

Hybridization Model for Air Pollution Prediction Using Time Series Data

Roni Yunis^{*1}, Andri², Djoni³

^{1,2,3}Informatics Faculty, Universitas Mikroskil; Jl. Thamrin No. 140, Medan

e-mail: ^{*1}roni@mikroskil.ac.id, ²andri@mikroskil.ac.id, ³djoni@mikroskil.ac.id

Abstract

In recent years, data science analysis, particularly time series predictions, has been widely employed across various industrial sectors. However, time series data presents high complexity, especially in seasonal patterns such as monthly, daily, or hourly fluctuations. Irregular fluctuations and external factors increasingly challenge accurate predictions. Therefore, this research proposes a hybrid approach combining SVR-SARIMA, SVR-Prophet, LSTM-SARIMA, and LSTM-Prophet to enhance time series prediction accuracy. This study followed the OSEM methodology approach: gathering data, cleaning data, exploring data, developing models, and interpreting crucial aspects of problem-solving. Seasonal effect predictions indicated a rise in SO₂ and NO₂ during dry and rainy seasons until the next two years (average daily increments of 0.0831 µg/m³ for SO₂ and 0.0516 µg/m³ for NO₂). Estimates suggest a decrease in the order of three particles. The evaluation showed that the SVR model performed better compared to the other three models (RMSE 7.765, MAE 5.477, and MAPE 0.261). The best-performing hybrid model was LSTM-Prophet (99.74% accuracy) with RMSE 12.319, MAE 12.057, and MAPE 0.259 values.

Keywords— time series data, air pollution, OSEM, hybrid, LSTM-Prophet

1. INTRODUCTION

Reducing air pollution levels can lessen symptoms of heart, lung, and acute respiratory disorders such as hay fever, asthma, pneumonia, bronchopneumonia, and others. Air pollution is one of the major environmental dangers to health. According to a 2018 WHO report, 90% of people on Earth breathe contaminated air, with Southeast Asian and Eastern Mediterranean regions having average air pollution levels that are five times higher than WHO guidelines [1]. Numerous things, including burning fossil fuels in power plants, industrial smoke and exhaust, burning agricultural land, and vehicle exhaust emissions, can contribute to air pollution. The degree of urbanization is another element influencing the amount of air pollution.

The WHO provides guidelines with thresholds for major air pollutants that are harmful to health. Particulate matter (PM), ozone (O₃), nitrogen dioxide (NO₂), carbon monoxide (CO), and sulfur dioxide (SO₂) are some of these contaminants [2]. There are two forms of PM: PM_{2.5} and PM₁₀. The size of these two particles sets them apart. PM_{2.5} is less than 2.5 µm in size. Because PM_{2.5} and PM₁₀ particles can enter the lung cavity directly, they are extremely hazardous particles [2]. According to the Air Quality Index, Indonesia's PM_{2.5} and PM₁₀ pollution levels are currently 6.1 times higher than the WHO norm. Jakarta, with an average AQI of 124, ranks among the top 10 most polluted cities in the world as of September 22, 2022, at 16:11. The severity of the situation makes it necessary to apply sophisticated analytical tools like time series analysis to accurately predict and manage pollution levels. A methodical way to identify patterns, trends, and seasonal fluctuations in pollution data across time is using time series analysis. By utilizing this analytical methodology, decision-makers and environmental authorities can lower health risks associated with prolonged exposure to polluted air, attenuate pollution spikes, and make well-informed decisions and focused responses.

This demonstrates the importance of using data science analysis, particularly time series analysis for predictive analysis of periodically recorded data, to better comprehend and investigate events. Since the standard statistical technique cannot be optimized, this scientific approach is necessary for a fuller understanding of the dynamic nature of air pollution and its alterations over time. Numerous researchers have employed a range of machine learning models and techniques, such as Prophet, SARIMA, Vector Autoregressive Model (VAR), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM), to aid in the predictive analysis of time series data. Research contrasting these models, specifically LSTM with VAR, ARIMA, and SVR, has consistently demonstrated LSTM's most accuracy over the other models [3-5]. Furthermore, research focusing on air pollution prediction in California utilizing SVR and a Radial Basis Function (RBF) kernel showcased an impressive accuracy rate of 94.1% [6]. Similarly, in a comparative study between SARIMA and Prophet, the Prophet model exhibited higher accuracy levels [7]. These results demonstrate the effectiveness of sophisticated analytical methods in predicting air pollution levels, particularly when combined with time series analysis and machine learning. By using these approaches, researchers and decision-makers can reduce the harmful consequences of air pollution on the environment and public health by implementing focused treatments and making better-informed decisions.

Some researchers have attempted to suggest hybridized models for prediction in other studies, such as research [8] on the LSTM-ARIMA hybrid model and research [9] on the Prophet-SVR hybrid model, with the hybrid model's accuracy showing a significant value higher than that of the single model test. The findings suggest that hybridized models are more accurate than single models; however, they are all based on time series data and do not attempt to identify seasonal trends within them. Understanding the seasonal trends within time series data is crucial for accurate predictions, emphasizing the need for a comprehensive approach that integrates various models to address the dynamic nature of air pollution. To produce predictions, hybridized models, which combine elements of both linear and non-linear models, can yield results that are more accurate than those of a single model [9]. Non-linear models can be utilized in hybrid model implementations to capture relationships that linear models are unable to capture [10]. Moreover, the latest study emphasizes how well hybrid models capture both linear and non-linear connections in time series data [11]. Hybridized techniques, which incorporate components from both linear and non-linear models, have the potential to produce forecasts that are more accurate than single-model approaches. In hybrid model implementations, non-linear models in particular are essential because they capture complex linkages that linear models could miss [12].

Time series data can be divided into three main categories: residual, trend, and seasonal [13]. Data influenced by periods, such as the wet and dry seasons of the year, or by seasonal characteristics, like the days of the week, months, or quarters of the year, clearly show seasonal trends. We can simplify the model by modeling each component independently. The study intends to 1) examine the trend of air pollution levels and forecast future pollution levels by detecting seasonal patterns; 2) conduct experiments to evaluate and compare the predictive abilities of Prophet, SARIMA, SVR, and LSTM models; and 3) develop a hybridization model that combines the best features of each model to improve their performance when combined. The results of the study can be used as a basis for choices about mitigating the risks brought on by air pollution, especially in large cities like Jakarta.

2. RESEARCH METHODS

2.1. Research Design

By breaking down the data and incorporating seasonal influences, the suggested hybridized model can be utilized to forecast air pollution. The models that were employed with time series data were put through several trials and tests as part of the experimental study design. After normalizing the data, the time series data was decomposed into residuals, trends, and seasonal patterns using the Seasonal Decomposition of Time Series (STL) method. In this study,

the data was divided into seasonal periods, as is common in Indonesia, where seasonal patterns comprise the rainy and dry seasons. The methodology of splitting the dataset into training and testing sets and then breaking the data down into seasonal periods is explained in Figure 1. Using the Prophet, SARIMA, SVR, and LSTM prediction models, prediction analysis is carried out once the data has been prepared.

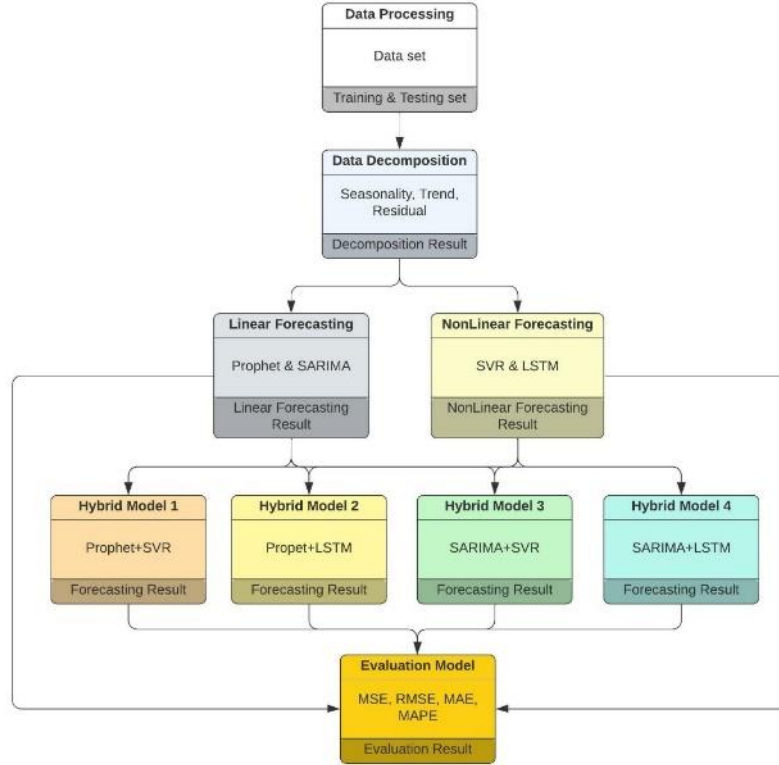


Figure 1. Research Design

Hybridization is the next step after testing each prediction model. The evaluation metrics of mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used for each prediction model.

$$\begin{aligned}
 MSE &= \frac{1}{n} \sum_{i=1}^n (y_i^{\hat{}} - y_i)^2 \\
 RMSE &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{\hat{}} - y_i)^2} \\
 MAE &= \frac{1}{n} \sum_{i=1}^n |y_i^{\hat{}} - y_i| \\
 MAPE &= \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i^{\hat{}} - y_i}{y_i} \right|
 \end{aligned} \tag{1}$$

Where $y_i^{\hat{}}$ is the predicted value and y_i is the actual value.

2.2. Model

This study employed four models that were hybridized to enhance the accuracy of the prediction model in supporting the predictions.

A) Prophet

Facebook developed the Prophet model, which is a member of the generalized additive model [13]. Three elements make up this model: trend, seasonality, and holiday [7].

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (2)$$

where $g(t)$ is a trend function that represents non-periodic variations in the time series' value? A seasonal or changing function, such as a daily, weekly, or annual function, is represented by $s(t)$. The effect of holidays that fall on a potentially erratic schedule for one or more days is represented by $h(t)$, while changes that the model is unable to account for are represented by ϵ_t . With t being normally distributed, the expected value is represented by the value of $y(t)$.

B) SARIMA

Similar to ARIMA, the GAM model seasonal auto-regressive integrated moving average (SARIMA) is applied to time series data that exhibits seasonal characteristics [14]. ARIMA(p, d, q)(P, D, Q), where p, d, q and P, D, Q stand for continuity difference and seasonal auto-regression, respectively, are typically used to describe seasonal expressions [10].

$$y_t = c + \sum_{n=1}^p \alpha_n y_{t-n} + \sum_{n=1}^q \theta_n \epsilon_{t-n} + \sum_{n=1}^P \phi_n y_{t-sn} + \sum_{n=1}^Q \eta_n \epsilon_{t-sn} + \epsilon_t \quad (3)$$

The SARIMA model makes it possible to distinguish data with seasonal frequencies (e.g., 12 months, 24 hours).

C) SVR

Encouragement An expansion of SVM used to solve regression issues is vector regression. To find the optimal function, $f(x)$, the data can be transformed to a higher dimension using the kernel function in SVR [9]. The function $f(x)$ can be modified by adding the kernel function to create the liner equation function shown below:

$$f(x) = \omega^T x + b = \sum_i^n (\alpha_i - \alpha_i^*) \kappa(x_i, x) + b \quad (4)$$

It is possible to use sigmoid, polynomial, radial basis, and linear kernel functions.

D) LSTM

An expanded version of the RNN model, Long Short-Term Memory is linked in a temporal sequence and features an intricate recursive structure. The hidden layer state $H(t)$, which varies over time, and the cell state $C(t)$, which preserves long-term memory, are two crucial characteristics of LSTM. The input gates $I(t)$, forgotten gate $F(t)$, and output gate $O(t)$, which contains the preceding layers $H(t)$ and $C(t)$, define the cell state $C(t)$. The input data and the state of cell $C(t)$ define the state of $V(t)$ [5, 15].

$$\begin{aligned} f_t &= \sigma(w_f[h_{t-1}, x_t] + b) \\ i_t &= \sigma(w_i[h_{t-1}, x_t] + b_i) \\ C_t &= \tanh(w_c \times [h_{t-1}, x_t] + b_c) \\ C_t &= f_t \times C_{t-1} + i_t \times C_t \\ O_t &= \sigma(w_o[h_{t-1}, x_t] + b) \\ h_t &= O_t \times \tanh(C_t) \end{aligned} \quad (5)$$

Where w and b stand for the weight matrix and bias vector, respectively, that were acquired during model training.

E) Hybrid Model

The hybrid model suggested in this study utilized a modified model configuration [10, 16]. It incorporated the weighted average of the two prediction models, as represented below:

$$y = w_1 * f(x) + w_2 * g(x) + e \quad (6)$$

Where:

- y is the value to be predicted (output).
- x is the input or feature.
- $f(x)$ is the result of model 1 (non-linear).
- $g(x)$ is the result of model 2 (liner).
- $w1$ and $w2$ are weights used to determine the contribution of each model.
- e is a constant.

This configuration combined the outputs of the two models and determined the relative contributions of each model to the result by assigning a weight to each model. In this experiment, we used $e = 0$ and assigned a weight of 50% to both $w1$ and $w2$ to ensure a balanced preference for each model. This equal weighting technique tries to prevent bias towards any one model and encourages fairness in the entire evaluation process since each prediction model contributes the same amount.

The selection of models for air pollution prediction is a crucial decision that should be based on the strengths and capabilities of each model in addressing specific challenges. The Prophet algorithm is chosen for its effectiveness in handling data with strong trends and seasonal patterns, as well as its ability to deal with missing or irregularly spaced data [17]. SARIMA is selected as a classical model capable of handling time series data with complex seasonal and trend patterns [18]. SVR is preferred for its capacity to handle nonlinear relationships in data, often encountered in air pollution prediction contexts [19]. LSTM is chosen for its capability to capture complex temporal patterns and nonlinear relationships within time series data, making it suitable for modeling air pollution influenced by multiple factors [20].

Analyzing the strengths and weaknesses of each model reveals that Prophet excels in handling complex trends and seasonality but lacks flexibility with non-periodic patterns [17]. SARIMA performs well with complex seasonal and trend patterns but struggles with data exhibiting unstable or changing trends [18]. SVR is adept at handling nonlinear relationships but can be sensitive to parameter tuning and computationally intensive [19]. LSTM, while effective in capturing complex temporal patterns, is prone to overfitting and requires substantial data for effective training [20].

To enhance the accuracy and stability of air pollution forecasts, a hybridization process is proposed, which combines predictions from Prophet, SARIMA, SVR, and LSTM using ensemble methods or weighting each prediction based on the confidence in the respective model [17]. By leveraging the strengths of each model through hybridization, the accuracy and stability of air pollution forecasts can be improved. In conclusion, the selection of models such as Prophet, SARIMA, SVR, and LSTM for air pollution forecasting is based on their unique strengths and capabilities in handling different aspects of the data. By combining these models through hybridization, it is possible to create a more robust forecasting system that capitalizes on the strengths of each model.

2.3. OSEM Framework

A framework called OSEM can be used to make data analysis easier [21]. The utilization of OSEM in Figure 2 of this study aims to facilitate the appropriate planning and management of the previously specified research design and activities. The steps involved in conducting the research include gathering data, cleaning it, analyzing, and visualizing it, and modeling and interpreting the outcomes of any predictive analysis that has been done.

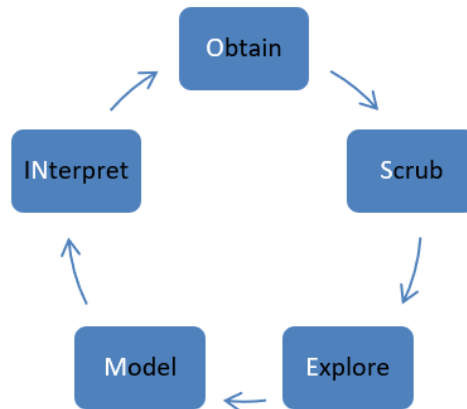


Figure 2. Research Phases

3. RESULT AND DISCUSSION

The findings of the analyses that were conducted following the identified stages of the study were discussed in the section that follows. The following are the outcomes:

3.1. Obtain

The Air Pollution Standard Index (ISPU) data, which is saved in the `dataspku.csv` file, is a time series of data spanning from 2016 to 2021. It is sourced from the DKI Jakarta Provincial Environment Office and is made available through Jakarta Open Data (<https://data.jakarta.go.id>). The data set spans from December 31, 2021, to January 1, 2016. Table 1 displays the 10,960 observation rows and 10 columns that make up the variables of the ISPU dataset. Since the missing value data is quite minimal, the deletion approach is employed to deal with missing data. Numerical types were applied to some particle variable data types.

Table 1. Dataset Variable

Variable	Definition	Type Data	Examples
date	Date of air quality measurement (format: yyyy-mm-dd)	Date	2021-10-01
stasiun	Location of air quality measurements at the station	Character	DKI(Bundaran HI)
pm10	Particulate matter is one of the parameters measured	Numeric	57
so2	Sulfide (in the form of SO ₂) is one of the measured parameters	Numeric	30
co	Carbon Monoxide is one of the parameters measured	Numeric	11
o3	Ozone is one of the parameters measured	Numeric	32
no2	Nitrogen Dioxide is one of the parameters measured	Numeric	38
max	The highest measured value of all parameters measured at the same time	Numeric	81
critical	Parameters whose measurement results are the highest	Character	PM10
category	Categories of air pollution standard index calculation results	Character	SEDANG

3.2. Scrub

The scrub process is used to clean up and modify the data so that the analysis process may be completed correctly based on the features of the variables and data types in the dataset. During this process, the variables `pm10`, `so2`, `co`, `o3`, and `no2` were converted to numeric types, while the date variable was changed to a date type to reflect daily data. Among these variables, several have missing values or are unavailable (NA). Specifically, the `pm10`, `so2`, `co`, `o3`, `no2`,

max, critical, and category variables, totaling 1365 rows, contain NA values. We reduced the original dataset with 10,960 rows to 9595 observation rows and 10 columns, covering the period from January 1, 2016, to December 31, 2021, after removing the NA values. The percentage of data deleted due to NA is 12.47% or 1365 rows of data. Notably, the variables with the most missing values are max and critical, but their absence does not interfere with the analysis process.

3.3. Explore

The prediction model chooses variables after cleansing the data. The Particle data distribution from the obtained dataset is shown in Figure 3. Figure 4 indicates that SO₂ particles rank higher from 2016 to 2021, which directs the emphasis of this study's prediction analysis. High amounts of SO₂ are associated with long-term respiratory health problems, which is in line with the goal of the study. Although PM₁₀ and NO₂ were considered, Figure 3 highlights the persistent presence of SO₂, giving it priority in the analysis.

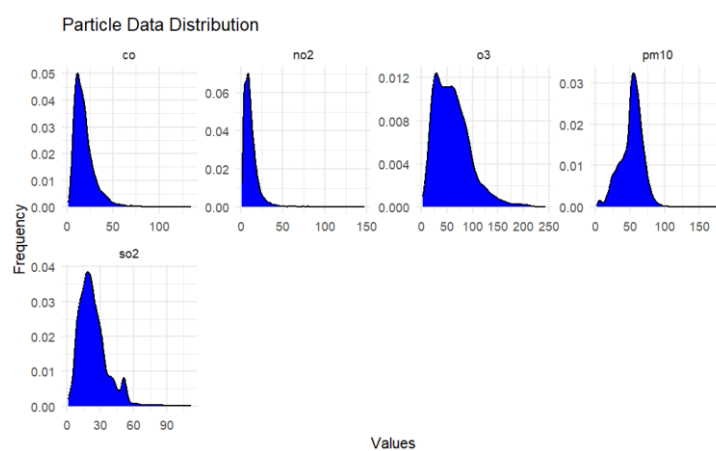


Figure 3. Particle Data Distribution

The categories in Figure 5 help to explain why, from 2016 to 2021, DKI Station 4 (Lubang Buaya) saw the highest concentration of SO₂ particles, with 55973 $\mu\text{g}/\text{m}^3$ in the medium group.

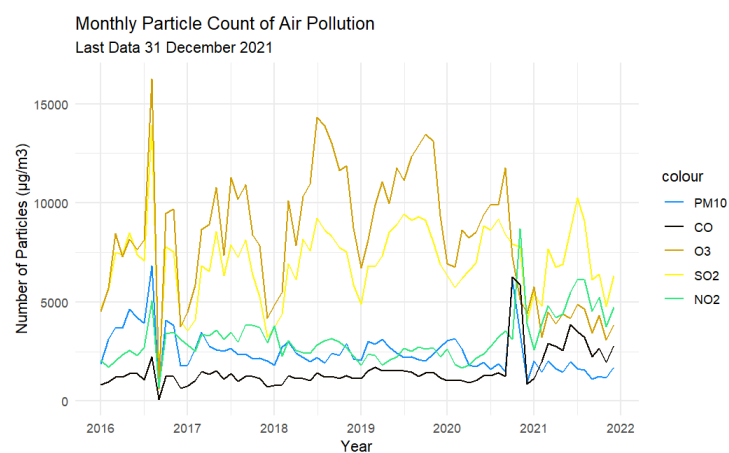


Figure 4. Monthly Particle Count of Air Pollution

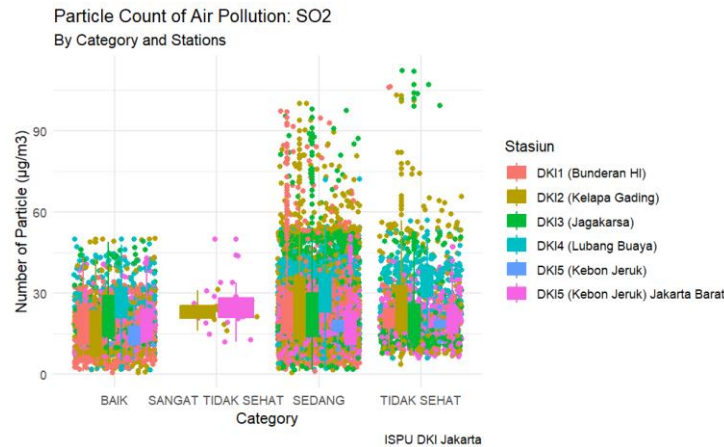


Figure 5. Category of Particle SO_2

This study seeks to examine the seasonal attributes of SO_2 particle concentrations by dividing the data into wet and dry seasons. In Indonesia, the dry season occurs from April to September, while the rainy season spans from October to March, forming the basis for this division. The seasonal data division aims to determine whether the rainy or dry season influences the increase in air pollution particles in Jakarta. The average amount of SO_2 particles in the air from 2016 to 2021, along with the lowest and highest values ($1.19 \mu\text{g}/\text{m}^3$ and $50.87 \mu\text{g}/\text{m}^3$, respectively), and the standard deviation ($11.81 \mu\text{g}/\text{m}^3$), show that the levels are highest from June to September, which could mean that they reach their highest point during the dry season. This study aims to determine whether seasonal fluctuations have an impact on air pollution levels in Jakarta.

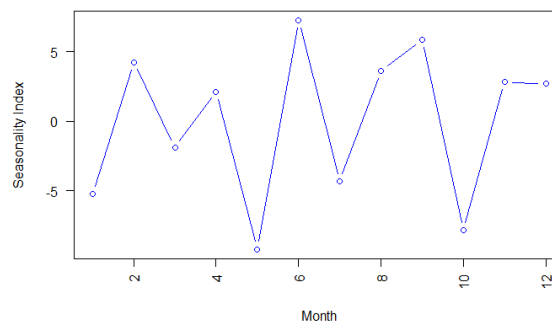


Figure 6. Seasonality Index Particle SO_2

3.4. Model

The earliest stages of the SO_2 particle prediction analysis model experiment involve data segmentation, model implementation, and model evaluation. The dataset is partitioned into training and testing sets in a 70:30 ratio to reduce the likelihood of overfitting during the training of the model. The research utilized the following hardware and software specifications: an Intel(R) Core (TM) i5-10210U CPU @ 1.60GHz, 2.11 GHz, 512GB SSD, NVIDIA GeForce MX130, RStudio/R 2022.12.0.353/4.2.2, and Anaconda 22.11.1. Both the Prophet and SARIMA linear models revealed a non-linear trend in the rise of SO_2 particles over the next two years during the model testing process. Model_4(2,1,1) (2,1,0) emerged as the most superior SARIMA model, achieving a BIC value of 319.68 and an AIC value of 310.53. In the same way, tests using non-linear models like Long Short-Term Memory (LSTM) with 100 epochs and a batch size of 1 and the Support Vector Regression (SVR) model with a Radial Basis Function (RBF) kernel always show that the amount of SO_2 particles is going up. Figure 7 presents the comparison results of a single model that forecasts SO_2 particles.

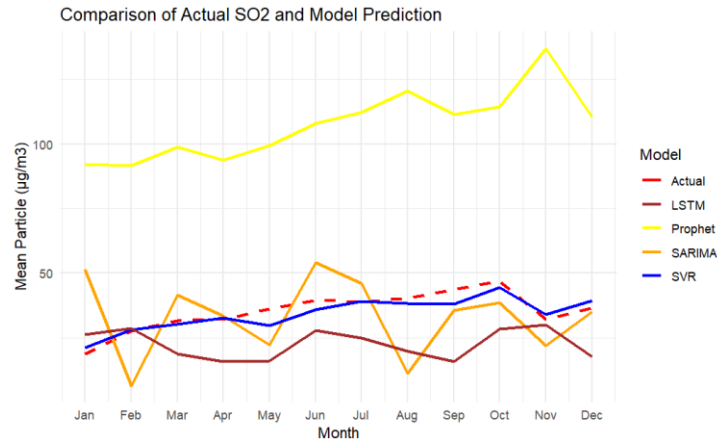


Figure 7. Comparison of Actual SO₂ and Model Predictions

The visualization in Figure 7 demonstrates that the SVR model exhibits predictions that closely align with the real values, whereas the Prophet model displays findings that deviate significantly from the actual values in comparison to SARIMA and LSTM. This demonstrates that SVR exhibits higher accuracy in predicting the actual values.

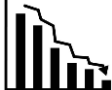

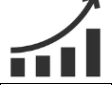
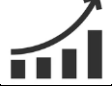
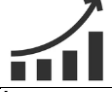
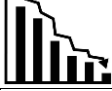
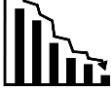
Table 2. Comparison of Single Model Evaluation Metrics

Model	MSE	RMSE	MAE	MAPE
Prophet	2059.89261	45.386040	37.539256	1.7203042
SARIMA	2183.81287	46.731284	41.130228	7.0980860
SVR	60.31069	7.765996	5.477602	0.2607191
LSTM	99.73182	9.986582	9.082258	0.4483081

According to Table 2's assessment, the SVR model demonstrates superior performance compared to the other three models, showcasing lower average MAE and RMSE values. With significantly reduced MSE and RMSE values, the SVR model outperforms others and exhibits a lower prediction error rate. A lower MAE signifies a smaller deviation between expected and actual values, while a reduced MAPE implies a modest prediction error. The analyses of air pollution-causing particles, considering Indonesia's dry and rainy seasons, projected an increase in SO₂ and NO₂ particles over the next two years, accompanied by a decline in the other three particles. When predicting particles, it's essential to consider seasonal fluctuations and long-term trends, which may affect model accuracy and result reliability. However, it's crucial to acknowledge that these model assessments might not capture the full complexity of real-world air pollution patterns, considering factors like seasonal variations, unpredictable weather, regulatory changes, and human behavior shifts, such as transportation habits. Table 3 summarizes the test results and predictions for the five particles.

Table 3. Estimated Average/Day Particles in Dry and Rainy Seasons

No	Particle	Season	Status	Average/Day (µg/m ³)
1	PM ₁₀	Rain		0,0057
		Dry		0,0025
2	CO	Rain		0,0495

		Dry		0,0631
3	SO ₂	Rain		0,0414
		Dry		0,1248
4	NO ₂	Rain		0,0196
		Dry		0,0836
5	O ₃	Rain		0,0792
		Dry		0,1891

To ensure consistency in prediction lengths across the SVR, SARIMA, and Prophet models, linear interpolation is employed before merging them with the LSTM model. Subsequently, all models have the same prediction object length and a hybrid approach with equal weights (50%) is applied. After creating hybrid models, their predictions are compared against actual values to assess accuracy. Hybrid models are favored for their ability to mitigate individual model flaws and enhance forecast precision by combining multiple models to alleviate bias and variation. Leveraging the strengths of each model, such as LSTM's temporal pattern handling and Prophet's seasonality simulation, contributes to improved predictions. A comparison of the expected values for the SVR-SARIMA and LSTM-SARIMA hybrid models is shown in Figure 8. The results suggest that the LSTM-SARIMA model demonstrates higher accuracy and better alignment between predictions and actual values compared to the SVR-SARIMA model.

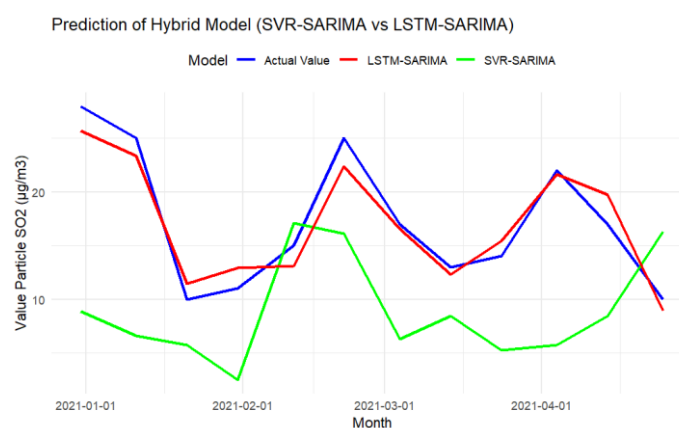


Figure 8. Comparison of Hybrid Model (SVR-SARIMA and LSTM-SARIMA)

The following Figure 9 presents a comparison of the prediction values of the hybrid SVR-Prophet and LSTM-Prophet models. The results indicate that the LSTM-Prophet model outperforms the SVR-Prophet model in terms of accuracy and alignment between predicted and actual values.

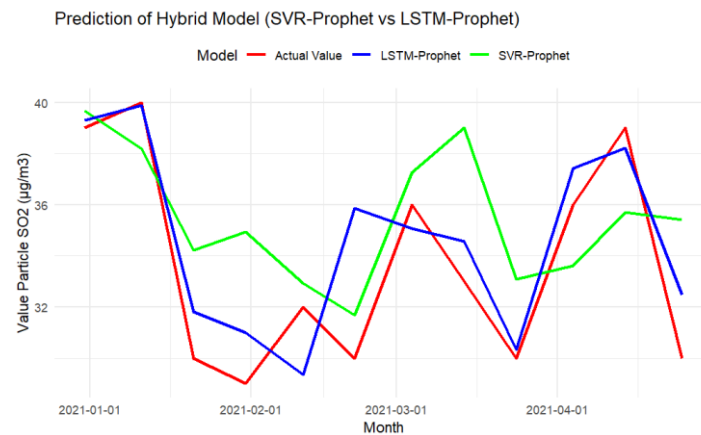


Figure 9. Comparison of Hybrid Model (SVR-Prophet and LSTM-Prophet)

The LSTM-Prophet hybrid model outperforms the other three hybrid models in terms of prediction, as shown by lower values for RMSE, MAE, and MAPE in Table 4.

Table 4. Comparison of Hybrid Model Evaluation Metrics

Matrix	SVR-Prophet	SVR-SARIMA	LSTM-Prophet	LSTM-SARIMA
MSE	766.7909	316.7712	151.7476	236.1367
RMSE	27.69099	25.62913	12.31859	15.36674
MAE	25.70055	20.55425	12.05666	11.21276
MAPE	0.52473	2.46700	0.25879	0.56959

3.5. iNterpret.

The following explanation can be given considering the outcomes of several tests that have been conducted on the model:

- Estimations for the next two years suggest a growth in SO_2 and NO_2 particle concentrations, particularly evident in the dry and wet seasons. During both seasons, the average daily rise in SO_2 particles is $0.0831 \mu\text{g}/\text{m}^3$, and for NO_2 particles, it is $0.0516 \mu\text{g}/\text{m}^3$. The increase in both particles is more pronounced during the dry season, likely attributed to heightened fuel combustion, industrial activities, and energy consumption. However, the precise influence of these factors on elevated NO_2 and SO_2 levels during the dry season warrants further investigation, contingent upon comprehensive dataset support.
- Based on independent testing of the Prophet, SARIMA, SVR, and LSTM models, the results indicate that SVR performs better, with an RMSE value of 7.765, MAE of 5.478, and MAPE of 0.261.
- The LSTM-Prophet hybrid model demonstrates excellent accuracy, achieving a prediction performance of 99.74%. With an RMSE value of 12.319, an MAE value of 12.057, and a MAPE value of 0.259, it outperforms the other three hybrid models.
- Hybridization with non-linear models like SVR and LSTM can enhance the performance of the SARIMA and Prophet models. The results showed that the Prophet and SARIMA models alone were not as effective as the SVR-SARIMA, SVR-Prophet, LSTM-SARIMA, and LSTM-Prophet models.
- SVR and LSTM excel in short-term predictions and pattern detection across various time intervals, while Prophet and SARIMA are adept at analyzing long-term data, especially Prophet's automatic detection of seasonal patterns. Combining LSTM with Prophet effectively addresses seasonal variations by leveraging LSTM's capacity for capturing nonlinear relationships, resulting in improved predictions for datasets with complex temporal and seasonal patterns.

4. CONCLUSION

From the study findings, we can conclude the following:

1. The root means square error (RMSE) of 7.765 and the mean absolute error (MAE) of 5.478 indicates the higher individual performance of the SVR model.
2. The LSTM-Prophet hybrid model demonstrated exceptional accuracy, obtaining an impressive accuracy rate of 99.74%, surpassing comparable hybrid models.
3. The SARIMA and Prophet models show enhanced performance when integrated with non-linear models like SVR and LSTM.
4. The Prophet and SARIMA models demonstrated exceptional proficiency in analyzing long-term data and identifying seasonal patterns, while the SVR and LSTM models had outstanding performance in predicting short-term data.
5. We anticipate a projected rise in SO₂ and NO₂ levels over the next two years, spanning both the dry and wet seasons.
6. The LSTM-Prophet hybrid model is an efficient solution for addressing seasonal fluctuations, which can present a challenge for any model. This hybrid model can enhance prediction accuracy for data with complex variations.
7. Constraints in model selection and assumption-making, uncertainties in long-term predictions, and the impact of local context on result interpretation limit the study.
8. In the future, researchers who want to make models that can predict air pollution will need to include meteorological and environmental data, do a lot of testing, keep the models up to date with new data and types of pollution, deal with data that doesn't make sense, and test other hybridized models again, with a focus on improving the configuration and performance of LSTM models through hyperparameter tuning.

5. ACKNOWLEDGMENTS

We express our gratitude to Universitas Mikroskil for granting the flagship research grant No. 106/UM.348/LP/08/PN/2022, under the terms of the research contract. We also express our gratitude to Universitas Mikroskil's Institute for Research and Community Service (LPPM) for its support during the research process.

REFERENCES

- [1] N. Osseiran and C. Lindmeier, "9 out of 10 people worldwide breathe polluted air, but more countries are taking action," *WHO*, 2018. [Online]. Available: <https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action>. [Accessed: 26-Dec-2022].
- [2] WHO, *WHO global air quality guidelines*. 2021.
- [3] A. Vidiyanto, A. Sindunata, and N. Yudistira, "Air Pollution Particulate Matter (PM2.5) Forecasting using Long Short Term Memory Model," *ACM Int. Conf. Proceeding Ser.*, pp. 139–145, 2021, doi:10.1145/3479645.3479662.
- [4] F. Hamami and I. A. Dahlan, "Univariate Time Series Data Forecasting of Air Pollution using LSTM Neural Network," *2020 Int. Conf. Adv. Data Sci. E-Learning Inf. Syst. ICADEIS 2020*, pp. 12–16, 2020, doi: 10.1109/ICADEIS49811.2020.9277393.
- [5] J. Arumugam, S. Sabarichvarane, and V. Venkatesan, Prasanna, "A Comparative Study of Bitcoin Price Prediction Using SVR and LSTM," *IJCRT*, vol. 10, no. 9, pp. 742–749, 2022, doi: 10.3390/math7100898.

- [6] M. Castelli, F. M. Clemente, A. Popović, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” *Complexity*, vol. 2020, no. M1, 2020, doi: 10.1155/2020/8049504.
 - [7] K. K. R. Samal, K. S. Babu, S. K. Das, and A. Acharaya, “Time series based air pollution forecasting using SARIMA and prophet model,” *ACM Int. Conf. Proceeding Ser.*, pp. 80–85, 2019, doi: 10.1145/3355402.3355417.
 - [8] E. Dave, A. Leonardo, M. Jeanice, and N. Hanafiah, “Forecasting Indonesia Exports using a Hybrid Model ARIMA-LSTM,” *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 480–487, 2021, doi: 10.1016/j.procs.2021.01.031.
 - [9] L. Guo, W. Fang, Q. Zhao, and X. Wang, “The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality,” *Comput. Ind. Eng.*, vol. 161, no. June, p. 107598, 2021, doi: 10.1016/j.cie.2021.107598.
 - [10] S. Xu, H. Kai, and T. Zhang, “Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach,” *Transp. Res. Part E*, vol. 122, no. December 2018, pp. 169–180, 2019, doi: 10.1016/j.tre.2018.12.005.
 - [11] S. Bhanja and A. Das, “A hybrid deep learning model for air quality time series prediction,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 3, pp. 1611–1618, 2021, doi: 10.11591/ijeecs.v22.i3.pp1611-1618.
 - [12] S. Du, T. Li, Y. Yang, and S. J. Horng, “Deep Air Quality Forecasting Using Hybrid Deep Learning Framework,” *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2412–2424, 2021, doi:10.1109/TKDE.2019.2954510.
 - [13] S. J. Taylor and B. Letham, “Forecasting at Scale,” *PeerJ Prepr. 5e3190v2*, vol. 35, no. 8, pp. 48–90, 2017.
 - [14] U. A. Bhatti, Y. Yan, M. Zhou, S. Ali, A. Hussain, and ..., “Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM2.5): An SARIMA and Factor Analysis Approach,” *Ieee ...*, 2021, doi: 10.1109/ACCESS.2021.3060744.
 - [15] S. Fan, D. Hao, Y. Feng, K. Xia, and W. Yang, “A hybrid model for air quality prediction based on data decomposition,” *Inf.*, vol. 12, no. 5, 2021, doi: 10.3390/info12050210.
 - [16] S. Prajapati *et al.*, “Comparison of Traditional and Hybrid Time Series Models for Forecasting COVID-19 Cases,” 2021, doi: 10.21203/rs.3.rs-493195/v1.
 - [17] A. Hasnain, Y. Sheng, M. Z. Hashmi, U. A. Bhatti, and ..., “Time series analysis and forecasting of air pollutants based on prophet forecasting model in Jiangsu province, China,” *Frontiers in ...* frontiersin.org, 2022, doi: 10.3389/fenvs.2022.945628.
 - [18] S. Mahajan, L. J. Chen, and T. C. Tsai, “Short-term PM2.5 forecasting using exponential smoothing method: A comparative analysis,” *Sensors (Switzerland)*, vol. 18, no. 10, pp. 1–15, 2018, doi: 10.3390/s18103223.
 - [19] B. C. Liu, A. Binaykia, P. C. Chang, M. K. Tiwari, and C. C. Tsao, “Urban air quality forecasting based on multidimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang,” *PLoS One*, vol. 12, no. 7, pp. 1–17, 2017,
-

doi: 10.1371/journal.pone.0179763.

- [20] C. J. Huang and P. H. Kuo, “A deep cnn-lstm model for particulate matter (Pm2.5) forecasting in smart cities,” *Sensors (Switzerland)*, vol. 18, no. 7, 2018, doi: 10.3390/s18072220.
- [21] K. Kumari, M. Bhardwaj, and S. Sharma, “OSEMN Approach for Real Time Data Analysis,” *Int. J. Eng. Manag. Res.*, vol. 10, no. 02, pp. 107–110, 2020, doi: 10.31033/ijemr.10.2.11.