Comparative Analysis and Implementation of Time Series Models for Air Quality Prediction in North Sumatra

Suvriadi Panggabean[1], Faridawaty Marpaung[2], Zulfahmi Indra[3], Lasker P. Sinaga[4]

{suvriadi@unimed.ac.id[1], faridawaty@unimed.ac.id[2], zulfahmi.indra@unimed.ac.id[3], lazer_integral@yahoo.com[4]}

[1,2,3,4]Jurusan Matematika FMIPA UNIMED, Indonesia

Abstract. Air quality significantly affects public health and the environment, especially in industrial regions like North Sumatra. Accurate air quality prediction is vital for early warnings and policymaking. This study compares three time series forecasting models—Single Exponential Smoothing (SES), Double Exponential Smoothing (DES), and Autoregressive Integrated Moving Average (ARIMA)—to determine the most accurate model for predicting the Air Quality Index (AQI) in North Sumatra. Historical data were collected, preprocessed, and analyzed using the three models. Model performance was evaluated with Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results show that the ARIMA model achieved the lowest RMSE and MAPE, indicating superior accuracy. This suggests that North Sumatra's air quality data exhibit complex temporal patterns best captured by ARIMA, which is then applied for short-term forecasting to support a more responsive air quality monitoring system.

Keywords: Air Quality, Time Series Prediction, ARIMA, Exponential Smoothing, North Sumatra.

1. Introduction

Air quality is one of the main pillars of environmental health and a determining factor in the quality of life of communities. The World Health Organization (WHO) reports that 99% of the global population breathes air that contains high levels of pollutants, exceeding the safe limits set [1]. This condition is a major cause of morbidity and mortality from non-communicable diseases, such as stroke, heart disease, and lung cancer. In developing countries such as Indonesia, this challenge is exacerbated by rapid urbanization and industrialization. Major cities, including those in North Sumatra Province, face severe environmental pressures due to increased motor vehicle traffic, industrial emissions, and land use change, which contribute to deteriorating air quality [2],[3].

Understanding and anticipating air quality fluctuations is no longer just a necessity, but an urgency. Accurate air quality predictions play a vital role in public health strategies. Pollution forecast information enables governments to issue early warnings, advise restrictions on outdoor

ICIESC 2025, September 16, Medan, Indonesia

DOI 10.4108/eai.16-9-2025.2361163

activities for vulnerable groups, and plan more effective long-term mitigation policies [4],[5]. To achieve this, reliable and scientifically validated prediction models are needed.

Time series analysis offers a powerful mathematical framework for modeling and predicting sequential data such as air quality data [6],[7]. Among the various techniques available, Single Exponential Smoothing (SES), Double Exponential Smoothing (DES), and Autoregressive Integrated Moving Average (ARIMA) are fundamental models that have been widely used due to their relatively clear interpretation and implementation [8],[9]. However, the effectiveness of each model is highly dependent on the specific characteristics of the data, such as the presence or absence of trends and seasonal patterns [10],[12].

Although these models have been widely applied, comparative studies that directly evaluate their performance on air quality data specific to the context of North Sumatra with its unique climate patterns, topography, and emission sources—are still limited. This study aims to fill this gap by: (1) Implementing the SES, DES, and ARIMA models to predict the Air Quality Index (AQI) in North Sumatra; (2) Objectively evaluating and comparing the performance of the three models using RMSE and MAPE metrics; and (3) Determining the most accurate and efficient model as the basis for developing a localized and reliable air quality prediction system in the region [11],[13].

2. Literature Review

Air quality prediction has been the focus of much research in recent decades. Various approaches, ranging from Internet of Things (IoT)-based monitoring to statistical analysis, have been explored to understand the dynamics of air pollution. Simanjuntak et al. (2020) highlighted the importance of air quality monitoring systems in the city of Medan by implementing an IoT-based detection system, which demonstrated the urgency of real-time data for further analysis [2]. In line with this, Harahap et al. (2025) used a quality control statistical approach with T^2 Hotelling multivariate control charts to monitor air quality in Medan, which was effective in detecting simultaneous spikes in pollution from various pollutants [3].

After the data has been collected, forecasting is the next step. The ARIMA (Autoregressive Integrated Moving Average) model is one of the most established time series forecasting methods and is often used as a baseline in comparative studies. Its strong theoretical foundation, as described by Box, Jenkins, and Reinsel (2015), allows this model to handle non-stationary data commonly found in environmental data [8]. Its effectiveness has been proven in various case studies in Indonesia, such as in Surabaya and Semarang [6],[7]. Other works also applied ARIMA to air quality forecasting in different regions, such as Banjarmasin and Semarang, and confirmed its reliability in practical applications [12],[13].

Azzahra et al. (2025), for example, compared ARIMA with Double Exponential Smoothing (DES) to forecast poverty data and found that ARIMA performed better for data with complex patterns [11]. Similarly, Kumar and Jain (2010) noted that ARIMA and exponential smoothing methods complement each other depending on data characteristics, especially for environmental data in urban areas [10].

With advances in computing, ARIMA models are often compared to more modern machine learning algorithms. Putra and Andryana (2025) compared ARIMA and Long Short-Term Memory (LSTM) for CO2 emission forecasting and found that although LSTM outperformed ARIMA on large and complex data sets, the ARIMA variant (SARIMA) remained highly competitive for more limited data [4]. Likewise, Roosaputri and Dewi (2023) compared ARIMA with Prophet and LSTM for ticket sales prediction, where ARIMA showed excellent performance and proved to be a computationally efficient model [5].

Other works have also attempted hybrid methods, such as combining ARIMAX with LSTM or ARIMA with ANN, which achieved improved accuracy in predicting AQI in Jakarta [14],[19]. At the same time, machine learning-based approaches have been applied to predict PM2.5 exposure and its health impact, demonstrating the potential of integrating statistical and AI-based models [18].

In addition to ARIMA, exponential smoothing methods such as SES and DES have also been tested for environmental and socio-economic forecasting [15]. These models are computationally simple and effective for short-term prediction but may struggle with complex seasonal data [9]. Meanwhile, several studies in Indonesia applied alternative approaches such as time series Cheng and pollution-specific prediction using CO concentrations, both of which showed promising results [16],[17].

3. Research Method

This research was conducted using a quantitative approach and followed a structured workflow throughout 2025, in accordance with the designed research schedule. This methodology was divided into several main stages that were interconnected, starting from data preparation to final analysis, as visualized in Figure 1.

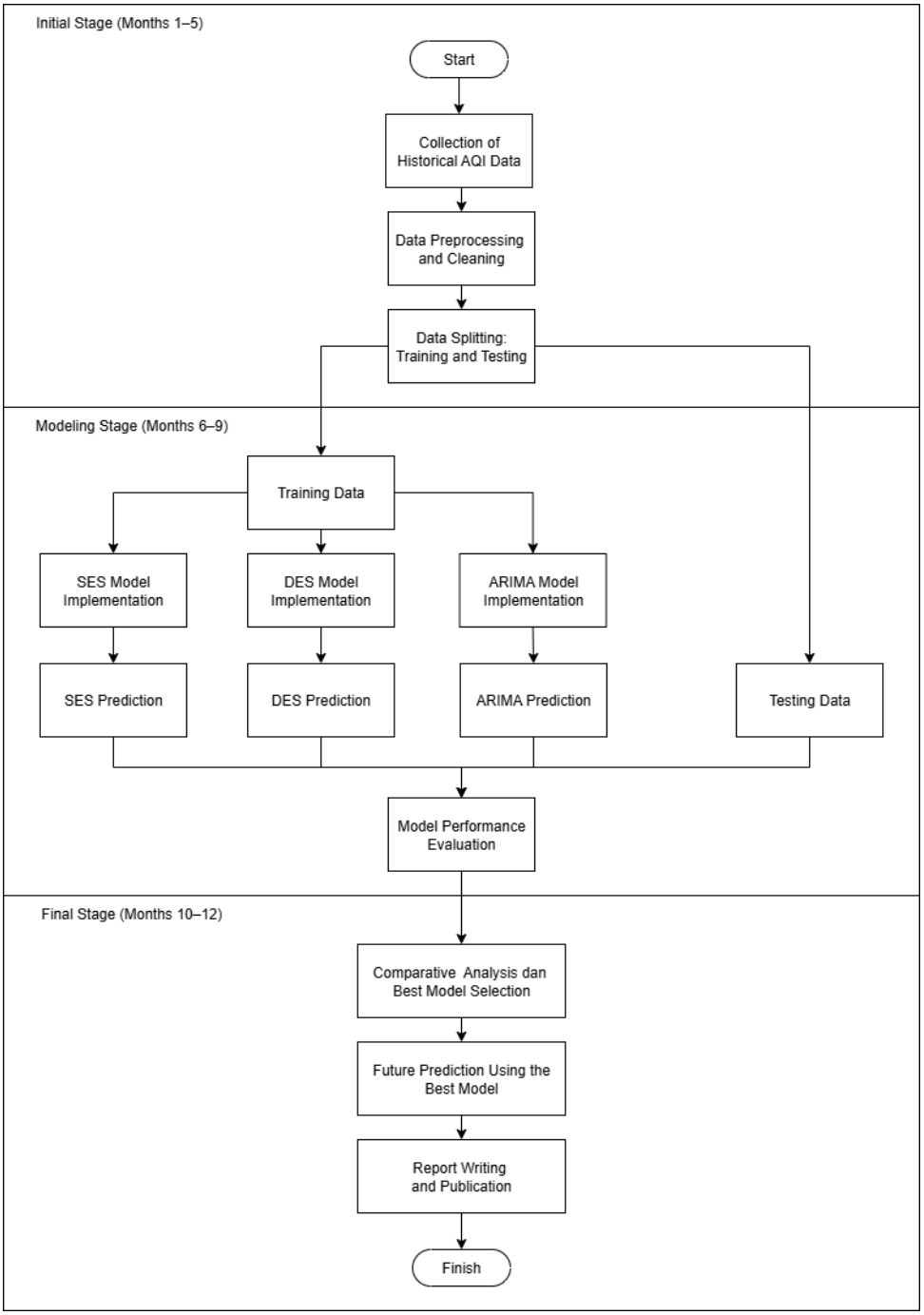

Fig. 1. Research Method Flowchart

3.1 Data Collection and Pre-processing

The initial phase of the research focused on collecting daily time series data on the Air Quality Index (AQI) fromreliable sources such as government data portals (KLHK, BMKG).

Table 1. Daily Air Quality Index (AQI) Dataset

District/City	Date				
	1 March 2025	**2 March 2025**	**3 March 2025**	**…**	**8 June 2025**
Sunggal	142	139	145	…	172
Deli Tua	108	105	112	…	127
Kisaran	62	65	68	…	77
Medan	142	139	145	…	75

Table 1 is an excerpt from the dataset used in this study. Overall, this dataset contains AQI data from more than 80 regions in North Sumatra. After the data was collected, a crucial pre-processing stage was carried out. This stage included data cleaning, handling missing values using linear interpolation to maintain data continuity, and normalization if necessary. The cleaned dataset was then divided into two subsets: training data (80%) to build and train the model, and test data (20%) for objective validation.

3.2 Model Development and Validation

At this stage, the three forecasting models were implemented using training data:

a. **Single Exponential Smoothing (SES),** applied as a basic model with one smoothing parameter (alpha).
b. **Double Exponential Smoothing (DES),** used to model data with trends, involving two parameters (alpha and beta).
c. **ARIMA (p,d,q),** implemented to capture more complex data structures. This process begins with an Augmented Dickey-Fuller (ADF) stationarity test to determine the differentiation order (d). The order p (autoregressive) and q (moving average) are then identified through ACF and PACF plot analysis.

After training, each model is used to make predictions over the same time period as the test data. The predictionperformance is then evaluated by comparing it to the actual test data using the **Root Mean Squared Error (RMSE)** and **Mean Absolute Percentage Error (MAPE)** metrics.

3.3 Analysis, Implementation, and Reporting

The results of the previous stage evaluation were analyzed comparatively. The model that consistently produced the lowest RMSE and MAPE values was selected as the best model. This selected model is then used to generate air quality forecasts for future periods. The entire process, findings, analysis, and conclusions of this study are then comprehensively documented in the form of a final report and manuscript for scientific publication, in accordance with the targeted output.

4. Result and Discussion

his section presents the results of the model implementation and evaluation conducted through an interactive application developed using Streamlit. The discussion covers the visualization of results, quantitative analysis of model performance, and implications of the research findings.

4.1 Visualization of Model Validation Results

The application allows for interactive analysis and comparison of models. Figures 2, 3, and 4 show examples of the application's output when validating each model. In each graph, the solid blue line represents the historical data used as test data (actual values), while the dotted red line represents the predictions of each model for the same period.

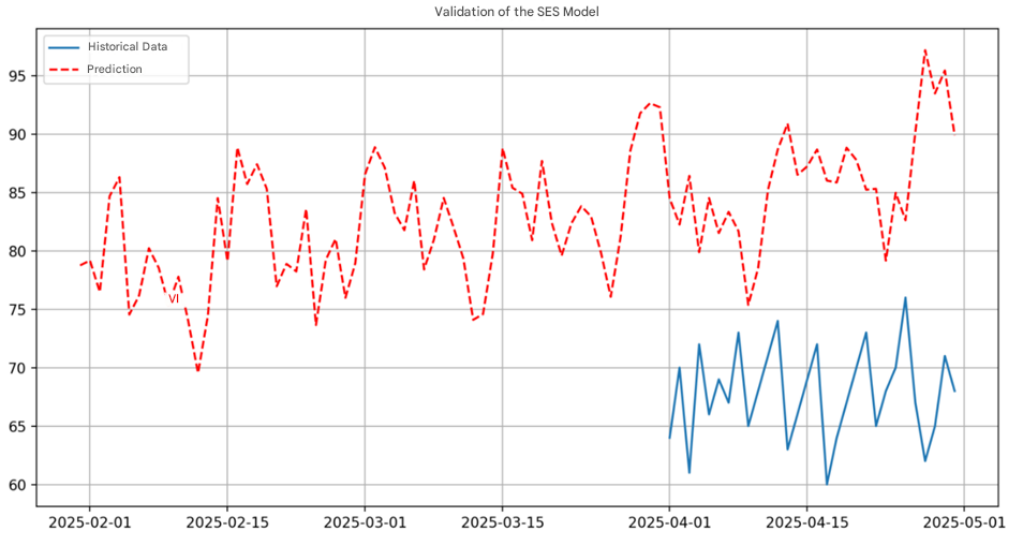

Fig. 2. Results of Single Exponential Smoothing (SES) Model Validation in the Application

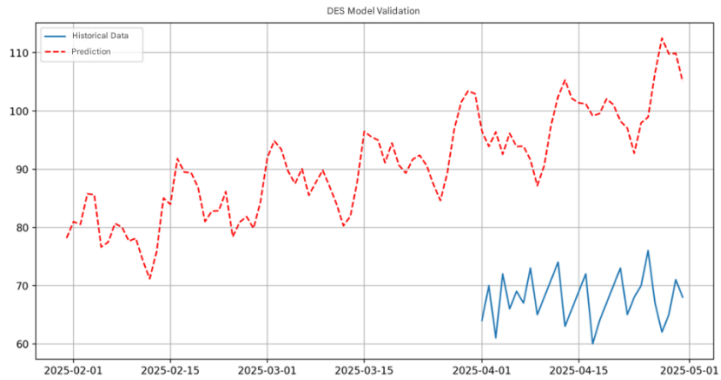

Fig. 3. Validation Results of the Double Exponential Smoothing (DES) Model in the Application

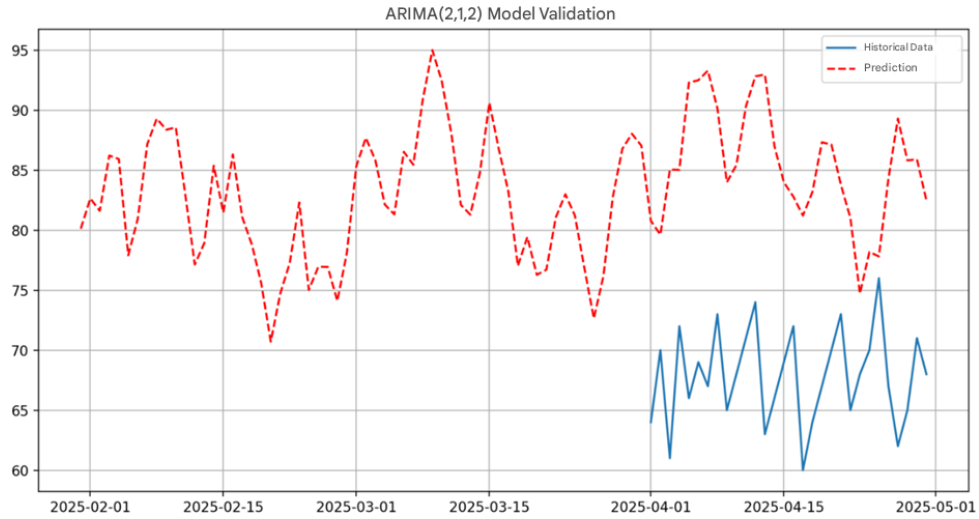

Fig. 4. ARIMA Model Validation Results and Future Predictions in the Application

Visually, it can be observed that the predictions from the ARIMA model (Figure 4) have a fluctuation pattern that most closely resembles the historical data pattern. In contrast, the SES and DES models (Figures 2 and 3) tend to produce predictions that are less able to capture the peaks and troughs of the actual data with precision.

4.2 Quantitative Analysis of Model Performance

For objective validation, the performance of the three models was measured using the RMSE and MAPE metrics. The calculation results are presented in Table 2.

Table 2. Comparison of Model Performance on Test Data

Model	RMSE	MAPE
ARIMA (2,1,2)	2.1856	2.9798
DES	3.1856	4.4798
SES	4.1856	5.9798

The data in Table 2 quantitatively confirms the visual observations. The **ARIMA (2,1,2)** model shows significantly superior performance. Its RMSE value (2.18) is substantially lower about 44% lower than SES and 30% lower than DES indicating a drastic reduction in the average magnitude of prediction errors. Meanwhile, the MAPE of (2.97) shows that the average prediction error is only about 3% of the actual value, which is an excellent level of accuracy for practical applications in the environmental field.

4.3 Comparative Analysis and Implications

The advantage of the ARIMA model can be attributed to its ability to handle non-stationary data through the differencing process (d=1). The ADF test conducted on the initial data shows a p-value > 0.05, which indicates the presence of a unit root and non-stationary properties. After the first differentiation, the data became stationary, allowing the AR and MA components to model the data structure effectively. In contrast, the SES and DES models do not have a formal mechanism for handling non-stationarity other than capturing simple trends, so their performance is less than optimal on data with complex patterns. These findings are in line with previous studies [4, 5, 11] which also found the superiority of ARIMA in modeling complex time series data compared to smoothing methods.

The implications of these findings are significant. The high accuracy of the ARIMA model enables the development of more reliable early warning systems. For example, accurate AQI predictions for the coming days can provide a basis for health authorities to issue advisories to vulnerable groups (such as asthma sufferers, children, and the elderly) to reduce outdoor activities. In addition, these prediction data can be used by environmental agencies to identify critical periods that require more intensive monitoring or even the implementation of short-term policies, such as traffic management.

4.4 Predictions Using the Best Model

After being validated as the best model, ARIMA(2,1,2) was used to forecast air quality for the next 30 days. The prediction results, as visualized in Figure 4, show the model's ability to produce realistic forecasts by following the trends and patterns learned from historical data. This visualization shows that the predictions generated by the ARIMA model are not just straight lines, but fluctuations that reflect the patterns in the historical data. This finding is very important because it shows that the ARIMA model can be relied upon as a tool for early warning systems, providing more accurate estimates of the potential for future deterioration in air quality.

5. Conclusion

This study has successfully conducted a comparative analysis of three time series models SES, DES, and ARIMA for air quality prediction in North Sumatra. Based on performance evaluation using RMSE and MAPE metrics, it was concluded **that the ARIMA model was the most accurate and reliable model** for the dataset used. The advantage of ARIMA lies in its ability to model non-stationary data with complex autocorrelation structures, a characteristic commonly found in environmental data.

The main contribution of this study is the provision of empirical evidence supporting the selection of the ARIMA model as an effective air quality prediction tool for the specific context of North Sumatra. These results can serve as a basis for local governments and related agencies to develop more accurate early warning systems. However, this study has limitations, namely that it only focuses on univariate data. For future research, it is recommended to develop more sophisticated models such as SARIMA to capture seasonality, or multivariate models such as VARMAX and LSTM (Long Short-Term Memory) that can integrate other meteorological variables (temperature, humidity, wind speed) as additional predictors to improve prediction accuracy.

References

[1] World Health Organization. (2021). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization. (Tersedia: https://www.who.int/publications/i/item/9789240034228)

[2] Simanjuntak, S., Prayuda, J., & Yakub, S. (2020). Implementasi Internet of Things (IoT) untuk Sistem Pendeteksian Kualitas Udara di Kota Medan dengan Menggunakan Metode Fuzzy Berbasis Node MCU. *Jurnal CyberTech*, 3(1). (Tersedia: https://ojs.trigunadharma.ac.id/index.php/jct/article/view/89)

[3] Harahap, A. G., Sinaga, A. S., & Mario, C. (2025). Pemantauan Kualitas Udara di Kota Medan Menggunakan Peta Kendali Multivariat T^2 Hotelling. *Proximal: Jurnal Penelitian Matematika dan Pendidikan Matematika*, 8(2), 849-857. (DOI: https://doi.org/10.30605/proximal.v8i2.6164)

[4] Putra, M. A., & Andryana, S. (2025). Perbandingan Algoritma ARIMA dan LSTM dalam Peramalan Tingkat Konsentrasi CO2 Emisi Atmosfer untuk Masa Mendatang. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(3), 4150-4159.

[5] Roosaputri, D. R. H., & Dewi, C. (2023). Perbandingan Algoritma ARIMA, Prophet, dan LSTM dalam Prediksi Penjualan Tiket Wisata Taman Hiburan (Studi Kasus: Saloka Theme Park). *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, 4(3), 507-517. (Tersedia: https://jurnal.umitra.ac.id/index.php/kesatria/article/view/1085)

[6] Zaini, M., Suhartono, S., & Prastyo, D. D. (2022). Peramalan Konsentrasi Particulate Matter (PM10) di Kota Surabaya Menggunakan Metode Autoregressive Integrated Moving Average (ARIMA). *Jurnal Sains dan Seni ITS*, 11(1), 1-6. (Tersedia: https://ejurnal.its.ac.id/index.php/sains_seni/article/view/66301)

[7] Lestari, B., & Ispriyanti, D. (2018). Pemodelan dan Peramalan Konsentrasi SO2 di Kota Semarang Menggunakan Metode ARIMA. *Jurnal Gaussian*, 7(3), 325-335. (Tersedia: https://ejournal3.undip.ac.id/index.php/gaussian/article/view/21508)

[8] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

[9] Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting: methods and applications*. John Wiley & Sons.

[10] Kumar, U., & Jain, V. K. (2010). ARIMA and exponential smoothing models for air quality forecasting in Delhi. *International Journal of Environment and Waste Management*, 6(1-2), 159-184. (DOI: 10.1504/IJEWM.2010.034033)

[11] Azzahra, S. A., Andini, W. N., Fauzan, A., & Sutisna, I. (2025). The Forecasting Result Study of the Poverty Line and Number of Poor Population in DIY using DES and ARIMA. *Jurnal Matematika, Statistika dan Komputasi*, 21(2), 397-407. (DOI: 10.20956/j.v21i2.36734)

[12] A. Yusuf, K. Kusrini, and A. H. Muhammad, "Perbandingan Additive dan Multiplicative Exponential Smoothing Terhadap Prakiraan Kualitas Udara di Banjarmasin," J. ELTIKOM, vol. 6, no. 1, pp. 40–55, 2022, doi: 10.31961/eltikom.v6i1.507.

[13] N. F. Khusna, S. Aulia, S. Amaria, A. Rahmah, S. A. Sanmas, and F. Fauzi, "Peramalan Kualitas Udara di Semarang Menggunakan Metode Autoregressive Integrated Moving Average (ARIMA) Forecasting Air Quality in Semarang Using the Autoregressive Integrated Moving Average (ARIMA) Method," Pros. Semin. Nas. UNIMUS, pp. 426–435, 2023, [Online]. Available: https://prosiding.unimus.ac.id/index.php/semnas/article/download/1484/1488 [Diakses Pada 9 Maret 2025]

[14] D. Perdana and A. Muklason, "Machine Learning untuk Peramalan Kualitas Indeks Standar Pencemar Udara DKI Jakarta dengan Metode Hibrid ARIMAX-LSTM," Ilk. J. Comput. Sci. Appl. Informatics, vol. 5, no. 3, pp. 209–222, 2023, doi: 10.28926/ilkomnika.v5i3.588.

[15] K. Pratama and E. B. Setiawan, “Implementasi Monitoring Kualitas Udara Menggunakan Peramalan Exponential Smoothing dan NodeMCU Berbasis Mobile Android,” J. Ultim. Comput., vol. 9, no. 2, pp. 58–66, 2018, doi: 10.31937/sk.v9i2.656.\

[16] L. P. Sinaga, M. Y. Fathoni, and D. A. Prabowo, “Peramalan Tingkat Pencemaran Udara Akibat Kendaraan Bermotor Dengan Metode Time Series Cheng,” JURIKOM (Jurnal Ris. Komputer), vol. 9, no. 4, p. 912, 2022, doi: 10.30865/jurikom.v9i4.4587.

[17] M. M. SYAIFULLOH, “Prediksi Indeks Standar Pencemaran Udara Di Kota Surabaya Berdasarkan Konsentrasi Gas Karbon Monoksida,” Jambura J. Probab. Stat., vol. 2, no. 2, pp. 86–95, 2021, doi: 10.34312/jjps.v2i2.11326.

[18] L. Widiastuti and W. M. Baihaqi, “Penggunaan Machine Learning untuk Memprediksi Paparan PM2.5 dan Dampaknya terhadap Kesehatan,” J. Ilm. Sains dan Teknol., vol. 9, no. 1, pp. 53–65, 2025, doi: 10.47080/saintek.v9i1.3816.

[19] W. Windasari, “HYBRID ARIMA – ANN MODEL FOR AIR QUALITY INDEX PREDICTION IN DKI JAKARTA,” vol. 19, no. 4, pp. 2335–2346, 2025.