Forecasting particulate matter concentration using deep learning and weather data

Ahmad Furqon[1], Santoso Soekirno[1], Naufal Ananda[2], Haryas Subyantara Wicaksana[3], Ahmad Meijlan Yasir[4], Tonny Wahyu Aji[4], Aviv Maghridlo[4], Fadhli Luthfiansyah[4], Marzuki Sinambela[4], Suko Prayitno Adi[4]

[1]Department of Physics, Faculty of Mathematics and Science, University of Indonesia, Depok, West Java, Indonesia
[2]Department of Instrumentation and Calibration, Region II of Meteorological Climatological and Geophysical Agency, Banten, Indonesia
[3]Directorate of Instrumentation Calibration, Agency for Meteorology Climatology and Geophysics, Jakarta, Indonesia
[4]Department of Instrumentation, State College of Meteorology Climatology and Geophysics, Tangerang, Banten, Indonesia

Article Info

Article history:

Received Mar 31, 2026
Revised Jul 13, 2026
Accepted Aug 28, 2026

Keywords:

Air quality
Long short-term memory
Multilayer perceptron
Particulate matter
Weather

ABSTRACT

Particulate matter (PM2.5) concentration is a critical air quality indicator with severe health impacts. Predicting its dynamics is challenging due to complex interactions with meteorological variables. This study evaluates long short-term memory (LSTM) and multilayer perceptron (MLP) networks to identify the most accurate forecasting architecture. We integrated continuous data from automatic weather stations (AWS) and air quality sensors. Data preprocessing involved temporal synchronization, outlier removal, and multivariate imputation by chained equations (MICE). Key input features past particulate matter, temperature, and relative humidity were selected via Pearson correlation and trained using a 24-hour sliding window. Results demonstrate that the single LSTM model outperforms both the MLP and hybrid architectures. The LSTM achieved the lowest root mean square error (RMSE) of 6.862 µg/m³ and a mean absolute percentage error (MAPE) of 37.38%. Although the MLP yielded a slightly higher coefficient of determination (0.810), it exhibited significant magnitude bias by treating sequential data as static features. Ultimately, this research provides a robust, data-driven framework for real-time air quality early warning systems in tropical regions.

Corresponding Author:

Santoso Soekirno
Department of Physics, Faculty of Mathematics and Science, University of Indonesia
Depok, West Java, Indonesia
Email: santoso.s@sci.ui.ac.id

1. INTRODUCTION

Air quality significantly impacts global public health and ecological balance [1]. Among atmospheric pollutants, particulate matter ≤ 2.5 µm (PM2.5) is particularly critical, contributing to approximately one-eighth of global pollutant-related deaths due to severe respiratory and cardiovascular risks [2], [3]. The accumulation and dispersion of PM2.5 are strongly driven by complex, nonlinear meteorological factors such as temperature, relative humidity, and wind speed [4]. Conventional statistical models, such as autoregressive integrated moving average (ARIMA) and linear regression, struggle to capture these dynamic interactions [5]. Similarly, traditional engineering solutions like Kalman filters or standard machine learning algorithms often require rigorous manual feature engineering and assume Gaussian noise distributions. These assumptions rarely hold true in highly volatile real-world atmospheric sensor streams. Consequently, deep sequential neural networks offer a superior engineering alternative by autonomously learning hierarchical temporal features directly from raw multi-sensor data.

In recent years, the development of machine learning methods has provided new opportunities in air quality modeling and prediction. The long short-term memory (LSTM) model is considered optimal in predicting PM2.5 concentrations because it is designed to overcome the problem of vanishing gradients and can retain important information in long time series [6]. In predicting PM2.5, hourly intervals have a root mean square error (RMSE) in the range of 6.3-13.1 μg/m³ [7]. On the other hand, the multilayer perceptron (MLP) is a neural network model that is also effective in handling nonlinear relationships between variables, although with a simpler approach than LSTM. This model can predict PM2.5 concentrations with evaluation metrics in the form of an RMSE of 3.41-8.12 μg/m³ and an MAE of 2.21-4.36 μg/m³ [7].

While previous studies have explored LSTM and MLP architectures for air quality forecasting [8], many simply feed raw meteorological variables into neural networks without rigorously evaluating the necessity of each feature or addressing the temporal irregularities inherent in sensor data. Most existing LSTM-based works rely on standardized, clean datasets from temperate regions and overlook the complex thermohygrometric interactions typical of tropical climates. Furthermore, there is a lack of systematic comparative evaluations that isolate the exact performance trade-offs between sequential LSTM, static feed-forward MLP, and hybrid architectures under identical data pre-processing pipelines. To bridge this critical gap, our study introduces a unified framework that combines temporal synchronization, multivariate imputation by chained equations (MICE), and rigorous Pearson correlation feature selection. We explicitly evaluate how integrating localized automatic weather station (AWS) data optimizes PM2.5 prediction accuracy, providing a reproducible benchmark for tropical air quality modeling.

The specific contributions of this study are threefold: (i) a comparative evaluation of LSTM and MLP architectures for sequential air quality data; (ii) the integration of continuous AWS parameters as predictors; and (iii) the implementation of a 24-hour sliding window approach to capture daily periodicity. The significance of this research lies in providing a highly reliable, data-driven framework for early warning systems, particularly in tropical regions where complex meteorological dynamics strongly dictate pollutant accumulation.

2. METHOD

This section describes the research stages conducted to evaluate the LSTM and MLP algorithms in predicting PM2.5 concentration using meteorological data as input features [9]. The stages of this research method include literature study, data collection, data pre-processing, data imputation, determining predictors, designing input features and model hyperparameters, and evaluating the LSTM and MLP models. The first stage of this research method is a literature study, this stage is carried out by exploring, reading, the research began with a comprehensive literature review of scientific journals and articles. This step focused on establishing the theoretical foundation of PM2.5 dynamics, the underlying mechanics of LSTM and MLP algorithms, and identifying methodological gaps in prior studies utilizing integrated AWS and air quality sensor data.

The data collection phase of this study focused on capturing raw data from various sources to obtain comprehensive information about environmental conditions. In this context, data was collected using two primary instruments: AWS and an air quality sensor. The calibrated AWS instrument served as a source of meteorological data, covering variables such as air temperature and relative humidity. Meanwhile, the air quality sensor served as a source of historical data on PM2.5 and PM10 particulate matter concentrations. The data processing phase then transformed the raw data into a format ready for processing by a machine learning model. These steps included data cleaning to remove corrupted or anomalous data and time synchronization to ensure that the temporal resolution of the meteorological data and the air quality sensor data was aligned [10].

The fourth stage is data imputation, a special step in pre-processing to handle missing values. Sensors often experience technical failures. Missing data is filled using specific statistical techniques (such as linear interpolation or mean imputation) to maintain the continuity of the time series data without having to delete much important information [11]. The fifth stage is determining predictors. This stage is the process of selecting input variables (features) that have the most influence on the prediction target (PM2.5). Input feature selection uses Pearson correlation analysis to identify meteorological parameters that have a significant impact on fluctuations in air pollution concentrations [12].

The sixth stage, input feature and hyperparameter design, is a crucial phase in the technical structure design before the model evaluation process begins. At this stage, input feature design is carried out by determining the time window to determine the duration of historical data, or the number of past hours that will be used as the basis for future predictions. Simultaneously, hyperparameter settings are carried out to optimize model performance by determining key parameters, such as the number of hidden layers, the number of neurons, the learning rate, the batch size, the activation function, the segmentation portion, and the type of optimizer. This aims to obtain the most optimal LSTM and MLP model configurations to produce a high level of prediction accuracy.

The final step is to evaluate the LSTM and MLP models to measure the accuracy performance of both algorithms. The models are tested using previously unseen data. Predicted results are compared with actual values using evaluation metrics. Figure 1 shows a flowchart of the evaluation of the LSTM and MLP models in predicting PM2.5 integrated with AWS meteorological data.

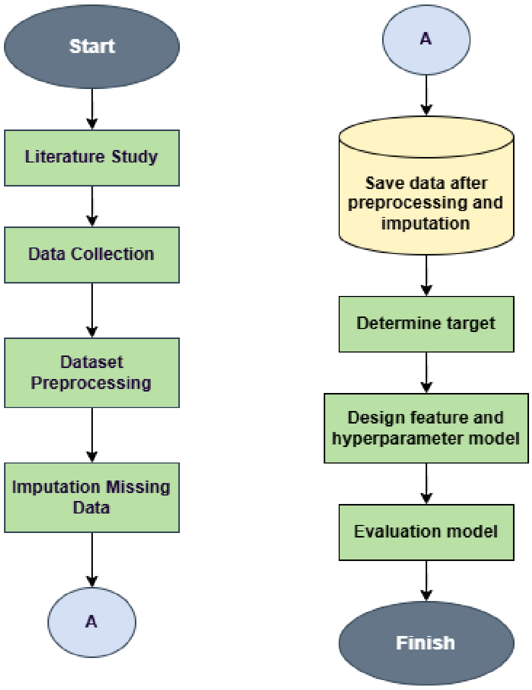

Figure 1. Flowchart research

Based on Figure 1, data collection was carried out by downloading weather and air quality data from AWS loggers and equipment at the Citeko Meteorological Station. The dataset utilized in this study spans an observation period from January 1 to July 20, 2023. Data preprocessing was carried out by synchronizing data and removing outliers. Data synchronization was carried out by matching the sampling intervals of meteorological data, PM10, and PM2.5 every 3 hours to ensure a consistent temporal resolution. Outlier checking was carried out using the range check and step check methods [13]. Empty data and outliers were imputed using MICE [14]. MICE is considered to have advantages in representing multivariate data, reducing bias and providing more stable data distribution after imputation [15].

After the data was filled in temporally, the PM2.5 predictor was determined and the dataset was split, allocating 80% for training data and 20% for testing data. The model configurations were detailed to ensure reproducibility. Specifically, the LSTM architecture consists of two hidden layers with 64 and 32 neurons, respectively. The MLP model was designed with a similar fully connected structure for a fair comparison. Both models utilized the Adam optimizer with a learning rate of 0.001, a batch size of 32, and were trained for a maximum of 100 epochs. The LSTM model hyperparameters were fine-tuned using the early stopping method to prevent overfitting. To ensure robustness and strict reproducibility, the model training process was executed across five independent runs using fixed random seeds in both the NumPy and TensorFlow libraries. The reported evaluation metrics represent the average performance across these multiple runs, thereby validating the stability of the models under varying initializations. After obtaining the optimal hyperparameters, the models were processed for evaluation.

Model performance was quantified using three standard statistical metrics: RMSE, mean absolute percentage error (MAPE), and the coefficient of determination (R^2). RMSE measures the average magnitude of prediction errors, giving higher penalty to large deviations. MAPE assesses forecasting accuracy as an intuitive percentage error. Finally, R^2 evaluates the proportion of variance in PM2.5 concentrations explained by the input features, reflecting the model’s ability to capture general dynamic trends.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_1 - \widehat{y_i})^2}{\sum_{i=1}^{n}(\overline{y_i} - \widehat{y_i})^2} \tag{1}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 \tag{3}$$

3. RESULTS AND DISCUSSION

This section presents the results of the PM2.5 concentration prediction system implementation and an in-depth analysis of the performance of the developed model. The discussion begins with the transformation of raw data into machine learning-ready features, followed by a comparative evaluation of LSTM, MLP, and hybrid LSTM-MLP models, and concludes with a discussion of the implications of the findings for understanding air pollution dynamics based on the research conducted.

3.1. Preprocessing data

Meteorological data sourced from AWS contains noise, outliers, and temporal inconsistencies that can interfere with the model's generalization process. Therefore, data preprocessing is essential, including data cleaning, handling missing data, correlation analysis for feature selection, and data scale normalization [16]. The first step is data integration and cleaning. The raw dataset consists of weather parameters at hourly resolution and air particle concentrations at three-hour resolution. The weather data is then aligned to three-hourly resolution to ensure consistency with the particulate data. Afterward, quality control (QC) procedures are performed as described in the methods section. The outlier detection utilized strict physical thresholds to ensure data validity. For meteorological parameters, data points outside the bounds of 15 °C to 45 °C for temperature and 30% to 100% for relative humidity were flagged as anomalies and removed. Following the removal, the MICE imputation validation was conducted by comparing the statistical distribution of the dataset before and after imputation to ensure the structural integrity of the time-series was maintained without introducing artificial bias. The visualization in Figure 2 shows the plot of temperature and humidity after QC, the three-hour average temperature, and PM2.5 concentration throughout the observation period.

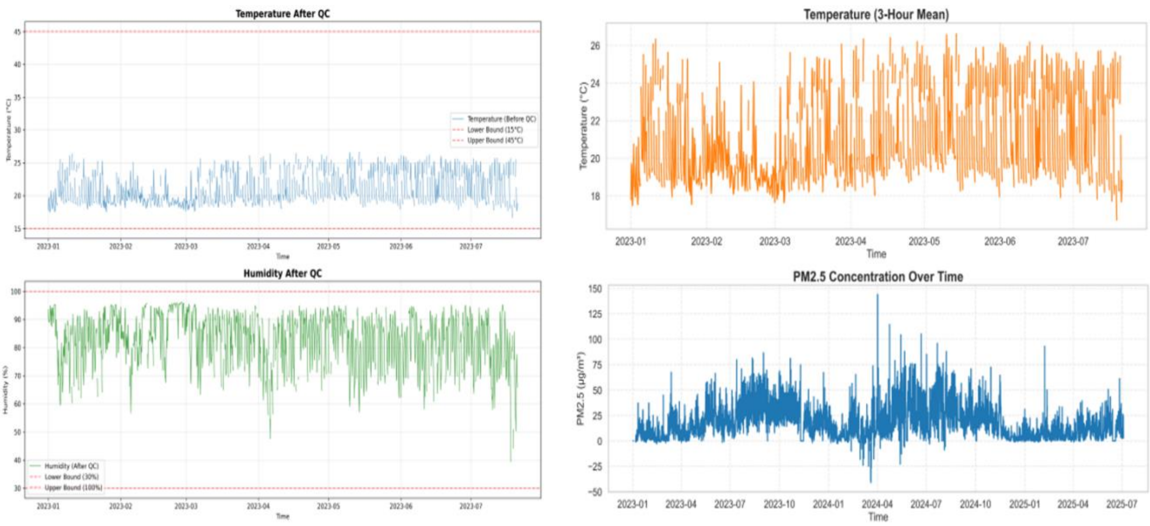

Figure 2. Plotting data humidity, temperature, PM10, and PM2.5

Based on this visualization, the daily fluctuation dynamics are visible with several gaps in missing values after data cleaning and QC. To handle this missing data, the imputation technique using the MICE algorithm was implemented due to its ability to utilize the multivariate correlation structure between atmospheric parameters to estimate the closest values, thus maintaining statistical distribution and minimizing bias in the imputation process [17]. The next stage, feature selection, is carried out to identify input parameters that have a positive contribution to the parameters to be predicted. This is crucial because irrelevant parameters can actually increase the computational burden and the risk of overfitting or underfitting. Pearson correlation matrix analysis was performed to determine the optimal input parameters, with the results shown in Figure 3.

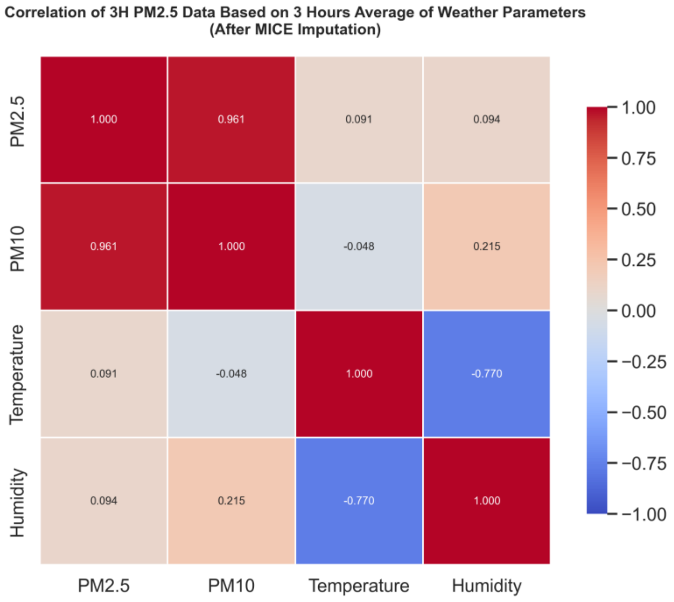

Figure 3. Pearson correlation matrix

Based on Figure 3, the correlation analysis shows a strong statistical relationship between the target variable (PM2.5) and certain parameters. Based on the correlation coefficient (r), it was decided to use only three main features as predictors: PM10, temperature, and humidity. PM10 shows a strong positive correlation with PM2.5, in line with the findings of Zhang *et al.* [17] who conducted a shapley additive explanation (SHAP) analysis of PM10 and PM2.5, which found that PM10 is the dominant predictor for PM2.5, considering that PM2.5 is a physical subset of PM10 with similar emission sources and dispersion patterns. Meanwhile, temperature and humidity also show a significant correlation. This is in accordance with research by Istiana *et al.* [4] which highlights the importance of the interaction between these two variables in tropical regions. The study showed that temperature and humidity have a greater impact on pollutant accumulation than rainfall. Conversely, other variables were eliminated from the input list due to their very weak correlations in the PM2.5 dataset in this study. This elimination of features with low correlations aims to reduce noise that could confuse the model in learning deterministic patterns [18].

After feature selection, data normalization was performed using the min-max scaling technique. The raw data has a highly heterogeneous scale range, where particulate concentrations can reach hundreds of µg/m³, while air temperatures range in the tens of degrees Celsius. This difference in scale causes the model gradient to oscillate and slows convergence. A transformation was applied to map all features to the [0-1] range to ensure that each feature has a proportional weight contribution in the model's loss function [19]. Furthermore, this normalization technique also aims to make the computational process and use of the model's memory capacity more efficient.

The model configuration was designed to capture temporal patterns in a multi-level manner. The first layer used LSTM and MLP with a two-hidden layer structure. Based on empirical experiments, the best hyperparameters were obtained in Figure 4. The final step in data preprocessing was transforming the time series into a format suitable for supervised learning, which was performed using a sliding window technique. This technique utilizes historical data over a specific time period (lookback) to construct an input sequence (X) that is used to predict the target value (y) at the next time step [20]. In this study, the sliding window was set to 24 hours, allowing the model to utilize information from the previous 24 hours to project PM2.5 concentrations one hour into the future. The lookback period takes into account the periodicity of PM2.5 concentration levels that repeat daily.

The sliding window process effectively transforms tabular data into 3D tensors with dimensions (samples, timesteps, features), which are then processed by the LSTM architecture to learn hierarchical temporal dependencies. For MLP models, the input tensors are flattened (reshaped) into a 2D format with dimensions (samples, timesteps x features) to be compatible with fully connected layers. The use of the sliding window technique has the advantage of preserving the temporal relationships present in pollutant data and comprehensively utilizing historical data. Furthermore, this approach also prevents data leakage by ensuring a clear separation between training and test data based on chronological order [21].

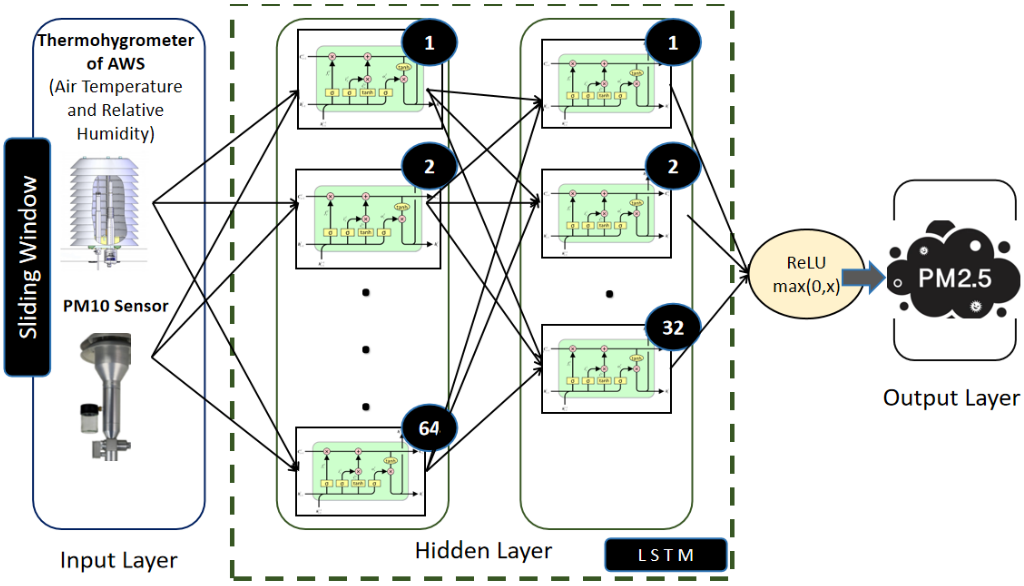

Figure 4. Model architecture

3.2. Evaluation LSTM and MLP model

Model performance evaluation was conducted by comparing the prediction results to a test set representing air pollution conditions the model had never encountered before. Visual analysis was performed sequentially, starting with the hybrid, LSTM, and MLP models, to identify patterns in the responsiveness of each architecture to actual data fluctuations.

The visualization in Figure 5 demonstrates the very solid performance of LSTM model. As illustrated by the alignment of the prediction and actual curves in Figure 5, this model demonstrates a high ability to minimize the difference between predicted and actual values, especially at fast time steps. The short- and long-term memory mechanisms in the LSTM allow it to memorize historical patterns more effectively than other models, resulting in a prediction curve that is quite precise with the actual data.

In contrast, the MLP model exhibits a different trend, as shown in Figure 6. Although its prediction line follows the up-and-down trend very well, there is a noticeable magnitude bias in some segments, where the model systematically predicts slightly higher or lower values (offset). This is understandable considering that MLPs lack a recurrent mechanism to capture sequential dependencies intrinsically, instead treating sliding window inputs as static features [17].

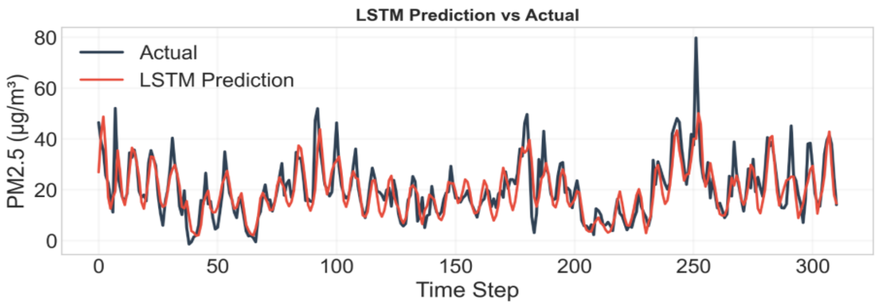

Figure 5. LSTM prediction vs actual

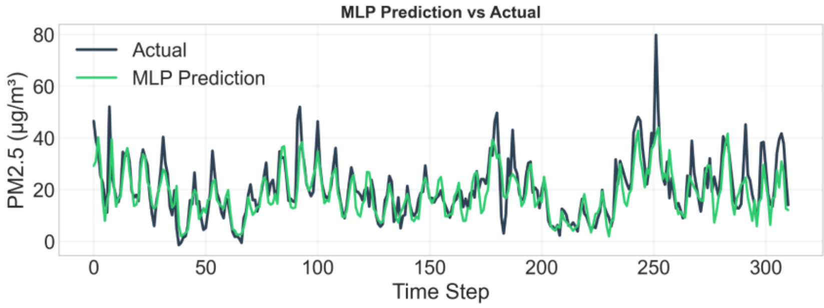

Figure 6. MLP prediction vs actual

Furthermore, the hybrid model evaluation is visualized in Figure 7. Based on Figure 7, the hybrid model demonstrates good generalization, with the predicted line following the general pattern of the actual line. However, there are slight deviations at some extreme peaks, where the model appears to struggle to capture sudden fluctuations with full precision. This indicates that combining the LSTM and MLP architectures as a hybrid model faces complex parameter optimization issues, resulting in its performance being lower than that of the single LSTM model in this study. To evaluate the overall model performance quantitatively, the evaluation matrix as shown in Table 1 was utilized.

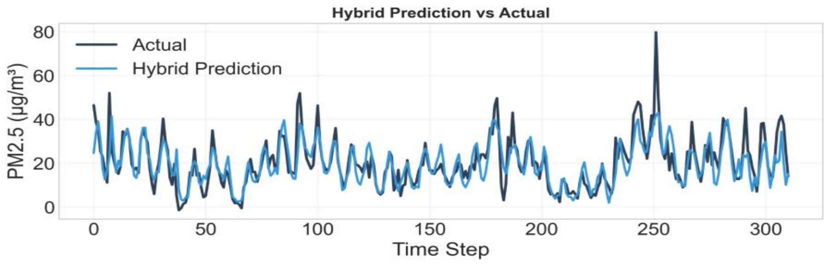

Figure 7. Hybrid LSTM-MLP prediction vs actual

Table 1. Model's metric evaluation

Model	Metric evaluation		
	RMSE	MAPE	R^2
LSTM	6.862	37.384	0.804
MLP	7.004	33.881	0.810
Hybrid	6.921	36.328	0.798

In terms of RMSE, Table 1 shows that LSTM achieved the best performance with a value of 6.862, which is the lowest among the three models studied. This indicates that LSTM has the smallest average absolute error deviation in the original units of pollutant concentration, which means this model produces the most accurate predictions and with minimal errors compared to the hybrid model (6.921) and MLP (7.004). This low RMSE value indicates that LSTM is more effective in handling data variability and provides more consistent results in predicting PM2.5 concentrations. Furthermore, in terms of MSE, LSTM again excels with a value of 47.081, which is lower than hybrid (47.901) and MLP (49.050). MSE gives greater weight to large prediction errors, so a lower MSE value for LSTM indicates that this model produces extreme prediction errors less frequently. LSTM models are better at minimizing large errors, which in turn increases prediction reliability in situations with high data variation. MLP and hybrid, although producing good results, show larger errors in some extreme predictions, which is reflected in the higher MSE values in Table 1.

Finally, for the coefficient of determination (R^2) in Table 1, MLP has a slight advantage with a value of 0.810, higher than LSTM (0.804) and hybrid (0.798). This R^2 value indicates that MLP is better able to follow the ups and downs of the data, reflecting its ability to capture data variability or fluctuations. However, the slightly higher R^2 value of MLP is not enough to overcome LSTM's superiority in terms of absolute prediction accuracy, as reflected in the lower RMSE and MSE values. Therefore, although MLP is better at matching data patterns, LSTM is superior in predicting more accurate values overall. Overall, LSTM was the best model in this study. Although MLP had a marginal advantage in the correlation coefficient (R^2), the difference was not practically significant. In the context of an air quality early warning system, minimizing the magnitude of prediction errors (RMSE and MSE) is far more critical than simply trend correlation. The superiority of LSTM in minimizing these errors aligns with the study by Kim *et al*. [19], which emphasized that recurrent neural network architectures are better able to capture the stochastic dynamics of air pollutants.

3.3. Discussion

This study found the LSTM architecture to be superior in minimizing PM2.5 prediction errors compared to MLP and hybrid models. Theoretically, this superiority is based on the LSTM's internal cell state mechanism, which can store relevant information long-term and forgetting unnecessary noise through a forget gate. This capability is crucial in the context of air pollution data, which is time-series with high autocorrelation. Recurrence-based deep learning approaches such as LSTM are inherently better suited to capturing non-linear temporal dynamics than MLP [21], which treat sliding window inputs as static feature vectors without an explicit time dimension [22]. This phenomenon explains why MLPs, despite having high correlation, fail to achieve the same level of absolute precision as LSTMs.

The observation that the single LSTM architecture outperformed the hybrid LSTM-MLP model provides valuable insights into neural network design for localized environmental time series. Although hybrid models conceptually attempt to leverage the sequential memory of LSTMs alongside the non-linear mapping capabilities of fully connected layers, our results indicate a stability and optimization trade-off. We attribute the hybrid model's lower performance to architectural over-parameterization. In a localized dataset with a 24-hour lookback window, adding dense feed-forward layers after recurrent layers introduces redundant weight matrices that distort the temporal gradients learned by the LSTM cells. This complexity increases the model's susceptibility to overfitting on training noise and creates suboptimal local minima during backpropagation. Consequently, the simpler, dedicated LSTM architecture maintains superior weight convergence and better generalization when predicting unseen extreme pollution peaks [23].

In addition to architectural differences, our findings emphasize the critical role of specific meteorological drivers in predicting PM2.5 dynamics. The empirical selection of air temperature and relative humidity while eliminating weakly correlated variables like wind speed demonstrates that thermal and moisture interactions dominate local particulate behavior in this study area. From an atmospheric physics perspective, high relative humidity accelerates the hygroscopic growth of aerosols and promotes the formation of secondary inorganic particulates. Simultaneously, temperature variations directly influence the height of the atmospheric boundary layer and thermal inversion events, which trap pollutants near the surface. By supplying the LSTM with these continuous thermohygrometric parameters, the network effectively learns the non-linear thermodynamic thresholds that trigger sudden pollutant accumulation, significantly improving accuracy over pure autoregressive (PM-only) historical inputs [24].

From an engineering and practical deployment perspective, the architectural superiority of the single LSTM model extends beyond predictive accuracy to computational efficiency. In real-world air quality monitoring, forecasting models must be integrated into internet of things (IoT) edge devices or centralized cloud platforms that process continuous, high-frequency sensor streams. While massive hybrid architectures demand substantial memory footprints and high computational latency, our optimized LSTM model maintains a compact parameter size after convergence. This lightweight profile enables low-latency real-time inference, making it exceptionally well-suited for deployment on edge artificial intelligence (Edge-AI) microcontrollers embedded within AWS.

Despite its robust performance, this study has limitations that present opportunities for methodological scaling. First, while MICE imputation effectively maintains statistical distribution during missing data intervals [25], artificial imputation inherently carries estimation uncertainty during unprecedented extreme pollution events. Second, the current implementation validates the architecture using a single-station dataset, which does not account for spatial pollutant transport. However, our modular preprocessing, imputation, and sliding-window framework is inherently scalable to multi-station and regional networks. To generalize this methodology across broader geographic areas, the pipeline can be readily extended into spatiotemporal graph neural networks (GNN) or convolutional LSTMs (ConvLSTM). In such scaled architectures, multiple AWS and air quality stations would serve as spatial nodes, allowing the network to simultaneously capture localized temporal dependencies and cross-regional pollutant dispersion patterns.

Based on the above analysis, further research is recommended to explore the integration of attention mechanisms or transformers that can provide dynamic weights at specific time-steps that are most influential, rather than simply assigning uniform weights to the entire sliding window. Furthermore, the development of a spatiotemporal model that integrates multi-station data using a grid-embedded approach is highly recommended to capture cross-regional pollution distribution patterns in complex urban areas [26]. Overall, this study successfully demonstrates that the LSTM architecture with appropriate data pre-processing can provide reliable PM2.5 prediction accuracy for early warning systems. These results provide a strong technical foundation for policymakers to adopt AI-based models to mitigate the health impacts of air pollution.

4. CONCLUSION

This study provided a comprehensive comparative evaluation of deep learning architectures for forecasting hourly PM2.5 concentrations by integrating localized AWS and air quality sensor data. The evaluation demonstrates that the single LSTM model is the most robust and reliable architecture, achieving the lowest error rates with an RMSE of 6.862 µg/m³ and an MSE of 47.081. The superiority of the LSTM lies in its internal gating and cell state mechanisms, which effectively capture complex temporal dynamics and long-term sequential dependencies. Conversely, while the standard MLP model achieved a slightly higher R^2 (0.810), it exhibited a noticeable magnitude bias by treating time-series sequences as static features. Furthermore, combining architectures into a hybrid LSTM-MLP model degraded prediction accuracy due to increased structural complexity and overfitting risks. Our findings also confirm that selecting PM10, air temperature, and relative humidity as core predictors effectively minimizes input noise while capturing essential atmospheric interactions.

To build upon these findings, future research should explore the integration of Attention mechanisms or Transformer-based architectures to dynamically weight critical sequential time steps. Additionally, expanding the methodology to incorporate multi-year datasets and multi-station spatiotemporal networks will further enhance model generalization. Implementing such spatial-temporal frameworks will be crucial for developing scalable, real-time air quality early warning systems in complex urban environments.

ACKNOWLEDGMENTS

We would like to express our gratitude to BMKG and Citeko Meteorological Station for their support in providing data and instrument that have been essential for this research project.

FUNDING INFORMATION

This research was funded by Faculty Mathematics and Natural Science Universitas Indonesia for this work, under project scheme of Hibah Penelitian Pascasarjana with number PKS-27/UN2.F3.D/PPM.00.02/2025.

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
Ahmad Furqon		✓	✓			✓		✓		✓			✓	✓
Santso Soerkirno	✓	✓		✓	✓	✓		✓	✓			✓		
Naufal Ananda	✓	✓				✓				✓	✓		✓	✓
Haryas Subyantara Wicaksana	✓	✓		✓	✓		✓		✓					✓
Ahmad Meijlan Yasir			✓		✓				✓	✓	✓			
Tonny Wahyu Aji			✓		✓	✓	✓		✓		✓			
Aviv Maghridlo			✓		✓		✓		✓		✓			
Fadhli Luthfiansyah	✓			✓		✓				✓	✓			
Marzuki Sinambela		✓				✓	✓	✓		✓		✓		
Suko Prayitno Adi	✓			✓	✓			✓	✓	✓		✓		

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

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author [SS], upon reasonable request

REFERENCES

[1] M. T. Udristioiu, Y. EL Mghouchi, and H. Yildizhan, "Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning," *Journal of Cleaner Production*, vol. 421, p. 138496, 2023, doi: 10.1016/j.jclepro.2023.138496.

[2] I. N. G. Suyasa, N. M. Marwati, and N. K. Rusminingsih, "Measuring Air Quality Over Denpasar City Indonesia in 2021," *Journal of Health and Medical Sciences*, vol. 4, no. 4, Dec. 2021, doi: 10.31014/aior.1994.04.04.191.

[3] G. Maule, "Air Pollution and Interstitial Lung Disease: Insights from the Global Burden of Disease Database," *European Medical Journal*, Mar. 2026, doi: 10.33590/emj/dkvu1711.

[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] Y. Wang *et al*., "A comparative study of three models to analyze the impact of air pollutants on the number of pulmonary tuberculosis cases in Urumqi, Xinjiang," *PLoS ONE*, vol. 18, no. 1, p. e0277314, Jan. 2023, doi: 10.1371/journal.pone.0277314.

[6] S.E. Hasyim, Nurazizah, M.Y. Pratama, U. R. Gurning, B. Khairunnisa, "PM 2.5 Prediction Using the Long Short-Term Memory Algorithm," *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, vol. 2, no, 2, pp. 63-69, 2025.

[7] S. Hoffman and R. Jasiński, "The Use of Multilayer Perceptrons to Model PM2.5 Concentrations at Air Monitoring Stations in Poland," *Atmosphere*, vol. 14, no. 1, p. 96, 2023, doi: 10.3390/atmos14010096.

[8] F. Mohammadi, H. Teiri, Y. Hajizadeh, A. Abdolahnejad, and A. Ebrahimi, "Prediction of atmospheric PM2.5 level by machine learning techniques in Isfahan, Iran," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-52617-z.

[9] S. Soekirno *et al.*, "Design of drought early warning system based on standardized precipitation index prediction using hybrid ARIMA-MLP in Banten province," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, p. 1878, 2024, doi: 10.11591/ijai.v13.i2.pp1878-1887.

[10] Hartanto, S. Humaidi, E. Frida, N. Ananda, and M. Sinambela, "Spatial evaluation rainfall estimation on weather radar using Marshall–Palmer reflectivity–rainfall rate in Banten," *Journal of Water and Land Development*, no. 62, pp. 193–200, 2024, doi: 10.24425/jwld.2024.151567.

[11] M. Perini and M. Nikolic, "In-Database Data Imputation," *Proceedings of the ACM on Management of Data*, vol. 2, no. 1, pp. 1–27, 2024, doi: 10.1145/3639326.

[12] T. Istiana *et al*., "Fine particulate matter concentration forecasting using long short-term memory network and meteorological inputs," *Global Journal of Environmental Science and Management*, vol. 10, no. 4, pp. 1759–1774, 2024, doi: 10.22034/gjesm.2024.04.16.

[13] C. A. Fiebrich, Y. R. Morgan, A. G. Mccombs, P. K. Hall, and R. A. Mcpherson, "Quality assurance procedures for mesoscale meteorological data," *Journal of Atmospheric and Oceanic Technology*, vol. 27, no. 10, pp. 1565–1582, 2010, doi: 10.1175/2010JTECHA1433.1.

[14] G. Akbar, P. Prajitno, Ariffudin, and N. Ananda, "Multivariate Imputation Chained Equation on Solar Radiation in Automatic Weather Station," *Jurnal Penelitian Pendidikan IPA*, vol. 10, no. 7, pp. 3633–3639, 2024, doi: 10.29303/jppipa.v10i7.7679.

[15] R. Pahlepi, S. Soekirno, and H. S. Wicaksana, "Solar Radiation Intensity Imputation in Pyranometer of Automatic Weather Station Based on Long Short Term Memory," *Ultima Computing : Jurnal Sistem Komputer*, pp. 35–40, 2023, doi: 10.31937/sk.v15i2.3348.

[16] K. S. Hernández, D. Nieves, J. S. Pérez-Carrasquilla, P. Montoya, M. D. Zuluaga, and M. Ramírez, "Forecasting PM2.5 concentrations using machine learning approaches: added value of low-cost monitoring and regional modeling," 2024, *Copernicus GmbH*. doi: 10.5194/egusphere-egu24-1214.

[17] Y. Zhang, Q. Sun, J. Liu, and O. Petrosian, "Long-Term Forecasting of Air Pollution Particulate Matter (PM2.5) and Analysis of Influencing Factors," *Sustainability (Switzerland)*, vol. 16, no. 1, p. 19, 2024, doi: 10.3390/su16010019.

[18] D. M. P. W. Dissanayake, R. M. K. T. Rathnayake, and L. L. G. Chathuranga, "Crop Yield Forecasting using Machine Learning Techniques - A Systematic Literature Review," *KDU Journal of Multidisciplinary Studies*, vol. 5, no. 1, pp. 54–65, 2023, doi: 10.4038/kjms.v5i1.62.

[19] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," *Applied Sciences (Switzerland)*, vol. 9, no. 1, p. 204, 2019, doi: 10.3390/app9010204.

[20] D. M. Asad, M. Shahriar, and A. H. Chowdhury, "Load Forecasting for Dhaka City Using RNN, LSTM, and GRU Architectures with Meteorological and Temporal Data," 2025, *IEEE*. doi: 10.1109/ECCE64574.2025.11013779.

[21] B. Eren, İ. Aksangür, and C. Erden, "Predicting next hour fine particulate matter (PM2.5) in the Istanbul Metropolitan City using deep learning algorithms with time windowing strategy," *Urban Climate*, vol. 48, p. 101418, 2023, doi: 10.1016/j.uclim.2023.101418.

[22] Y. Wu, C. Qian, and H. Huang, "Enhanced Air Quality Prediction Using a Coupled DVMD Informer-CNN-LSTM Model Optimized with Dung Beetle Algorithm," *Entropy*, vol. 26, no. 7, p. 534, 2024, doi: 10.3390/e26070534.

[23] Z. L. Thakker and S. H. Buch, "Effect of Feature Scaling Pre-processing Techniques on Machine Learning Algorithms to Predict Particulate Matter Concentration for Gandhinagar, Gujarat, India," *International Journal of Scientific Research in Science and Technology*, pp. 410–419, 2024, doi: 10.32628/ijsrst52411150.

[24] M. A. Boucher, J. Quilty, and J. Adamowski, "Data Assimilation for Streamflow Forecasting Using Extreme Learning Machines and Multilayer Perceptrons," *Water Resources Research*, vol. 56, no. 6, 2020, doi: 10.1029/2019WR026226.

[25] D. Yulizar, D. Aofany, S. Soekirno, N. Ananda, M. A. Prabowo, and S. P. Adi, "Visualization of Rainfall Classification using Rain Gauge based on Website," 2024, *IEEE*. doi: 10.1109/AIMS61812.2024.10512600.

[26] L. K. Kalyanam and S. Katkoori, "Sigmoid-based Neuron Pruning Technique for MLPs on IoT Edge Devices," 2023, *IEEE*. doi: 10.1109/ICPEEV58650.2023.10391875.

BIOGRAPHIES OF AUTHORS

Ahmad Furqon received bachelor of instrumentation engineering at Indonesian School of Meteorology, Climatology and Geophysics in 2016. He worked as technician in Indonesian Meteorology, Climatology and Geophysics Agency since 2017. He is recently studying in Department of Physics, Faculty of Mathematics and Science, University of Indonesia. He can be contacted at email: ahmad.furqon@bmkg.go.id.

Santoso Soekirno is an expertise in the sensors and electronics area and actively lecturing the sensors and electronics instrumentation, he is also busy tutoring the students in completing their final projects, especially those related to the sensors and transducers. He was born in Jogjakarta on 4 April, 1961. He received his bachelor and master degrees from University of Indonesia and his doctor degree from the University of Besancon, France, in the field of the LC components with the dissertation title "Frequency control parameters for oscillator sensor, a study on LC component". He can be contacted at email: santoso.s@sci.ui.ac.id.

Naufal Ananda received Bachelor of Instrumentation Engineering at Indonesian School of Meteorology, Climatology and Geophysics in 2019 and Master of Instrumentation and Control Engineering at Bandung Institute of Technology in 2024. He worked as calibration engineer in Indonesian Meteorology, Climatology and Geophysics Agency since 2020 and also lecture in School of Meteorology, Climatology and Geophysics of Indonesia and Universitas Terbuka. His research field comprises data visualization, website program, data science and meteorological-geophysics instrument. He can be contacted at email: naufal.ananda@bmkg.go.id.

Haryas Subyantara Wicaksana received Bachelor of Instrumentation Engineering at Indonesian School of Meteorology, Climatology and Geophysics in 2019 and Master of Instrumentation and Control Engineering at Bandung Institute of Technology in 2024. He worked as calibration technician in Indonesian Meteorology, Climatology and Geophysics Agency since 2020. His research field comprises machine learning, measurement and calibration for meteorological and geophysics instrument. He can be contacted at email: haryas.wicaksana@bmkg.go.id.

Ahmad Meijlan Yasir is a Diploma IV student in the Department of Instrumentation, State College of Meteorology Climatology and Geophysics, Tangerang, Indonesia. His research interests include IoT-based environmental instrumentation, information system, atmospheric data analysis, and machine learning applications in meteorology and climatology. He can be contacted at email: yasirahmad220504@gmail.com.

Tonny Wahyu Aji is an is an Undergraduate Student in the Diploma IV Instrumentation for Meteorology, Climatology, and Geophysics program at School of Meteorology, Climatology and Geophysics. His research focuses on instrumentation and IoT systems for environmental monitoring, applications of artificial intelligence in data analysis, and technological innovations to support meteorological and air quality observations. He can be contacted at email: tonny.wahyuaji@gmail.com, tonny.wahyu.aji.student@stmkg.ac.id.

Aviv Maghridlo is an undergraduate student in the Diploma IV Instrumentation for Meteorology, Climatology, and Geophysics program at the School of Meteorology, Climatology, and Geophysics. His interests include environmental instrumentation, IoT-based monitoring systems, basic applications of physics and networking, and introductory machine learning and deep learning for data analysis. He focuses on practical technologies to support meteorological and environmental observations. He can be contacted at email: aviv1.qwerty@gmail.com.

Fadhli Luthfiansyah earned a Bachelor in Instrumentation for Meteorology, Climatology, and Geophysics from the State College of Meteorology, Climatology, and Geophysics (STMKG) in 2019 and a Master's degree in Information Technology from the Faculty of Computer Science, University of Indonesia, in 2024. He worked as IT specialist in Indonesian Meteorology, Climatology and Geophysics Agency (BMKG) since 2020 and also lecture in STMKG. He is actively engaged in digital transformation initiatives, IT consulting services, and the development of technology education. He can be contacted at email: fadhli.luthfiansyah@bmkg.go.id.

Marzuki Sinambela affiliated with the State College of Meteorology, Climatology, and Geophysics (STMKG), where he conducts research in the fields of physics and computational instrumentation. He earned his Master's degree in Electrical Engineering and his Ph.D. in Physics Science from the University of North Sumatera. His research focuses on the application of advanced machine learning and deep learning techniques, particularly signal processing–based classification algorithms, data science, and AI-driven computational modeling, to address complex problems in meteorology, climatology, and geophysics. His academic interests include instrumentation for atmospheric and geophysical observations and data-driven modeling of environmental systems. He can be contacted at email: sinambela.m@gmail.com.

Suko Prayitno Adi finished Associate degree of Geophysicsin 1993 and Bachelor degree of Applied Statistic at Universitas Terbuka UPBJJ–UT Jakarta in 1992. He continued his study in Sam Ratulangi Manado University, and reached Master degree of Regional Development science in 1998.He received Doctoral degree of Public Administration from Universitas Negeri Makasar (UNM). He has specialist research: applied meteorological and climatological data scince in regional development, and implementation of tsunami risk and mitigation. Recently, he is a lecturer and dean of STMKG (Meteorological, Climatological and Geophysics Academy of Indonesia). He can be contacted at email: suko.prayitno.adi@stmkg.ac.id.