VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

MULTIVARIATE GRU FOR PM2.5 MULTI-HORIZON FORECASTING IN JAKARTA: FEATURE INTERACTION ANALYSIS AND BASELINE EVALUATION

Wawan Gunawan[1*], Saruni Dwiasnati[1], Yudo Devianto[1], Dwiki Jatikusumo[1], Jordy Lasmana Putra[2], Rachmat Adi Purnama[2]

Universitas Mercu Buana, Jakarta, Indonesia[1]
www.mercubuana.ac.id[1]
wawan.gunawan@mercubuana.ac.id[*], yudo.devianto@mercubuana.ac.id,
saruni.dwiasnati@mercubuana.ac.id, dwiki.jatikusumo@mercubuana.ac.id

Universitas Bina Sarana Informatika, Jakarta, Indonesia[2]
www.bsi.ac.id[2]
jordy.jlp@bsi.ac.id, rachmat.rap@bsi.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License.

Submitted: 20 May 2026, **Revised:** 07 November 2026, **Accepted:** 09 November 2026, **Published:** 20 November 2026

Abstract— *Air pollution in Jakarta remains a critical environmental and public health issue, particularly due to elevated PM2.5 concentrations. This study develops and evaluates a multivariate Gated Recurrent Unit (GRU) forecasting framework combined with simple interaction and ablation analysis for multi-horizon PM2.5 prediction in Jakarta, Indonesia. The dataset consists of 5,538 daily observations from five air quality monitoring stations in DKI Jakarta during 2010-2024, covering six pollutant variables: PM10, PM2.5, SO_2, CO, O_3, and NO_2. The preprocessing pipeline includes missing value handling, chronological data splitting, MinMaxScaler normalization fitted only on training data, and sliding window transformation with a 30-day lookback. Three baseline models were used for comparison: Naïve persistence, Moving Average-7, and Vector Autoregression (VAR). Results show that VAR remains superior at t+1, achieving RMSE = 11.875 and MAE = 9.321, while GRU obtains RMSE = 12.708 and MAE = 9.930. The Diebold-Mariano test confirms VAR's significant advantage at t+1 (DM = -3.6008, p = 0.00032). However, GRU consistently outperforms VAR from t+2 to t+7, with the strongest improvement at t+5, reducing RMSE by 5.23% and MAE by 6.31%. Significant GRU superiority is observed at t+3, t+5, and t+6. Ablation analysis indicates that several engineered features introduce redundancy or noise. These findings suggest that VAR is suitable for immediate forecasting, while GRU provides stronger medium-range PM2.5 prediction capability for Jakarta.*

Keywords: *Gated Recurrent Unit, interaction analysis, Jakarta air quality, multivariate time series forecasting, PM2.5 prediction.*

INTRODUCTION

Air pollution in Jakarta constitutes a severe environmental and public health crisis. According to the Centre for Research on Energy and Clean Air (CREA), PM2.5 concentrations across the Jabodetabek metropolitan area ranged from 30 to 55 μg/m^3 in 2024 [1], representing six to eleven times the World Health Organization (WHO) annual guideline of 5 μg/m^3. Based on previous research related to sentiment analysis, pollution in Jakarta has given negative results of 75.2% based on tweets on the platform X [2]. Indonesia was ranked 15th most polluted country globally in 2024 [3], with the Indonesia Air Quality Life Index (AQLI) estimating a potential gain of more than 2 years in life expectancy for Jakarta residents if PM2.5 were reduced to WHO levels [4]. The concentration of

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

PM2.5 and PM10 particulates often exceeds the WHO recommended threshold, increasing the risk of cardiovascular disease, respiratory disorders, and even premature death [5], [6]. Air-pollution dynamics in Jakarta are strongly associated with meteorological conditions, including humidity, rainfall, wind speed, temperature, and sunshine duration, with pollutant-specific relationships reported for particulate matter, SO_2, and O_3 [7]. These relationships interact with Jakarta's broader urban and anthropogenic pollution context [8]. The health burden is substantial, quantified over 9,700 premature deaths annually attributable to PM2.5 in Jakarta, along with more than 5,000 hospitalizations and 6,100 childhood stunting cases per year, with an estimated economic burden equivalent to 2.2% of Jakarta's gross regional domestic product [5]. These statistics underscore the urgent need for robust data-driven prediction systems to support proactive policy interventions.

Multiple modeling approaches have been applied to air quality forecasting. Statistical models such as ARIMA [9] and VAR [10] offer interpretability but are limited in capturing non-linear temporal dependencies, most studies focus on prediction accuracy without analyzing the structure of the relationships between variables. Last research applied an ANN with five parameters to forecast SO_2, CO, PM10, O_3, and NO_2 in DKI Jakarta, achieving RMSE values of 5.474, 8.392, 9.477, 18.250, and 5.171 respectively[11], although the results are good, the research is still carried out separately for each pollutant variable and has not explicitly emphasized multivariate interactions. The RF-LSTM-Stacking model for single PM2.5 prediction in Beijing-Tianjin-Hebei achieved R^2=0.95 and RMSE=7.74 $\mu g/m^3$, increasing accuracy by up to 17% [12], but have not yet modeled the simultaneous relationships between pollutants.

The ST-iTransformer model successfully outperformed LSTM, CNN-LSTM, and other Transformer models for spatiotemporal prediction in Beijing [13], but requires a more complex architecture and greater data and computing support. Recent studies have increasingly adopted deep learning architectures for PM2.5 forecasting because of their ability to capture nonlinear and complex temporal dependencies. A recent review by Zhou et al. [14] highlights the growing use of recurrent, convolutional, attention-based, and hybrid deep learning architectures for PM2.5 concentration prediction. Recent models, including LSTM/BiLSTM-based spatiotemporal architectures [15], GRU-based approaches [16], [17], and transformer-based architectures [13], have demonstrated superior performance for complex time series. Deep learning has also demonstrated strong potential for PM2.5 forecasting, including decomposition-assisted deep learning frameworks developed for urban air-quality prediction [18]. Beyond conventional single-output forecasting, recent studies have also investigated multi-output and spatiotemporal neural forecasting. Previous study proposed a neural-network-based multi-output spatiotemporal approach for air-pollution forecasting, demonstrating the importance of jointly considering temporal and spatial dependencies [19].

Nevertheless, such approaches have primarily been evaluated outside Jakarta and do not specifically address pollutant-interaction diagnostics and statistically validated multi-horizon PM2.5 forecasting in the Jakarta context. However, most prior work on Jakarta's air quality employs *univariate* prediction targets, lacks multi-step ahead evaluation beyond three horizons, and does not systematically validate statistical significance of model comparisons. Furthermore, no existing study provides a structured *multivariate* GRU baseline coupled with explicit inter-pollutant interaction analysis for Jakarta's multi-station data across an extended historical period (2010–2024).

This study addresses these gaps through four specific contributions: (1) no prior study has established a structured multivariate diagnostic baseline for PM2.5 in Jakarta, a tropical megacity with distinct pollution dynamics driven by monsoon seasonality and vehicular-dominated emissions. This study presents a curated, validated multi-station PM2.5, NO_2, SO_2, CO, and O_3 dataset from five DKI Jakarta monitoring stations spanning 2010-2024 with full preprocessing provenance; (2) A structured statistical baseline suite (Naïve, MA-7, VAR) as rigorous performance benchmarks; (3) existing studies rarely employ formal statistical hypothesis testing to validate model superiority; this work applies the Diebold-Mariano (DM) and Wilcoxon test as the primary validation instrument, A multivariate GRU model evaluated on seven prediction horizons (t+1 to t+7) using four error metrics and statistically validated; and (4) A simple interaction analysis framework comprising lagged cross-correlation, rolling correlation, and ablation study to identify inter-pollutant temporal relationships and feature importance. A diagnostic dimension absent from most GRU air quality studies generating actionable insights for future feature engineering in tropical urban contexts.

Comparison of previous studies with the proposed study can be seen in the Table 1.

CC BY NC

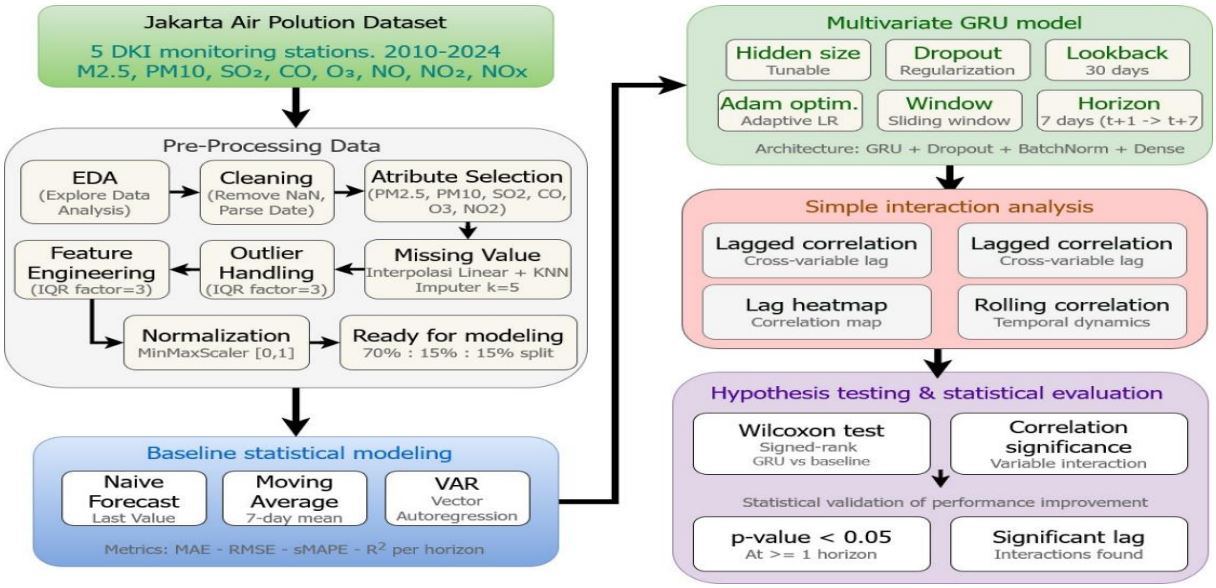
VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Table 1. Research Gaps

Study	Model	Location	Multi-horizon	DM	Ablation	Inter-pollutant
Anggraini et al. (2022) [11]	ANN	Jakarta	✗	✗	✗	✗
Karyadi et al. (2022) [17]	GRU/LSTM	Indonesia	✗	✗	✗	✗
Nie et al. (2024) [16]	GRU/LSTM	General	✗	✗	✗	✗
Zhang & Awang (2025) [13]	ST-iTransformer	Beijing	✓	✗	✗	✗
Proposed study	Multivariate GRU	Jakarta	✓ (7)	✓	✓	✓

Source: (Research Results, 2026) Based on [11], [13], [16], [17].



Source: (Research Results, 2026)

Figure 1. Research Workflow for Jakarta Air Pollution Forecasting Using the Multivariate GRU Model and Statistical Evaluation of Variable Interactions

MATERIALS AND METHODS

Gated Recurrent Unit (GRU) is a variant of RNN designed to address the vanishing gradient problem in long time series data. GRU has two main gates, namely the reset gate and the update gate, which regulate the flow of information efficiently with lower computational complexity compared to LSTM. Comparison of LSTM, GRU, and VARIMA for air quality time series prediction shows that GRU provides competitive performance with higher efficiency [20]. Karyadi et al. (2022) compared LSTM, Bidirectional LSTM, and GRU for air quality prediction in Indonesia [17], while Zhu et al. (2026) proposed a multi-scale spatiotemporal air quality forecasting model to capture spatial and temporal dependencies for AQI prediction [21].

Interactions between pollutants have become a growing research topic. Wu et al. (2025) showed that meteorological variables have different impacts on air quality across urbanization gradients [22]. The simple interaction analysis approach used in this study uses three complementary methods: (1) Pearson lag correlation to identify temporal relationships; (2) rolling correlation to detect time-varying interaction dynamics; and (3) ablation study to measure the relative contribution of each pollutant variable to the accuracy of GRU predictions. This approach is non-causal, and it does not claim causality, but rather provides initial clues about the structure of temporal relationships between pollutants that can serve as the foundation for a more in-depth causal analysis.

The research workflow was systematically designed to develop an air quality prediction model based on a Multivariate GRU using the Jakarta Air Pollution Dataset, which contains monitoring data from five DKI Jakarta air quality stations over the period 2010-2024. The research process begins with data preprocessing, including exploratory data analysis, data cleaning, attribute selection, missing value imputation, outlier handling, feature engineering, normalization, and data splitting for model development. Subsequently, baseline

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

statistical models, namely Naive Forecast, Moving Average, and Vector Autoregression (VAR), are implemented as comparative benchmarks. The processed dataset is then used to develop the Multivariate GRU model by considering several key parameters, including hidden size, dropout, lookback period, optimizer, sliding window, and forecasting horizon. To further investigate the relationships among air pollution variables, a simple interaction analysis is conducted using lagged correlation, lag heatmap, and rolling correlation.

Model hyperparameters were selected through a combination of literature-guided initialization and empirical validation on the training set. The hidden unit sizes of 64 and 32 for GRU Layer 1 and Layer 2 respectively follow the cascade reduction principle commonly adopted in stacked RNN architectures. The lookback window of 30 days was chosen based on the significant autocorrelation structure observed in the data: PM2.5. A dropout rate of 0.2 was selected who recommend rates between 0.2 and 0.5 for recurrent networks. The Adam optimizer was chosen for its adaptive learning rate properties, with the initial learning rate of 0.001. All parameter settings can be seen in Table 2.

Table 2. Parameter setting

Parameter	Value	Reference
GRU Layer 1 units	64	GRU is competitive in air quality data sets with mid-range units. [16]
GRU Layer 2 units	32	stacked GRU Indonesia air quality with cascade reduction ½ [17]
Dropout rate	0.2	rate 0.1–0.3 is optimal for recurrent networks [23]
Lookback window	30 days	The 30-day window captures the sub-monthly cycle of urban PM2.5. [18]
Batch size	128	batch 128 balance between speed and stability [24]
Initial learning rate	0.001	Adam Optimizer recommended [25]
LR reduction factor	0.5	Empirically selected; ReduceLROnPlateau is commonly employed in recent deep-learning air-quality forecasting to adapt the learning rate when validation performance plateaus [26]
EarlyStopping patience	25	patience 20–30 is sufficient for air pollution time series [6]
Maximum epochs	100	100 epoch planned [27]

Source: (Research Results, 2026)

Finally, hypothesis testing and statistical evaluation are performed using the Wilcoxon signed-rank test and correlation significance analysis to validate the statistical significance of the model performance improvement and the detected lag-based interactions. The complete research workflow is presented in Figure 1.

1. Study Area and Data Source

Air quality data were obtained from five monitoring stations included: DKI1 (Bundaran HI, Central Jakarta), DKI2 (Kelapa Gading, North Jakarta), DKI3 (Jagakarsa, South Jakarta), DKI4 (Lubang Buaya, East Jakarta), and DKI5 (Kebon Jeruk, West Jakarta), representing the five administrative districts of Jakarta [28], while data for 2022-2024 were directly retrieved from the DKI Jakarta Regional Environmental Management Agency (BPLHD) monitoring portal [29].

The study period spans January 1, 2010, to December 31, 2024, covering 5,538 daily observations after aggregation. Six pollutant variables were selected based on ISPU (Indonesia Air Pollution Standard Index) requirements and convergence across recent literature (2023-2025): PM10, PM2.5 (forecast target), SO_2, CO, O_3, and NO_2. Variables NO, NOx, THC, NMHC, and CH4 were excluded due to redundancy or incomplete coverage across all stations. Although meteorological variables (temperature, humidity, wind speed, precipitation, and atmospheric pressure) are well-established drivers of PM2.5 dynamics, they were intentionally excluded from this study.

First, this study is explicitly framed as a pollutant-only diagnostic baseline study, establishing a reproducible benchmark using consistently available monitoring data before introducing additional data sources. Second, the ablation results independently corroborate this framing: the removal of simplified meteorological proxies consistently improved model performance, suggesting that coarse meteorological representations may introduce noise rather than signal without high-quality measured data. The choice of PM2.5 as the primary forecast target reflects its documented dominance as the most health damaging pollutant in Jakarta [5].

2. Preprocessing Pipeline

The preprocessing pipeline consists of eight sequential steps designed to ensure data integrity, prevent information leakage, and maximise the signal-to-noise ratio of the input features. Figure 1 illustrates the complete workflow.

Step 1: Initial inspection confirmed that all six pollutant variables (PM2.5, PM10, SO_2, CO, O_3, NO_2) exhibit right-skewed, heavy-tailed distributions. Temporal coverage, station completeness, and inter-variable correlation structures were assessed prior to any transformation.

CC BY NC

Step 2: Raw records were standardised to daily temporal resolution via date parsing. Structurally invalid entries (negative concentrations, duplicate timestamps) were removed. Variables NO, NOx, THC, NMHC, and CH_4 were excluded due to cross-station coverage gaps exceeding 40% and documented redundancy with retained variables under ISPU standards.

Step 3: Six pollutant variables, namely PM2.5, PM10, SO_2, CO, O_3, and NO_2, were retained based on their relevance to Indonesia's Air Pollution Standard Index (ISPU) and their frequent use in recent Jakarta air quality studies. PM2.5 was selected as the primary forecasting target because it has been widely reported as one of the most health-damaging pollutants in Jakarta [5]. Syuhada et al. quantified the health and economic impacts of PM2.5 and O_3 exposure in Jakarta, supporting the importance of PM2.5-focused modelling.

Step 4: Missing values were handled using a two-stage strategy differentiated by gap length, following the approach validated for Jakarta urban pollution data [30], [31]:

Stage 1: applied first to short gaps ≤ 7 consecutive days, leveraging the strong temporal autocorrelation of daily pollutant series (PM2.5 lag-1 r = 0.754) to produce smooth, physically plausible estimates.

Stage 2: applied to longer gaps where linear interpolation would introduce systematic bias. KNN was selected over mean, median, and univariate interpolation methods for three reasons.

a. KNN preserves multivariate inter-pollutant correlations that univariate methods discard, a critical advantage given the documented co-emission patterns of PM2.5, CO, and NO_2 in Jakarta's traffic-dominated atmosphere [31].

b. Benchmarked six imputation methods on urban air quality sensor data and found KNN superior on MAE and RMSE metrics when missing rates are below 30%, consistent with the missing rate observed in this dataset [30].

c. Recent comparative studies show that the effectiveness of imputation methods depends on the missing-data rate, variable distribution, and correlation structure of the dataset. KNN imputation provides a computationally efficient multivariate approach by exploiting similarity among neighboring observations, whereas simple mean or median imputation may introduce distributional bias [30], a principle that applies directly to the correlated pollutant variables in this study.

The parameter k = 5 was adopted as the empirically validated optimum in the sensor environmental literature: values below k = 3 are sensitive to local outliers, while values above k = 20 converge towards mean imputation, negating KNN's local-structure advantage[30].

Step 5: Outliers were identified using the IQR method with a multiplier of **k = 3.0** [32], corresponding to Tukey's (1977) *outer fence* criterion for "far out" values, explicitly distinct from the standard inner fence of k=1.5. This conservative threshold was chosen for two domain-specific reasons. First, PM2.5 concentration data exhibits well-documented right-skewed, heavy-tailed distributions in which extreme values represent physically valid severe pollution episodes rather than instrument errors [6], [12]. Applying the standard 1.5 threshold would risk discarding genuine haze and pollution spike events, precisely the observations most critical for public health early warning model learning. Second, recent methodological guidance warns against "hypothesis-blind removal" in environmental datasets, where aggressive outlier exclusion can distort model generalisation on rare but representative conditions.

Step 6: Sliding window sequences of length 30 days (lookback = 30) were constructed from the cleaned daily series, producing input tensors of shape [samples × 30 × 6]. The 30-day lookback was selected to capture sub-monthly meteorological cycles, including Jakarta's dominant weekly anthropogenic activity cycle, while remaining computationally tractable.

Step 7: All six pollutant variables were scaled to the range [0, 1] using MinMaxScaler. Critically, the scaler was fitted exclusively on the training partition and applied in transform-only mode to the validation and test sets. This design choice prevents a common form of data leakage in which statistical properties of future data (mean, range) influence the representation of training inputs [33].

Step 8: The dataset was partitioned using a strict forward chronological split without shuffling, yielding a 70%:15%:15% ratio: training (January 2010 – December 2019, n = 3,876 sequences), validation (January 2020-December 2022, n = 830 sequences), and test (January 2023-December 2024, n = 832 sequences). Random shuffling was explicitly rejected: time series observations exhibit temporal dependency such that random assignment introduces data leakage by allowing future observations to contaminate the training set, producing artificially optimistic performance estimates that do not generalise to deployment conditions [33].

3. Statistical Baseline Models

Three baseline models were established as performance benchmarks: (1) *Naïve (Persistence)*

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

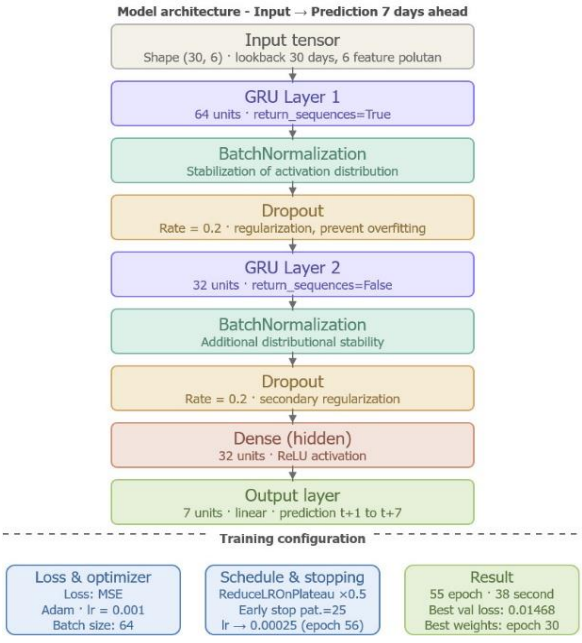
Model: predicts the last observed value for all horizons; (2) *Moving Average-7 (MA-7)*: predicts the mean of the last 7 observed values; and (3) *Vector Autoregression (VAR)*: fits a multivariate linear autoregressive model with lag order selected via the Akaike Information Criterion (AIC), yielding an optimal lag of 7, precisely capturing Jakarta's weekly activity cycle. Prior to VAR model fitting, the Augmented Dickey-Fuller (ADF) test was applied to all six pollutant time series to verify the stationarity assumption. All variables rejected the unit root null hypothesis at the 5% significance level ($p < 0.05$), confirming stationarity of the aggregated daily series. This result validates the direct application of VAR without differencing transformations. VAR predictions on the test set were generated using a rolling one-step-ahead forecasting approach, where the model is re-queried at each time step with the actual historical context. The three models were evaluated over seven prediction horizons (t+1 to t+7) using the MAE, RMSE, sMAPE, and R^2 metrics.

4. GRU Multivariate Architecture

The proposed forecasting model architecture is presented in Figure 2. The model receives a multivariate input tensor with a shape of (30, 6), representing a 30-day lookback window across six air pollutant features (PM2.5, PM10, SO_2, CO, O_3, NO_2), a window length selected to fully capture Jakarta's documented 7-day anthropogenic activity cycle and bi-weekly meteorological regime while remaining computationally tractable [18]. The architecture consists of two stacked GRU layers following the established funnel design pattern, the first GRU layer contains 64 units with return_sequences=True to capture low-level temporal patterns [34], followed by Batch Normalization to improve training stability and accelerate model convergence [35] and Dropout (rate = 0.2) to reduce overfitting [34], the second GRU layer contains 32 units with return_sequences=False, compressing the sequence into a compact temporal representation, which is again stabilized via Batch Normalization and Dropout. This representation is then passed to a hidden Dense layer with 32 units and ReLU activation, chosen for its ability to reduce vanishing gradients and improve convergence in regression tasks [36], before reaching the output layer. The final output layer contains seven linear units corresponding to the simultaneous multi-step prediction horizons t+1 to t+7, implementing a direct MIMO forecasting strategy that avoids error accumulation inherent in recursive approaches [12]. The model was trained using the Mean Squared Error (MSE) loss function, which effectively penalizes large deviations critical for peak pollution events [37], and the Adam optimizer with an initial learning rate of 0.001 and a batch size of 64, consistent with recent GRU-based PM2.5 forecasting studies [38]. A ReduceLROnPlateau scheduler (factor = 0.5, patience = 20) was applied to adaptively reduce the learning rate when validation loss stagnated, triggering at epoch 55 and reducing lr to 0.00025; early stopping with patience = 25 and best-weight restoration was applied to prevent overfitting. As shown in Figure 2, the best validation loss of 0.014684 was achieved at epoch 30, with total training time of 38 seconds [16].

5. Simple Interaction Analysis

Inter-pollutant temporal interactions were analyzed through four complementary methods: (1) *Lagged Pearson Cross-Correlation*: computed between each pollutant variable and PM2.5 target at lags 1-14 days; pairs with $|r| > 0.30$ (moderate-to-strong) were retained; (2) *Correlation Heatmap*: visualizes the full lag feature correlation matrix for structural overview; (3) *Rolling Correlation*: sliding window correlation to detect temporal non-stationarity in inter-pollutant relationships; and (4) *Ablation Study (Leave-One-Out)*: the GRU model was retrained leave-one-feature-out ablation across 17 input features, using a smaller architecture for computational efficiency; the RMSE change (ΔRMSE%) quantifies each variable's marginal contribution to predictive accuracy.



Source: (Research Results, 2026)

Figure 2. Proposed Multivariate GRU Architecture for Seven-Day-Ahead Jakarta Air Pollution Forecasting.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

6. Evaluation Metrics and Statistical Testing

Model performance was evaluated using four metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Symmetric Mean Absolute Percentage Error (sMAPE), and the coefficient of determination (R^2). Two statistical significance tests were employed to formally compare model accuracies: (1) The Wilcoxon signed-rank test (non-parametric), comparing the distributions of absolute errors between GRU and VAR; and (2) The Diebold-Mariano (DM) test, specifically designed for comparing predictive accuracy of two forecasting models on a time series, using squared error loss and Newey-West variance estimation with lag correction. Significance threshold was set at $\alpha = 0.05$.

RESULTS AND DISCUSSION

1. Dataset Characteristics

The final dataset contains 27,690 daily records spanning January 1, 2010, to December 31, 2024. The sliding window transformation produces 3,876 training sequences (shape: 3876×30×6), 830 validation sequences, and 832 test sequences. Zero data leakage was confirmed by verifying no index overlap between training and test sets.

The fifteen-year coverage includes a full COVID-19 lockdown period (2020-2021), which provides valuable variation in anthropogenic emission patterns for model learning. Data leakage verification confirmed a train-test overlap of 0 points, meaning no future information leaked into the training data. After MinMaxScaler normalization, fitted only to the training data, a sliding window transformation yielded an input tensor of size [samples, 30, 6], a 30-day lookback for six pollutant variables. The six variables (PM10, PM2.5, SO_2, CO, O_3, NO_2) were selected based on ISPU standards and literature consensus for 2023-2024, with PM2.5 as the target for prediction due to its most significant health impacts.

2. Statistical Baseline Performance

Table 3. Full Baseline Model Evaluation

Model	Horizon	MAE	RMSE	sMAPE (%)	R^2
Naive	t+1	10.03189	13.09168	14.87316	0.535239
	t+2	12.28478	15.80473	18.15381	0.321007
	t+3	13.11817	16.82273	19.48446	0.230247
	t-4	13.44626	17.18226	20.0537	0.19659
	t+5	13.9567	17.89716	20.80093	0.128791
	t+6	14.09862	18.0979	21.04031	0.108007
	t+7	14.19662	18.45779	21.08033	0.072517
MA-7	t+1	10.235361	13.293427	15.236891	0.520805
	t+2	10.944816	14.085712	16.290634	0.460677
	t+3	11.148293	14.459324	16.619535	0.431338
	t+4	11.356074	14.716731	16.920427	0.410614
	t+5	11.476268	14.95024	17.079395	0.392075
	t+6	11.694958	15.113145	17.372768	0.377965
	t+7	11.883482	15.232423	17.617111	0.368338
VAR	t+1	9.320584	11.874785	13.856128	0.617625
	t+2	10.71066	13.622695	15.846274	0.495551
	t+3	11.14348	14.266977	16.542593	0.446366
	t+4	11.44153	14.612912	17.025076	0.418901
	t+5	11.663	14.930796	17.367228	0.393655
	t+6	11.74523	15.040226	17.481226	0.383953
	t+7	11.77577	15.118837	17.49041	0.377723

Source: (Research Results, 2026)

VAR consistently outperforms both Naïve and MA-7 across all horizons, with the most pronounced advantage at t+1, where VAR achieves RMSE = 11.875 compared to Naïve (RMSE = 13.092, a 9.4% improvement) and MA-7 (RMSE = 13.293, a 10.7% improvement) as shown in Table 3. The AIC-selected lag of 7 is noteworthy: it precisely captures Jakarta's weekly activity cycle, where pollutant concentrations exhibit recurring patterns driven by weekday–weekend emission differences. All R^2 values for Naïve and MA-7 remain positive across all horizons, indicating that both baseline models retain some predictive capability, though progressively weaker as the horizon extends.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

3. GRU Multivariate Forecasting Performance

Table 4. GRU vs. VAR (Best Baseline) Comparison Across All Horizons

Horizon	RMSE (VAR)	RMSE (GRU)	Δ RMSE (%)	MAE (VAR)	MAE (GRU)	Δ MAE (%)	R^2 (GRU)
t+1	0,51875	0,5493056	-7.03	09.32	0,4395833	-6.54	00.56
t+2	0,5847222	13.49	01.01	0,4659722	10.47	02.25	00.51
t+3	14.27	0,5909722	0,19375	11.14	0,4625	04.38	00.49
t+4	0,6256944	14.37	0,0986111	11.44	11.06	03.37	00.44
t+5	0,6479167	14.13	05.23	0,5041667	0,48125	06.31	00.46
t+6	15.04	14.27	0,2236111	0,5097222	11.12	05.33	00.45
t+7	15.12	14.46	04.00	0,5125	11.34	0,1743056	00.43

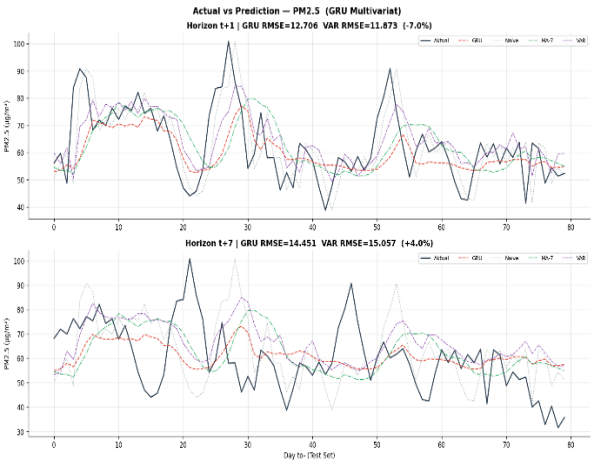
**Negative Δ% at t+1 indicates GRU underperforms VAR at the shortest horizon. From t+2 to t+7, positive Δ% confirms consistent GRU superiority, with the strongest improvement at t+5 (RMSE: 5.23%, MAE: 6.31%).*
Source: (Research Results, 2026)

Experimental results show in Table 4 that the multivariate GRU model generally outperforms VAR across most prediction horizons, although VAR remains superior at the very short-term horizon. At horizon t+1, GRU produces an RMSE of 12.708, while VAR achieves a lower RMSE of 11.875, resulting in a negative improvement of -7.03%. The same pattern is observed in MAE, where GRU records 9.930 compared to VAR's 9.321, corresponding to -6.54%. This indicates that VAR is still more effective for one-day-ahead PM2.5 forecasting, likely because very short-term air pollution dynamics are strongly influenced by immediate linear temporal dependence and recent pollutant persistence.

However, from t+2 to t+7, the GRU model consistently produces lower RMSE and MAE than VAR. The improvement becomes more evident at intermediate horizons, particularly at t+5, where GRU achieves an RMSE of 14.134 compared to VAR's 14.931, yielding the highest RMSE improvement of 5.23%. At the same horizon, GRU also reduces MAE from 11.663 to 10.927, corresponding to the highest MAE improvement of 6.31%. This suggests that GRU's nonlinear sequence-learning capability becomes more beneficial when the prediction horizon extends beyond immediate temporal dependence, allowing the model to capture more complex pollutant dynamics.

Overall, the results indicate that GRU is less competitive than VAR for very short-term prediction, but becomes more effective for medium-range forecasting from t+2 to t+7. The average improvement shows that GRU reduces RMSE by approximately 2.34% and MAE by approximately 2.96% compared with VAR. Meanwhile, the R^2 values gradually decline from 0.561 at t+1 to 0.431 at t+7, indicating that prediction becomes increasingly difficult as the forecasting horizon becomes longer. Nevertheless, the relatively stable R^2 values above 0.43 across all horizons suggest that the GRU model still maintains meaningful predictive capability for multi-day PM2.5 forecasting.

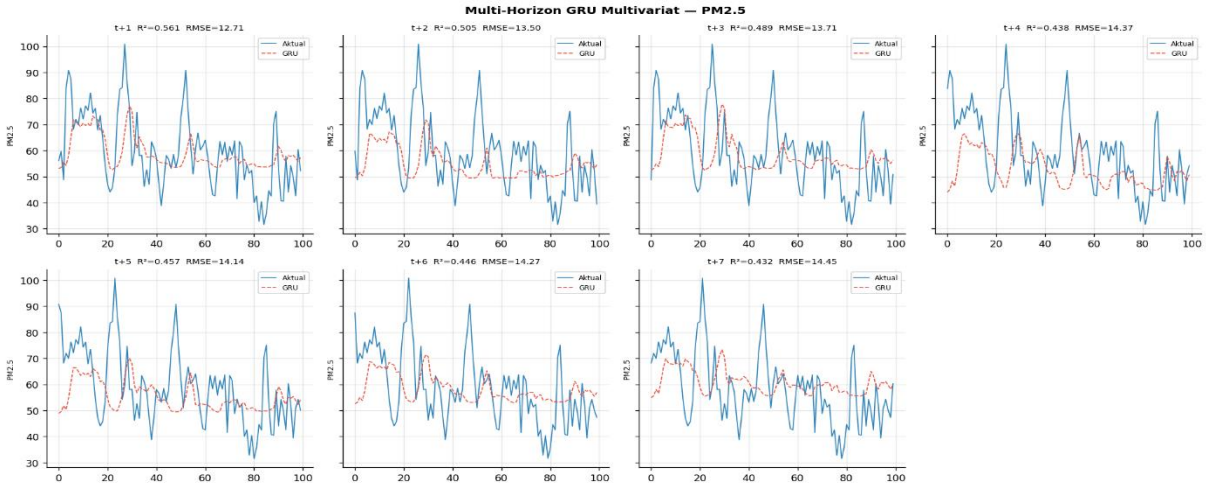
Figure 3 shows that at horizon t+1, the GRU model generally follows the actual PM2.5 trend but still underestimates several sharp peaks, especially around days 4-6, 26-29, and 51-53. This explains why VAR performs better at t+1, with a lower RMSE of 11.873 compared to GRU's 12.708. The Naïve model is more reactive but often shifted, while MA-7 produces smoother predictions because it relies on averaging. At horizon t+7, GRU becomes more competitive and provides a more stable medium-term prediction pattern. Although it still misses some extreme peaks, GRU achieves a lower RMSE than VAR, namely 14.455 versus 15.057, indicating a 4.00% improvement. This suggests that VAR is stronger for very short-term prediction, while GRU is more useful for longer multi-step PM2.5 forecasting.



Source: (Research Results, 2026)

Fig. 3. GRU vs. Baseline Predictions for PM2.5 on the Test Set (Top: t+1; bottom: t+7). Actual Values (Solid Blue); GRU (Red Dashed); Naïve (Grey Dotted); MA-7 (Green Dash-Dot).

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.



Source: (Research Results, 2026)

Fig. 4. Multi-Horizon GRU Forecast Performance for PM2.5. R^2 Values for Horizons t+1 Through t+7 Illustrate Degradation and the Anomalous Recovery at t+5.

Figure 4 illustrates the multi-horizon forecasting behavior of the multivariate GRU model for PM2.5. The R^2 value gradually decreases from t+1 (R^2 = 0.561, RMSE = 12.71) to t+7 (R^2 = 0.431, RMSE = 14.46), indicating that prediction uncertainty increases as the forecasting horizon becomes longer. However, the decline is relatively stable, showing that the GRU still maintains meaningful predictive capability across seven days. A small recovery appears at t+5, where RMSE decreases to 14.13 and R^2 improves to 0.457 compared with t+4 (R^2 = 0.439). This suggests that the GRU partially captures medium-range temporal patterns, possibly related to weekly or working-day emission cycles in Jakarta. Across all horizons, the GRU predictions tend to follow the general trend but underestimate sharp PM2.5 peaks, reflecting a mean-regression behavior commonly observed in sequence models.

4. Simple Interaction Analysis Results

Table 5. Top-10 Significant Lag Pairs (From 42 Total, |r| > 0.30)

Rank	Feature	Lag	r Pearson	Physical Interpretation
1	PM2.5	1	0,502083	Strong one-day autocorrelation indicates that today's PM2.5 concentration is highly influenced by yesterday's concentration. This reflects short-term pollutant persistence caused by carry-over emissions, limited dispersion, and stable atmospheric conditions.
2	PM10	1	4,65625	Strong lag-1 correlation between PM10 and PM2.5 suggests that coarse and fine particles are closely linked through shared emission sources, such as traffic, road dust, construction activity, and combustion processes.
3	PM2.5	2	4,245139	The two-day lag correlation shows that PM2.5 pollution episodes can persist for more than one day. This indicates that fine particulate matter may remain suspended under weak wind, limited rainfall, or stable boundary-layer conditions.
4	PM10	2	0,404167	PM10 remains strongly associated with PM2.5 at a two-day lag, confirming that particulate pollution from common sources can persist beyond daily fluctuations and contribute to short-term accumulation.
5	PM2.5	3	3,909028	The moderate-to-strong correlation at lag-3 indicates continued PM2.5 persistence, although the influence begins to weaken as meteorological dispersion, rainfall, and daily emission variability become more dominant.
6	PM10	3	3,753472	PM10 still shows a meaningful relationship with PM2.5 after three days, suggesting that particle-related pollution episodes may last several days, especially when atmospheric dispersion is limited.
7	PM2.5	4	3,669444	The four-day PM2.5 autocorrelation indicates medium-range temporal persistence. This suggests that pollution episodes in Jakarta may not dissipate immediately and can be sustained by repeated emissions and unfavorable weather conditions.
8	PM2.5	5	3,552778	The five-day correlation suggests a possible working-week emission pattern, where PM2.5 concentrations are influenced by recurring human activities such as commuting, traffic density, and industrial or commercial operations.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

Rank	Feature	Lag	r Pearson	Physical Interpretation
9	PM10	4	3,530556	PM10 at lag-4 remains moderately correlated with PM2.5, indicating that particulate pollution from dust, traffic, and combustion sources may have multi-day effects on fine particle concentration.
10	PM10	5	3,435417	The five-day PM10 relationship with PM2.5 suggests that shared particle sources may contribute to a weekly or working-day pollution cycle. Although weaker than shorter lags, the correlation remains physically meaningful.

Source: (Research Results, 2026)

Table 6. Ablation Study Results for Feature Contribution Analysis (Normalized RMSE)

Configuration	Removed Feature	RMSE	ΔRMSE (%)	Interpretation
Full model (all features)	None	0.192	0	*This is the baseline configuration using all 17 features. The model produces an RMSE of 0.192 and MAE of 0.151, which becomes the reference point for evaluating the impact of removing each feature.*
Without SO2	SO2	0.1914	-0.29	*Removing SO2 slightly improves the model, reducing RMSE by 0.29%. This suggests that SO2 contributes limited useful information for PM2.5 prediction.*
Without CO	CO	0.1905	-0.79	*Removing CO improves RMSE by 0.79%. Although CO is related to combustion emissions, its direct contribution to PM2.5 forecasting may be weak or redundant with other pollutants.*
Without PM25_lag2	PM25_lag2	0.1899	-1.11	*Removing the 2-day PM2.5 lag improves RMSE by 1.11%. This indicates that lag-2 information may overlap with other lag or rolling features and does not add strong independent predictive value.*
Without PM25_rstd14	PM25_rstd14	0.1895	-1.29	*Removing the 14-day rolling standard deviation improves performance by 1.29%. This suggests that long-window volatility may introduce noise rather than useful temporal information.*
Without PM25_rmean14	PM25_rmean14	0.1893	-1.37	*Removing the 14-day rolling mean improves RMSE by 1.37%. This means that the two-week average may over-smooth PM2.5 dynamics and reduce the model's ability to capture daily variation.*
Without month_cos	month_cos	0.1888	-1.63	*Removing the cosine-based monthly seasonal feature improves RMSE by 1.63%. This suggests that calendar-based seasonality may be too simple to represent real atmospheric seasonal effects.*
Without PM25_lag1	PM25_lag1	0.1885	-1.8	*Removing the 1-day PM2.5 lag improves RMSE by 1.80%. This is unexpected because lag-1 is usually important, but in this model it may be redundant with the sequential input structure already learned by GRU.*
Without O3	O3	0.1876	-2.26	*Removing O3 improves RMSE by 2.26%. This indicates that ozone may have a nonlinear or indirect relationship with PM2.5 and may not improve prediction when used directly.*
Without NO2	NO2	0.1874	-2.37	*Removing NO2 improves RMSE by 2.37%. Although NO2 is related to traffic emissions, its relationship with PM2.5 may depend on atmospheric chemistry and meteorological conditions not fully represented in the model.*
Without PM2.5	PM2.5	0.1874	-2.38	*Removing the PM2.5 feature improves RMSE by 2.38%. This may indicate redundancy because PM2.5 information is already represented through lag and rolling features.*
Without PM25_lag14	PM25_lag14	0.1874	-2.4	*Removing the 14-day PM2.5 lag improves RMSE by 2.40%. This suggests that a two-week historical value is too distant to provide stable predictive information for daily PM2.5 forecasting.*
Without PM10	PM10	0.1868	-2.72	*Removing PM10 improves RMSE by 2.72%. Although PM10 and PM2.5 may share emission sources, PM10 may introduce different particle dynamics and reduce model generalization.*
Without PM25_rmean7	PM25_rmean7	0.1861	-3.04	*Removing the 7-day rolling mean improves RMSE by 3.04%. This suggests that weekly averaging may smooth out important short-term PM2.5 fluctuations.*
Without PM25_rstd7	PM25_rstd7	0.1856	-3.33	*Removing the 7-day rolling standard deviation improves RMSE by 3.33%. This indicates that short-term volatility may add instability rather than useful forecasting information.*
Without is_hujan	is_hujan	0.1855	-3.39	*Removing the rainy-season dummy improves RMSE by 3.39%. This suggests that a simple binary rainy-season indicator is too coarse to represent actual rainfall and wet deposition processes.*
Without month_sin	month_sin	0.1851	-3.56	*Removing the sine-based monthly seasonal feature improves RMSE by 3.56%. This confirms that simplified seasonal encoding may not adequately capture real meteorological variation in Jakarta.*
Without PM25_lag7	PM25_lag7	0.184	-4.15	*Removing the 7-day PM2.5 lag gives the largest improvement, reducing RMSE by 4.15%. This indicates that weekly PM2.5 memory may be unstable or misleading for this dataset, and the model performs best when this feature is excluded.*

Source: (Research Results, 2026)

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

The ablation results shown in Table 6 indicate that the full-feature configuration does not provide the most optimal predictive performance. The removal of several variables reduces RMSE, suggesting that some engineered and pollutant-based features may introduce redundancy or physically unstable information into the GRU model. The largest improvement is obtained when PM2.5_lag7 is removed, implying that a fixed weekly PM2.5 memory is not sufficiently robust to represent the daily air pollution dynamics in Jakarta.

This condition is physically plausible because PM2.5 concentration is strongly influenced by short-term emissions, meteorological dispersion, rainfall events, and atmospheric stability rather than by a strict seven-day recurrence pattern. Seasonal variables such as month_sin and is_hujan also reduce model performance, indicating that calendar-based seasonality is too coarse to represent actual atmospheric processes. Therefore, the results emphasize that feature engineering for PM2.5 forecasting should not only increase the number of predictors but also ensure that each feature has strong physical relevance to emission processes, pollutant persistence, and meteorological mechanisms.

5. Feature Importance via Ablation Study

The ablation analysis used a leave-one-feature-out strategy across the 17-feature engineered input space. Starting from the full-feature configuration, the GRU model was retrained by removing one feature at a time, and the resulting change in RMSE relative to the full model was used to quantify the marginal predictive contribution of each feature. Table 6 shows the impact of removing each feature on the model's RMSE in normalized scale (MinMaxScaler output).

All feature-removal scenarios produce negative ΔRMSE values, meaning that removing any single feature improves the model compared with the full 17-feature configuration. This suggests that the full model contains redundant, noisy, or physically weak features. The most problematic feature is PM25_lag7, followed by month_sin, is_hujan, PM25_rstd7, and PM25_rmean7. Therefore, the model would likely benefit from feature selection and a more physically informed feature engineering strategy, especially by replacing simple seasonal indicators with actual meteorological variables such as rainfall, humidity, wind speed, wind direction, temperature, and atmospheric stability. The absence of measured meteorological covariates represents the most significant limitation of this study and the primary direction for future model improvement.

6. Statistical Significance Testing

Two statistical tests were performed to validate the performance differences between the models: the Wilcoxon test and the Diebold-Mariano test (DM Test). Table 7 summarizes the results.

Table 7. Statistical Significance Tests (GRU vs. VAR)

Horizon	W-stat	p-val (W)	DM Stat	p-val (DM)	DM Interpretation
t+1	188073	0.99	-3.60	0.01	VAR is significantly better
t+2	158957	0.05	0.49	0.63	There is no significant difference
t+3	151876	0.01	2.39	0.02	GRU is significantly better
t+4	162481	0.14	0.71	0.48	There is no significant difference
t+5	149339	0.01	2.19	0.03	GRU is significantly better
t+6	151163	0.01	2.16	0.03	GRU is significantly better
t+7	159031	0.06	1.65	0.09	There is no significant difference

Source: (Research Results, 2026)

The Diebold-Mariano test provides a clear statistical comparison between VAR and the multivariate GRU model across forecasting horizons. At the very short-term horizon t+1, the DM statistic is negative and significant (DM = -3.6008, p = 0.00032), indicating that VAR is significantly more accurate than GRU. This result confirms that the linear temporal structure captured by VAR remains more effective for one-day-ahead PM2.5 forecasting. At t+2, t+4, and t+7, the DM test does not detect any statistically significant difference between the two models, as the p-values are above 0.05. This suggests that the predictive performance of VAR and GRU is statistically comparable at these horizons.

In contrast, the GRU model demonstrates statistically significant superiority at intermediate horizons. At t+3 (DM = 2.3999, p = 0.0164), t+5 (DM = 2.1993, p = 0.02786), and t+6 (DM = 2.1609, p = 0.03071), the positive and significant DM statistics indicate that GRU produces significantly lower forecasting errors than VAR. These findings suggest that GRU begins to benefit from its nonlinear sequence-learning capability when the prediction horizon extends beyond immediate short-term dependence. The Wilcoxon signed-rank test supports this pattern, showing significant results at t+3 (p = 0.00409), t+5 (p = 0.00128), and t+6 (p = 0.00299), while no significant improvement is

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

detected at t+1, t+2, t+4, and t+7. Overall, the results indicate that VAR is superior for immediate one-day prediction, whereas GRU becomes statistically more reliable for selected intermediate horizons, particularly t+3, t+5, and t+6.

7. Why Does GRU Outperform VAR Mainly at Intermediate Horizons?

The results indicate that the performance difference between VAR and GRU is strongly horizon-dependent. At the very short-term horizon t+1, VAR remains superior, achieving RMSE = 11.875 and MAE = 9.321, while GRU produces RMSE = 12.708 and MAE = 9.930. This corresponds to a negative improvement of -7.03% in RMSE and -6.54% in MAE, indicating that GRU is less effective for one-day-ahead PM2.5 forecasting. The Diebold-Mariano test also confirms this result, with DM = -3.6008 and p = 0.00032, meaning that VAR is significantly better at t+1. However, the advantage shifts toward GRU from t+2 to t+7. At t+2, GRU slightly improves over VAR with RMSE = 13.495 compared to 13.623, giving a 1.01% RMSE improvement, and MAE = 10.470 compared to 10.711, giving a 2.25% MAE improvement. At t+3, the improvement becomes stronger, with GRU reducing RMSE from 14.267 to 13.714 (3.99%) and MAE from 11.143 to 10.655 (4.38%).

This horizon is also statistically significant based on the DM test (DM = 2.3999, p = 0.0164) and Wilcoxon test (p = 0.00409), showing that GRU is significantly better than VAR. At t+4, GRU still performs better numerically, reducing RMSE from 14.613 to 14.365 (1.82%) and MAE from 11.442 to 11.056 (3.37%). However, the DM test result (DM = 0.7121, p = 0.4764) indicates that this improvement is not statistically significant. The strongest GRU performance appears at t+5, where RMSE decreases from 14.931 to 14.134, producing the highest RMSE improvement of 5.23%. MAE also decreases from 11.663 to 10.927, giving the highest MAE improvement of 6.31%. This result is statistically significant, with DM = 2.1993, p = 0.02786, and is also supported by the Wilcoxon test (p = 0.00128).

The GRU advantage continues at t+6, where RMSE decreases from 15.040 to 14.269 (4.82%) and MAE decreases from 11.745 to 11.119 (5.33%). The DM test confirms that GRU is significantly better at this horizon (DM = 2.1609, p = 0.03071), while the Wilcoxon test also supports the result (p = 0.00299). At t+7, GRU remains better in terms of error metrics, with RMSE decreasing from 15.119 to 14.455 (4.00%) and MAE decreasing from 11.776 to 11.339 (3.71%). However, the improvement is not statistically significant according to the DM test (DM = 1.6557, p = 0.09777). Overall, GRU reduces the average RMSE from 14.210 for VAR to 13.877, corresponding to an average RMSE improvement of approximately 2.34%. It also reduces the average MAE from 11.114 to 10.785, giving an average MAE improvement of approximately 2.96%. These results show that GRU does not fail to outperform VAR; rather, its advantage emerges mainly at intermediate horizons, especially t+3, t+5, and t+6. VAR remains more suitable for immediate one-day prediction because it captures short-term linear dependence efficiently, while GRU becomes more useful when the forecasting horizon extends beyond immediate autocorrelation and requires nonlinear sequence learning. Therefore, VAR and GRU should be interpreted as complementary models, with VAR being stronger at t+1 and GRU providing better medium-range PM2.5 forecasting performance from t+2 to t+7.

CONCLUSION

This study developed and evaluated a multivariate GRU framework, complemented by simple interaction and ablation analysis, for multi-horizon PM2.5 forecasting in Jakarta. Using daily data from five DKI Jakarta monitoring stations (2010-2024) and six pollutant variables, the GRU was benchmarked against Naïve, Moving Average, and VAR baselines across seven forecasting horizons (t+1 to t+7). The comparison was validated with the Diebold-Mariano and Wilcoxon tests. The study shows that model performance depends on the forecasting horizon rather than on a single universal winner. VAR is the more accurate model for immediate one-day-ahead prediction, because very short-term PM2.5 is governed largely by linear persistence. The GRU becomes more advantageous at intermediate horizons, where nonlinear temporal learning yields statistically significant gains at several horizons. The interaction analysis indicates that the temporal structure of PM2.5 is driven mainly by its own recent history and by PM10. The ablation analysis shows that the full engineered feature set is not necessarily optimal, because several features, particularly weekly lag features and simple seasonal indicators, appear to add redundancy or noise rather than useful information.

This study makes four contributions. First, it provides a curated and fully preprocessed multi-station air quality dataset for Jakarta. Second, it establishes a statistical baseline suite as a reference for future work. Third, it presents a statistically validated, horizon-specific comparison between a multivariate GRU and VAR. Fourth, it offers a simple interaction and ablation framework that guides

CC BY NC
Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

future feature engineering. These findings support a hybrid approach to air quality management in Jakarta. VAR can serve immediate daily alerts, while the GRU can support medium-range planning such as health advisories, traffic management, and temporary emission control. Because the framework relies on official monitoring data, it can be integrated into existing environmental monitoring systems. The study has several limitations. It does not use measured meteorological variables, it aggregates data across stations, it uses daily rather than hourly resolution, and its ablation results come from a single run on a smaller architecture. Future research should therefore (1) incorporate meteorological variables, (2) retain station-level spatial resolution, (3) apply ablation-based feature selection, (4) test advanced architectures such as Temporal Fusion Transformer, CNN-GRU, or spatiotemporal Transformers, and (5) extend the data to hourly resolution.

DECLARATION OF AI USE

The authors declare that generative artificial intelligence (AI) tools were used in the preparation of this manuscript. ChatGPT (OpenAI) was used to assist with English language editing and grammar improvement. The authors reviewed and revised all AI-assisted content and take full responsibility for the accuracy, originality, and final content of the manuscript.

ACKNOWLEDGEMENTS

This research was supported by the Research and Community Service Institute (LPPM) of Universitas Mercu Buana, Jakarta, through the Domestic Collaborative Research (KDN) scheme 2025/2026, grant number 190-KDN-2025 and Universitas Bina Sarana Informatika for dataset contribution and collaborative support. Data used in this study were obtained from the Regional Environmental Management Agency (BPLHD) of DKI Jakarta Province.

REFERENCE

[1] CREA, "Indonesia Air Quality 2024: Annual Assessment," CREA — Centre for Research on Energy and Clean Air. Accessed: Apr. 20, 2026. [Online]. Available: https://energyandcleanair.org

[2] L. Hakim, M. V. Dalimunthe, C. Danuputri, and D. Widyaningrum, "Sentimen Analisis Mengenai Polusi Udara Menggunakan Algoritma Support Vector Machine dan Random Forest," *J. Ilm. FIFO*, vol. 15, no. 2, pp. 91–101, 2023, doi: https://doi.org/10.22441/fifo.2023.v15i2.001.

[3] AQLI, "Annual Update 2024," Energy Policy Institute, University of Chicago. Accessed: Apr. 20, 2026. [Online]. Available: https://epic.uchicago.edu/area-of-focus/aqli/

[4] IQAir, "World Air Quality Report 2024," IQAir AG, Goldach, Switzerland. Accessed: Apr. 20, 2026. [Online]. Available: https://www.iqair.com/world-air-quality-report

[5] G. Syuhada *et al.*, "Impacts of Air Pollution on Health and Cost of Illness in Jakarta , Indonesia," *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, pp. 1–14, 2023, doi: https://doi.org/10.3390/ijerph20042916.

[6] S. Sekar and Z. Wei, "Air quality prediction-based big data analytics using hebbian concordance and attention-based long short-term memory," *Sci. Rep.*, vol. 15, no. 1, pp. 1–22, 2025, doi: 10.1038/s41598-025-09508-8.

[7] T. Handhayani, "An integrated analysis of air pollution and meteorological conditions in Jakarta," *Sci. Rep.*, pp. 1–12, 2023, doi: 10.1038/s41598-023-32817-9.

[8] D. S. Soemarko *et al.*, "Linking Jakarta ' s Typical Indonesian Urban Context , Air Pollution , and Child Health," *Open Public Health J.*, vol. 16, pp. 1–6, 2023, doi: 10.2174/18749445-v16-e230831-2023-109.

[9] W. Nengsih, C. Widiasari, P. Madhona, H. C. Lubis, and I. Agus, "Sensor Integration and ARIMA-Based Forecasting in WAQMS for Environmental Monitoring in Riau Province , Indonesia," *Sci. J. Informatics*, vol. 12, no. 3, pp. 555–566, 2025, doi: 10.15294/sji.v12i3.24742.

[10] B. Wu, S. Zhao, Y. Liu, and C. Zhang, "Do meteorological variables impact air quality differently across urbanization gradients ? A case study of Kaohsiung , Taiwan , China," *Heliyon*, vol. 11, no. 2, pp. 1–11, 2025, doi: 10.1016/j.heliyon.2025.e41694.

[11] A. N. Anggraini, N. K. Ummah, Y. Fatmasari, and K. F. Hayati, "Air Quality Forecasting in DKI Jakarta Using Artificial Neural Network," *MATICS J. Ilmu Komput. dan Teknol. Inf.*, vol. 14, no. 1, pp. 1–5, 2022, doi: https://doi.org/10.18860/mat.v14i1.13863.

[12] X. Gao *et al.*, "PM2.5 concentration 7-day

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8526

prediction in the Beijing–Tianjin–Hebei region using a novel stacking framework," *Sci. Rep.*, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-07719-7.

[13] R. Zhang and N. Awang, "A novel ST-iTransformer model for spatio-temporal ambient air pollution forecasting," *J. Big Data*, vol. 12, no. 1, pp. 1–27, Dec. 2025, doi: https://doi.org/10.1186/s40537-025-01150-5.

[14] S. Zhou, W. Wang, L. Zhu, Q. Qiao, and Y. Kang, "Environmental Science and Ecotechnology Deep-learning architecture for PM 2 . 5 concentration prediction : A review," *Environ. Sci. Ecotechnology*, vol. 21, pp. 1–17, 2024, doi: 10.1016/j.ese.2024.100400.

[15] X. Wang *et al.*, "Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD – GAT – BiLSTM," *Sci. Rep.*, pp. 1–12, 2024, doi: 10.1038/s41598-024-68874-x.

[16] Y. N. Ng, H. Y. Lim, Y. C. Cham, M. A. A. Bakar, and N. M. Ariff, "Comparison Between LSTM , GRU and VARIMA in Forecasting of Air Quality Time Series Data," *Malaysian J. Fundam. Appl. Sci.*, vol. 20, pp. 1248–1260, 2024, doi: https://doi.org/10.11113/mjfas.v20n6.3411.

[17] Y. Karyadi and H. Santoso, "Prediksi Kualitas Udara Dengan Metoda LSTM , Bidirectional LSTM , dan GRU," *J. Tek. Inform. dan Sist. Inf.*, vol. 9, no. 1, pp. 671–684, 2022, doi: 10.35957/jatisi.v9i1.1588.

[18] N. Zaini, L. W. Ean, A. N. Ahmed, and M. A. Malek, "PM2 . 5 forecasting for an urban area based on deep learning and decomposition method," *Sci. Rep.*, pp. 1–13, 2022, doi: 10.1038/s41598-022-21769-1.

[19] K. K. R. Samal, K. S. Babu, and S. K. Das, "Multi-output Spatio-temporal air pollution forecasting using neural network approach," *Appl. Soft Comput.*, vol. 126, pp. 1–11, 2022, doi: 10.1016/j.asoc.2022.109316.

[20] Y. Nie, H. Ying, Y. Chyi, M. Aftar, and A. Bakar, "Comparison Between LSTM , GRU and VARIMA in Forecasting of Air Quality Time Series Data," *Malaysian J. Fundam. Appl. Sci.*, vol. 20, pp. 1248–1260, 2024.

[21] L. Zhu, L. Chen, and H. Chen, "Short-term air quality prediction using a multi-scale attention fusion model with 3DIGAT-CBAM-BiLSTM based on spatio-temporal correlation," *Expert Syst. Appl.*, vol. 298, pp. 1–23, 2026, doi: 10.1016/j.eswa.2025.129856.

[22] B. Wu, S. Zhao, Y. Liu, and C. Zhang, "Do meteorological variables impact air quality differently across urbanization gradients ? A case study of Kaohsiung , Taiwan , China," *Heliyon*, vol. 11, no. 2, p. e41694, 2025, doi: 10.1016/j.heliyon.2025.e41694.

[23] A. O˘guz and ¨Omer Faruk Ertu˘grul, "Determining the fullness of garbage containers by deep learning," *Expert Syst. Appl.*, vol. 217, no. December 2022, pp. 1–9, 2023, doi: 10.1016/j.eswa.2023.119544.

[24] Z. Liu, Z. Fang, and Y. Hu, "A deep learning-based hybrid method for PM 2 . 5 prediction in central and western China," *Sci. Rep.*, pp. 1–17, 2025, doi: https://doi.org/10.1038/s41598-025-95460-6.

[25] R. Lan *et al.*, "Multi-scale cross-layer feature interaction GAN for underwater image," *Digit. Signal Process.*, vol. 145, pp. 1–12, 2024, doi: 10.1016/j.dsp.2023.104354.

[26] Furizal, A. Maarif, I. Suwarno, A. Masitha, L. Aulia, and A.-N. Sharkawy, "Real-Time Mechanism Based on Deep Learning Approaches for Analyzing the Impact of Future Timestep Forecasts on Actual Air Quality Index of PM10," *Results Eng.*, vol. 24, no. November, pp. 1–19, 2024, doi: 10.1016/j.rineng.2024.103434.

[27] A. P. Putra and A. M. Harahap, "Implementasi Time Series Forecasting dengan Algoritma LSTM untuk Pemantauan dan Prediksi Produktivitas Kelapa Sawit Berdasarkan Hasil Panen," *JURIKOM (Jurnal Ris. Komputer)*, vol. 12, no. 2, pp. 46–56, 2025, doi: 10.30865/jurikom.v12i2.8495.

[28] T. Pohan, "Air Quality Index in Jakarta." [Online]. Available: https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021. [Accessed: 20 April 2026]

[29] B. D. Jakarta, "Sistem Informasi Kualitas Udara Pemerintah Provinsi DKI Jakarta," BPLHD DKI Jakarta. Accessed: Apr. 20, 2026. [Online]. Available: https://www.lingkunganhidup.jakarta.go.id

[30] J. Lee *et al.*, "Comparison of Models for Missing Data Imputation in PM-2 . 5 Measurement Data," *Atmosphere (Basel).*, vol. 16, pp. 1–18, 2025, doi: https://doi.org/10.3390/atmos16040438.

[31] M. Karmoude, B. Munhungewarwa, I. Chiraira, R. Mckenzie, M. Kumar, and B. Mellado, "Science of the Total Environment Machine learning for air quality prediction and data analysis : Review on recent

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

advancements , challenges , and outlooks," *Sci. Total Environ.*, vol. 1002, no. 1, pp. 1–18, 2025, doi: 10.1016/j.scitotenv.2025.180593.

[32] J. W. Tukey, "Exploratory Data Analysis: How to Check for Outliers." Accessed: Jan. 20, 2026. [Online]. Available: https://datatest.readthedocs.io/en/stable/how-to/outliers.html

[33] S. Albelali and M. Ahmed, "Hidden Leaks in Time Series Forecasting : How Data Leakage Affects LSTM Evaluation Across Configurations and Validation Strategies Salma," 2025.

[34] A. Oğuz and Ö. F. Ertuğrul, "Diagnostic Biomedical Signal and Image Processing Applications with Deep Learning Methods," in *Diagnostic Biomedical Signal and Image Processing Applications with Deep Learning Methods*, K. Polat and S. Öztürk, Eds., in Intelligent Data-Centric Systems. , Academic Press, 2023, pp. 1–40. doi: https://doi.org/10.1016/B978-0-323-96129-5.00003-2.

[35] J. Seo *et al.*, "PM2 . 5 Forecasting at U . S . Embassies and Consulates Worldwide Using NASA Model Powered by Machine," *Adv. Earth Sp. Sci.*, pp. 1–19, 2025, doi: 10.1029/2025EA004210.

[36] N. Zaini, L. W. Ean, A. N. Ahmed, and M. A. Malek, "PM2 . 5 forecasting for an urban area based on deep learning and decomposition method," *Sci. Rep.*, pp. 1–13, 2022, doi: 10.1038/s41598-022-21769-1.

[37] Z. Jiang, Y. Nie, and W. Gong, "Enhanced PM2.5 forecasting via bio-inspired hybrid LSTM optimization for high-precision results," *AIP Adv.*, vol. 15, no. 7, pp. 1–11, 2025, doi: 10.1063/5.0287553.

[38] X. Yue *et al.*, "Novel hybrid data-driven modeling based on feature space reconstruction and multihead self-attention gated recurrent unit: applied to PM2.5 concentrations prediction," *Sci. Rep.*, vol. 15, no. 1, pp. 1–21, 2025, doi: 10.1038/s41598-025-00911-9.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.