

[10] S. Gündoğdu and T. Elbir, "A data-driven approach for PM 2.5 estimation in a metropolis: random forest modeling based on ERA5 reanalysis data," *Environmental Research Communications*, vol. 6, no. 3, p. 035029, Mar. 2024, doi: 10.1088/2515-7620/ad352d.

[11] G. Cican, A.-N. Buturache, and R. Mirea, "Applying machine learning techniques in air quality prediction—A Bucharest city case study," *Sustainability*, vol. 15, no. 11, p. 8445, May 2023, doi: 10.3390/su15118445.

[12] A. Ayu Oktaviana, A. Fadli Assomadi, and J. Hermana, "Analysis correlation of meteorological factors with PM 2.5 concentrations in forecasting air quality of the city of Jakarta," *IOP Conference Series: Earth and Environmental Science*, vol. 1448, no. 1, p. 012006, Feb. 2025, doi: 10.1088/1755-1315/1448/1/012006.

[13] J. Handschuh, T. Erbertseder, and F. Baier, "Systematic evaluation of four satellite AOD datasets for estimating PM2.5 using a random forest approach," *Remote Sensing*, vol. 15, no. 8, p. 2064, Apr. 2023, doi: 10.3390/rs15082064.

[14] B. Mitra *et al.*, "Exploring spatial and temporal dynamics of red sea air quality through multivariate analysis, trajectories, and satellite observations," *Remote Sensing*, vol. 16, no. 2, p. 381, Jan. 2024, doi: 10.3390/rs16020381.

[15] M. I. Shiddiq, I. S. Sitanggang, and M. A. Agmalaro, "Temporal prediction model for CO and CO_2 pollutants using long short-term memory," in *Proceedings of the 2020 International Conference on Computer Science and Its Application in Agriculture (ICOSICA)*, 2020.

[16] I. J. Junior, J. B. Tukay, A. Okrah, G. Magara, and D. J. Masunga, "Precipitation nowcasting in Dar es Salaam: Comparative analysis of LSTM and bidirectional LSTM for enhancing early warning systems," *Journal of Geoscience and Environment Protection*, vol. 13, no. 04, pp. 327–342, 2025, doi: 10.4236/gep.2025.134018.

[17] D. Lopez-Betancur *et al.*, "Evaluation of optimization algorithms for measurement of suspended solids," *Water*, vol. 16, no. 13, p. 1761, Jun. 2024, doi: 10.3390/w16131761.

[18] H. A. D. Nguyen *et al.*, "Long short-term memory Bayesian neural network for air pollution forecast," *IEEE Access*, vol. 11, pp. 35710–35725, 2023, doi: 10.1109/ACCESS.2023.3265725.

[19] O. Kaya and M. Utku Kalay, "Spatio-temporal forecasting of bus arrival times using context-aware deep learning models in urban transit systems," *IEEE Access*, vol. 13, pp. 161423–161435, 2025, doi: 10.1109/ACCESS.2025.3609530.

[20] X. Liu, “Enhanced prediction of karst spring discharge using a hybrid LSTM-XGBoost model optimized with grid search,” *Machine Learning with Applications*, vol. 22, p. 100740, Dec. 2025, doi: 10.1016/j.mlwa.2025.100740.

[21] L. Liu, J. Liang, L. Ma, H. Zhang, Z. Li, and S. Liang, “Gas pipeline flow prediction model based on LSTM with grid search parameter optimization,” *Processes*, vol. 11, no. 1, p. 63, Dec. 2022, doi: 10.3390/pr11010063.

[22] S. Tırınk, “Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM,” *PLOS One*, vol. 20, no. 10, p. e0334252, Oct. 2025, doi: 10.1371/journal.pone.0334252.

[23] X. Bai, N. Zhang, X. Cao, and W. Chen, “Prediction of PM 2.5 concentration based on a CNN-LSTM neural network algorithm,” *PeerJ*, vol. 12, p. e17811, Aug. 2024, doi: 10.7717/peerj.17811.

[24] L. Li *et al.*, “Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke,” *Environment International*, vol. 145, p. 106143, Dec. 2020, doi: 10.1016/j.envint.2020.106143.

[25] M. Ciepiela, W. Sobczyk, and E. J. Sobczyk, “Modelling backward trajectories of air masses for identifying sources of particulate matter originating from coal combustion in a combined heat and power plant,” *Energies*, vol. 18, no. 3, p. 493, Jan. 2025, doi: 10.3390/en18030493.

[26] N. Zaini, L. W. Ean, A. N. Ahmed, M. Abdul Malek, and M. F. Chow, “PM2.5 forecasting for an urban area based on deep learning and decomposition method,” *Scientific Reports*, vol. 12, no. 1, p. 17565, Oct. 2022, doi: 10.1038/s41598-022-21769-1.

[27] G. P. Apangu *et al.*, “Sentinel-2 satellite and HYSPLIT model suggest that local cereal harvesting substantially contribute to peak Alternaria spore concentrations,” *Agricultural and Forest Meteorology*, vol. 326, p. 109156, Nov. 2022, doi: 10.1016/j.agrformet.2022.109156.

[28] C. Saravia and K. Trachte, “Spatiotemporal analysis of NH3 emission sources and their relation to land use types in the Eastern German Lowlands,” *Atmosphere*, vol. 16, no. 3, p. 346, Mar. 2025, doi: 10.3390/atmos16030346.

BIOGRAPHIES OF AUTHORS

Hendro Pratama Saragih received a B.Sc. degree in Informatics from Telkom University, Indonesia, in 2020 and is currently pursuing an M.Sc. degree in Computer Science at IPB University. He specializes in the fields of computational intelligence, optimization, and deep learning, with a particular focus on air quality forecasting using LSTM models. He has an extensive professional background as a data engineer at Bank Mandiri, where he develops generative AI and data, regulatory and compliance systems. His research interests include numerical modeling, having previously developed the MoSV 1D application for wave simulation. He can be contacted at email: hendrosaragih@apps.ipb.ac.id.

Imas Sukaesih Sitanggang received a PhD in Computer Science from the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, in 2013. She is a lecturer in the School of Data Science, Mathematics, and Informatics at IPB University, Indonesia. Her main research interests include spatial data mining and smart agriculture. She can be contacted at email: imas.sitanggang@apps.ipb.ac.id.

Hendra Rahmawan is a faculty member in the Computer Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University (Institut Pertanian Bogor). He received his bachelor’s degree from the Department of Computer Science, IPB University, and his master’s and doctoral degrees from the School of Electrical Engineering and Informatics, Institut Teknologi Bandung. His research interests include high-performance computing, internet of things (IoT), and computer networks. He can be contacted at email: hrahmawan@apps.ipb.ac.id.