BAREKENG: Journal of Mathematics and Its Applications
December 2025 Volume 19 Issue 4 Page 2335–2346
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19no4pp2335-2346

HYBRID ARIMA–ANN MODEL FOR AIR QUALITY INDEX PREDICTION IN DKI JAKARTA

Wahyuni Windasari[1*], **Augistri Putri Pradani**[2]

[1,2] *Departement of Data Science, Faculty of Science and Technology, Universitas Putra Bangsa*
Jln. Ronggowarsito 18, Kebumen, 54361, Indonesia

*Corresponding author's e-mail: * **wahyuni@fst.universitasputrabangsa.ac.id***

Article History:
Received: 21st November 2024
Revised: 18th February 2025
Accepted: 1st June 2025
Available online: 1st September 2025

Keywords:
Artificial Neural Network;
Air Quality Index;
Hybrid ARIMA-ANN;
$PM_{2.5}$.

ABSTRACT

Air pollution is a threat to all countries, including Indonesia. One area in Indonesia with poor air quality is DKI Jakarta. One step to minimize the decline in air quality in an area is to predict the air quality index in the future. In this study, a hybrid ARIMA-ANN analysis was conducted, combining the ARIMA method and Artificial Neural Networks to model air quality in DKI Jakarta. The time series data of the air quality index $PM_{2.5}$ sourced from the DKI Jakarta Environmental Service during January 19-30, 2023, which was observed every hour with a total of 288 data. The results of the study showed that the SAE and RMSE of the ARIMA model were 94.135 and 1.157, respectively, while the SAE and RMSE values of the hybrid ARIMA-ANN model were 61.094 and 1.15. The results of the study showed that the hybrid ARIMA-ANN model had a higher accuracy value compared to the single ARIMA model in describing DKI Jakarta air quality data. This study has limitations in that determining the network architecture in the ANN model is still done by trial and error, so it takes a relatively longer time.

How to cite this article:

W. Windasari and A. P. Pradani., "HYBRID ARIMA–ANN MODEL FOR AIR QUALITY INDEX PREDICTION IN DKI JAKARTA," *BAREKENG: J. Math. & App.*, vol. 19, no. 4, pp. 2335-2346, December, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Air pollution is one of the serious environmental problems, especially in urban areas. The presence of air pollution has a negative impact on human health and the environment. Diseases caused by air pollution are estimated to be comparable to other global health risks, such as smoking and unhealthy diets. The World Health Assembly also determined that air pollution is a risk factor for non-communicable diseases such as ischemic heart disease, chronic obstructive pulmonary disease, stroke, cancer, and asthma [1]. Air pollution is caused by particles and gases released into the atmosphere from various human activities, such as industrial facilities [2], household combustion devices, forest fires, and transportation are common sources of air pollution. The results of the study in Isfahan showed that 1374 tons of carbon monoxide, 727 tons of VOCs, 691 tons of hydrocarbons, 329 tons of PM_{10}, 319 tons of $PM_{2,5}$, 13.297 tons of nitrogen oxides, and 0.13 tons of sulfur oxides were released into the atmosphere on average each year by the railway system. In addition, the study of the pollutant emission inventory produced by aircraft showed that in 2016, an average of 8076 tons of carbon dioxide, 22.8 tons of carbon monoxide, 19.4 tons of nitrogen oxides, 3.8 tons of hydrocarbons, and 2.54 tons of sulfur oxides were released into the atmosphere each year [3]. Furthermore, research on pollutants in China shows that non-road transportation produces more $PM_{2,5}$ particulates than road transportation [4].

As a developing country, Indonesia also has problems with air pollution. Based on IQAir, Indonesia is placed 17th as the country with the worst air quality in the world in 2021, with Jakarta as the most polluted city in Indonesia [5]. In fact, for the first time in June 2022, Jakarta was declared the city with the worst air quality in the world, with an air quality index value reaching 185 AQI (Air Quality Index), which is classified in the red or unhealthy category. The AQI index is an index that describes air quality issued by the United States Environmental Protection Agency (EPA), with the range being between 0 and 500. This AQI index is calculated based on six main types of pollutants, namely Carbon monoxide, Sulfuric acid, Nitrogen dioxide, ground surface Ozone, Particulate Matter $PM_{2.5}$ and PM_{10}. From the six pollutants, Particulate Matter ($PM_{2.5}$) are the air particles that have a size less than 2.5 μm (micrometers). These particles can increase due to hot air, dust, and smoke because of fires, environmental pollution, and waste disposal. When inhaled, these particles, which are smaller than 3% of the human hair diameter, can settle on the surface and deep parts of the lungs, causing short-term health impacts related to respiratory diseases such as lung disease, heart disease, bronchitis, and asthma. Meanwhile, if exposed long-term, $PM_{2.5}$ can cause chronic heart and lung disease, decreased lung function in children, and even premature death. Research showed the influence of $PM_{2.5}$ on the total number of deaths due to lung cancer in 15 cities in Iran, amounting to 120 cases during the period 21 March 2015 to 19 March 2016 [6].

The existence of serious adverse effects, especially on health caused by air pollution, makes research related to air pollution something that needs to be done. Especially research related to the diameter of 2.5 μm ($PM_{2.5}$), which is considered the most significant air pollution [7]. Research related to air pollution has begun to attract researchers, especially those related to predicting Particulate Matter (PM) both $PM_{2.5}$ and PM_{10} using statistical models and machine learning approaches. Particulate Matter (PM) prediction using statistical models has been widely developed both through a causal approach using linear regression [8]-[11], a time series approach [12]-[15], and a deep learning approach [16]. The research results show that the ARIMA model is good enough to describe the training dataset PM_{10} in Pekanbaru, while the prediction results for testing data do not adequately describe the actual data [12]. The ARIMA model was deemed capable of being applied to estimate the concentration of $PM_{2.5}$ which experienced higher fluctuations in cold periods and lower in warm periods in Fuzhou, China [14]. Furthermore, the modeling of the PM_{10} dataset for Surabaya was carried out using Double Seasonal ARIMA (DSARIMA) with two observation stations, namely Kebon Bibit Wonorejo (SUF6) and Kebonsari Village (SUF7) [13]. The research shows that the forecasting results of sample data are better than out-of-sample data. This is because the forecasting in-sample data is the one-stage forward forecast so the small RMSE is obtained, while the forecasting on out-of-sample data is a direct n-stage forward forecast. Furthermore, the results of the study using a univariate time series model to make model of $PM_{2.5}$ concentration in three regions, namely Majalengka, Kuningan, and Cirebon shows that the univariate time series model can predict the concentration of $PM_{2.5}$ for the short term, while the prediction results for the long term have a different pattern [15]. Another study used time series models based on autoregressive integrated moving average (ARIMA), autoregression (AR), and moving average (MA) models applied to $PM_{2.5}$ concentration in Delhi and Bengaluru. The results showed that all three time series statistical models provided inaccurate model performance for modeling $PM_{2.5}$ concentration in Delhi, with MAPE values of AR, MA, and ARIMA models above 20% [17]. This research suggests the need for additional

topographic and other meteorological parameters to produce a better model. In addition to statistical and time series approaches, research related to air quality prediction has also begun to be developed using deep learning approaches such as LSTM. The results of research related to air quality prediction in Jakarta show that for univariate analysis, the ARIMA method has lower RMSE, MSE, MAE, and error ratio values than the LSTM method [16]. This is because the actual data has a value that does not fluctuate or is almost constant (has a fixed value) throughout the observation.

Nowadays, research on machine learning models to estimate the concentration of air pollutants, especially $PM_{2.5}$ is starting to be developed. Even some studies show that modeling with a machine learning approach or a hybrid of machine learning and statistics provides better results compared to statistical models. The ANN Backpropagation Machine Learning model is better than the ARIMA method [18]. One of the weaknesses of the ARIMA model is the inability of this classical statistical method to describe non-linear data patterns. On the other hand, the advantage of the ANN Backpropagation method is the model's ability to describe the correlation of non-linear patterns well. Used three machine learning approaches, including a new hybrid model based on long short-term memory (LSTM), a deep feedforward neural network (DFNN), and multiple additive regression trees (MART). The results showed that the LSTM model provided the best results in modeling $PM_{2.5}$ mass concentrations. The same results were also shown in several studies that examined the comparison of machine learning and statistical models [17], where both machine learning and hybrid approaches provided better results than statistical models [19]-[21]. Based on background and previous research, this research will carry out DKI Jakarta air quality modeling and compare the accuracy of classical statistical methods and the hybrid ARIMA-ANN method.

2. RESEARCH METHODS

In this research, data analysis was carried out using the hybrid ARIMA-ANN method. Hybrid ARIMA-ANN is a combination of two or more systems in one function. In this research, we use a combination of the different advantages of Artificial Neural Network (ANN) and the classical statistical method ARIMA. The ARIMA model has the advantage of representing linear patterns in data. Furthermore, the ANN model will estimate the residuals obtained from the ARIMA model by building a neural network [22].

2.1 ARIMA Model

The ARIMA model attempts to identify patterns in historical data. The ARIMA model has three components, each of which helps model a specific type of pattern. The Autoregressive (AR) component takes into account the pattern between a given period and the previous period. The Moving Average (MA) component measures the adaptation of new forecasts to previous forecast errors. The Integration (I) component indicates trends or integrative processes in the data. The ARIMA model is symbolized by ARIMA (p, d, q), where p is the order of autoregressive terms, d is the number of differences, and q is the number of moving averages. In general, the form of the ARIMA (p, d, q) model is given in the following **Equation (1)**:

$$\left(1 - a_1 B - a_2 B^2 - \cdots - a_p B^p\right)(1 - B)^d y_t = \mu + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q} \quad (1)$$

where error ε_t has an IID normal distribution with zero mean and constant variance σ_ε^2, B is the lag operator.

2.2 Neural Network Methods

Neural Network methods, also known as Artificial Neural Network (ANN), are a part of machine learning that in the process, resemble how the human brain's nerves work. The most frequently used ANN architecture is a multilayer network with the Backpropagation method. The advantage of the ANN Backpropagation method is the ability to formulate appropriate experience and knowledge, and the forecasting rules are flexible. On the other hand, the ANN Backpropagation method has disadvantages, including that the predictions obtained from this method can give invalid results if the input received is outside the range given during training, or the required training data is not sufficient.

The algorithm of the Backpropagation method consists of several artificial neural networks that are connected. The arrangement of artificial neural networks generally consists of three layers, namely:

1. Input layer

   The input layer is a layer consisting of input units that receive data patterns from outside. The unit in the input layer is called the neuro input.

2. Hidden layer

   The hidden layer is a layer that contains the units whose output cannot be directly observed. The units in the hidden layer are called hidden neurons.

3. Layer output

   A layer that consists of output units, which are solutions produced from the ANN.

2.3 Hybrid ARIMA-ANN

The hybrid model is a combination method of one or more models in the function of a system. The ARIMA and ANN models are models to solve linear or nonlinear problems. The purpose of the hybrid ARIMA ANN method consists of two things. First, the ARIMA model is used to analyze the linear part of the problem, and second, the ANN model is built to model the residuals of the ARIMA model. This is because the ARIMA model cannot capture the nonlinear structure of the data, the residual of the linear model will have information about nonlinearity. The results of the ANN can be used to predict errors for the ARIMA model.

To get a hybrid of the ARIMA model and Artificial Neural Network, namely by adding the models. The ARIMA model is formed from research data, while the Artificial Neural Network model is formed from the ARIMA model error. The general form, the hybrid ARIMA-ANN model is written as **Equation (2)**.

$$\hat{y}_t = \hat{x}_t + \hat{z}_t \tag{2}$$

$\hat{x}_t$: ARIMA value on time t
$\hat{z}_t$: ANN value on time t

The research stages using the hybrid ARIMA-ANN method are **[23]** :

1. Conduct descriptive statistical analysis and exploration of air quality data using time series plots.
2. Modeling of daily air quality in Jakarta using the ARIMA model **[24]**
   a. Testing data stationarity in variance and mean. In this study, the Box-Cox test was used to test data stationarity in variance. To test data stationarity in the mean, the Dickey-Fuller test was used.
   b. Identify ARIMA models based on the ACF and PACF plots.
   c. Estimating and testing the significance parameters model.
   d. Perform model diagnostic tests to check the independence and normality of the residuals.
   e. Overfitting the data to check the possibility of another model that is simpler than the previous model.
   f. Selecting the best ARIMA model based on SAE and RMSE criteria.
3. Error modeling of ARIMA using Artificial Neural Network (ANN).
   a. Determine the input variables for hybrid ARIMA-ANN modeling based on the ACF and PACF plots of the residuals in the selected ARIMA model.
   b. Building a Backpropagation ANN architecture.

      In the ANN method, there are no standard rules for determining the optimal ANN to be applied to the system. Henceforth, determining the architecture, especially determining the number of nodes in the hidden layer, is carried out by trial and error **[25]**.

   c. Initialize parameters, including the learning rate to determine the network learning constant and the epoch value to determine the maximum number of iterations.

d. Choose the best hybrid ARIMA-ANN model by looking at the SAE and RMSE values.

4. Compare the best results from the ARIMA and hybrid ARIMA-ANN methods and then select the best results based on the smallest SAE and RMSE value.

5. Summarize the results of the analysis and modeling.

2.4 Best Model Selection

The aim of selecting the best model is to get the right model to describe air quality data. This selection is based on accuracy values using Mean Square Error (MSE), Sum of Absolute Error (SAE), and RMSE. The MSE, SAE, and RMSE value is formulated as follows **Equation (3)**.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2 \tag{3}$$

$$SAE = \sum_{i=1}^{n}|y_t - \hat{y}_t|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2}$$

where :
y_t : actual data
$\hat{y}_t$: forecasting data
n : amount of data

3. RESULTS AND DISCUSSION

3.1 Data

This research uses daily data on the air quality monitoring in Jakarta using the parameter $PM_{2.5}$. The data that is used in this research is secondary data sourced from the Department of Environment (DOE). The data used is air quality data for the period 19 - 30 January 2023. Data is taken every hour every day, starting at 00.00 to 23.00, with a total of 288 data. The analysis used in the research is the hybrid ARIMA-ANN method, which is solved by the R program **[26]**.

3.2 Descriptive Statistics

Atmospheric Particulate Data $PM_{2.5}$ during 19 – 30 January 2023 is shown in the following **Table 1**.

Table 1. Descriptive Statistics value of Atmospheric Particulate $PM_{2,5}$

Criteria	$PM_{2,5}$ Value
Minimum	35.108
Maximum	87.343
Mean	54.144
Standard deviation	10.641

Based on **Table 1**, it is known that $PM_{2.5}$ has a minimum value of 35.108 and a maximum $PM_{2.5}$ value of 87.343. The average $PM_{2.5}$ during 19 – 30 January 2023 was 54.144, which is included in the moderate category.

3.3 ARIMA Model

Time series data can be modeled using ARIMA provided that the observed time series data is stationary. To see a rough estimate of the shape of the ARIMA model to be built, whether it has passed the

stationary test can be seen from the data plot. Stationary time series data can be seen from the absence of a trend or sharp increase or decrease in the data.

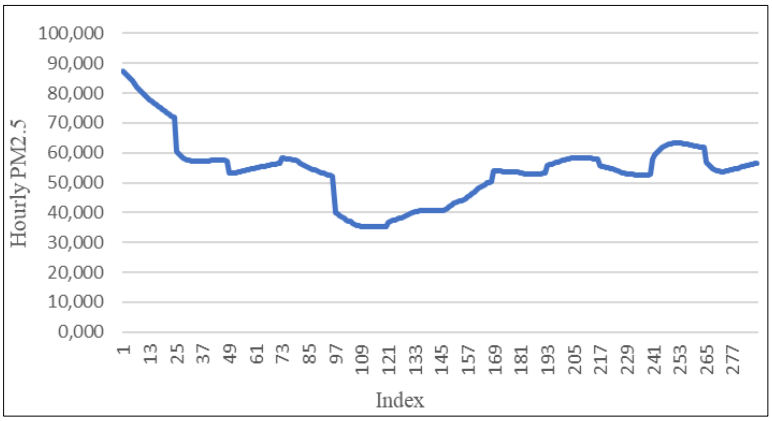
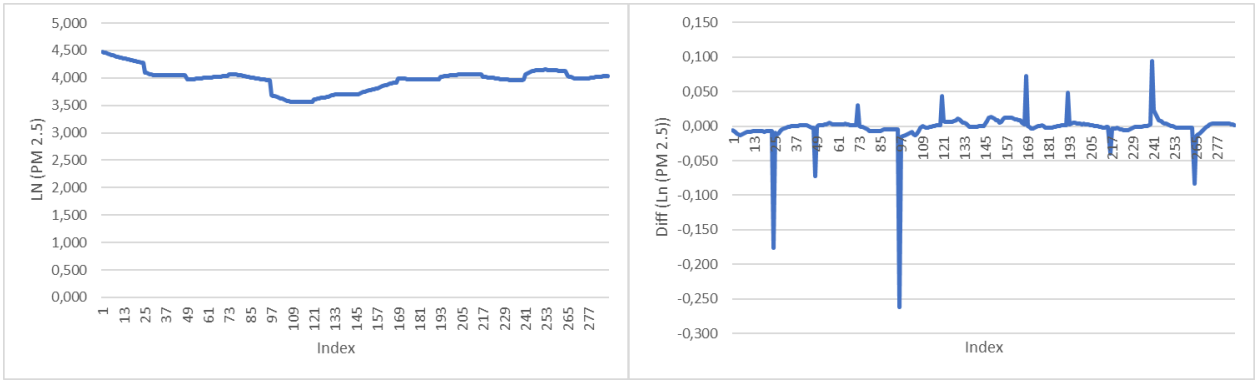


Figure 1. Plot Data $PM_{2.5}$

Based on **Figure 1**, we know that the plot data $PM_{2.5}$ tends to fluctuate and contain trends. This indicates that $PM_{2.5}$ is not stationary in either mean or variance. To get stationary data, before modeling the time series data, it is necessary to carry out a natural logarithm (Ln) transformation on the data $PM_{2.5}$



(a) **(b)**

Figure 2. Plot Data Ln($PM_{2,5}$) Before and After Differencing

The Ln-transformed data still tends to fluctuate. This indicates that there is still a trend pattern in the data seen in **Figure 2** (a). To obtain data that tends to be more stable, it is necessary to differentiate the Ln-transformed data. The plot of the data resulting from the differentiation of the Ln-transformed data in **Figure 2** (b) is already around the average value. This indicates that the first differencing data from the Ln ($PM_{2.5}$) is stationary. This can also be confirmed by the results of the unit root test using the ADF test. The ADF test statistical value, which is smaller than its critical region, indicates that the data is stationary.

Table 2. Result of Root Test ADF

	ADF_{test}	*Test Critical Value 5%*	**Prob.**
$PM_{2.5}$ Level	-3.067	-3.426	0.1162
Ln($PM_{2.5}$) Level	-2.329	-3.426	0.4165
Ln($PM_{2.5}$) *first different*	-14.727	-3.426	0.0000

Based on the test results that have been carried out at the best level, both the $PM_{2.5}$ data and the Ln($PM_{2.5}$) transformation data are not yet stationary. This can be seen from the value of ADF_{test} data $PM_{2.5}$ and transformed data Ln($PM_{2.5}$) which is greater than the value of the test critical value. Furthermore, for the first different level test results, the data results were stationary. This can be seen from the value of ADF_{test} which is smaller than the value of the test critical value, namely -14.727 < -3.426. The same results were also obtained from probability values that were smaller than the alpha value of 0.05. Because the data is stationary in the first difference, the d order in the ARIMA model has a value of one, while the p and q orders can be seen based on the ACF and PACF plots shown in **Figure 3**.

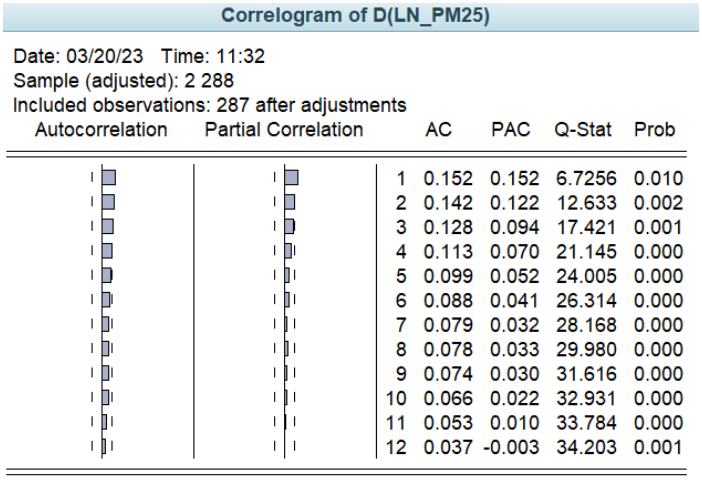


Figure 3. Plot ACF and PACF Data Ln($PM_{2.5}$), d=1

Model identification is carried out using the Bartlett test, where each lag on the ACF and PACF plots will be within the interval boundary line (5%). Lags that exceed the boundary line are identified as AR levels based on the PACF plot and MA levels based on the ACF plot. Based on the ACF and PACF plots in **Figure 3**, a cut-off occurs at the first lag. Considering the principle of model simplicity, the possible models are ARIMA (1,1,1), ARIMA (1,1,0), and ARIMA (0,1,1). Further, a parameter significance test was carried out in each model. The selected model is a model whose parameters pass the significance test.

Table 3. Parameter Significance Test Data Ln($PM_{2,5}$), d=1

ARIMA Model	AR (1)	MA (1)	Decision
ARIMA (1,1,1)	0.8878 (0.0000)	-0.7748 (0.0009)	✓
ARIMA (1,1,0)	0.1518 (0.0164)		✓
ARIMA (0,1,1)		0.1240 (0.1440)	×

Based on **Table 3**, it is obtained that the ARIMA model (1,1,1) and the ARIMA model (1,1,0) have significant parameters. The next step is to carry out diagnostic tests on the two selected results to determine the best model.

Table 4. Diagnostic Test

Test	ARIMA (1,1,1)	ARIMA (1,1,0)
Q-stat Test (residual independence test)	✓	×
Homoscedasticity test	✓	✓

Based on diagnostic tests, it can be seen that the ARIMA model (1,1,1) passes the Q-stat test and also the homoscedasticity test. This shows that there is no serial correlation in the data, and the residual data is random, so the ARIMA (1,1,1) model can be considered as the selected model. After obtaining the appropriate ARIMA model, the fit value for $PM_{2,5}$ air quality data can be obtained. The R^2 for $PM_{2,5}$ data is 98.62%. This value indicates that the ARIMA model is very good at describing the original data. In addition, the MSE value shows a very small result of 1.339. The MSE value between 10% and 20% indicates that the forecasting model's ability is good. Furthermore, the RMSE and SAE values also produce good values, namely 1.157 and 94.135, respectively.

3.4 Hybrid ARIMA-ANN Model

After obtaining the ARIMA model for $PM_{2,5}$ in Jakarta, the error of the model is obtained. The forecast results from the ARIMA model are used as linear components. The error from the ARIMA model is modeled with the Artificial Neural Network model, which is used as a non-linear component. In this research, the input used is forecasting results and errors from the selected model, namely ARIMA (1,1,1)[27]. The results of error calculations from the ARIMA model, which is used as input for the ANN network are given in **Table 5** below:

Table 5. Input dan Output Pattern of Model ANN

Pattern	Prediction	Error	Actual $PM_{2,5}$
1	4.461129152	-0.005369437	4.455759716
2	4.452778307	-0.007516696	4.445261611
3	4.441468497	-0.008520661	4.432947836
4	4.428363625	-0.008422309	4.419941316
5	4.414694608	-0.006961434	4.407733174
⋮	⋮	⋮	⋮
286	4.033675266	0.000803552	4.034478818

The ANN architecture design used in this study consists of one input layer, a hidden layer, and an output layer. The output layer consists of one neuron, namely the error from the ARIMA model at time t which is used as the target value. While the number of neurons in the input layer and hidden layer with the trial & error process is limited to being tried from lag 1 to 5. Based on [28], it is not certain that a large number of neurons can obtain better accuracy. The following are the results of the trial process for selecting the number of neurons in the input layer:

Table 6. Testing the Number of Neurons in the input layer

Input	MSE
1	0.0004696045
2	0.0004710513
3	0.0004684316
4	**0.0004661953**
5	0.0004673620

Based on the experimental results, it was obtained that the number of neurons in the input layer was 4 lags. This is assumed from the smallest MSE results obtained from experiments using 4 lags. This indicates that today's $PM_{2,5}$ value is influenced by the $PM_{2,5}$ value from the previous 4 days. After obtaining the data normalization results, data processing was then carried out using the ANN method. Then, to determine the neurons in the hidden layer, trial and error is also carried out starting from 1-5 neurons [29]. Below are given the error values of several neurons.

Table 7. Number of neurons in the hidden layer and MSE

Amount of Neuron	MSE
1	0.0004868759
2	0.0004781988
3	**0.0004776726**
4	0.0004872067
5	0.0004872406

From the test results in Table 7, the smallest MSE value was obtained with the number of neurons in the hidden layer as many as 3. Next, a visualization of Figure 4 is given regarding the hybrid ARIMA-ANN architecture from $PM_{2,5}$ data where there are 4 neurons in the input, 3 neurons in the hidden layer, and 1 output.

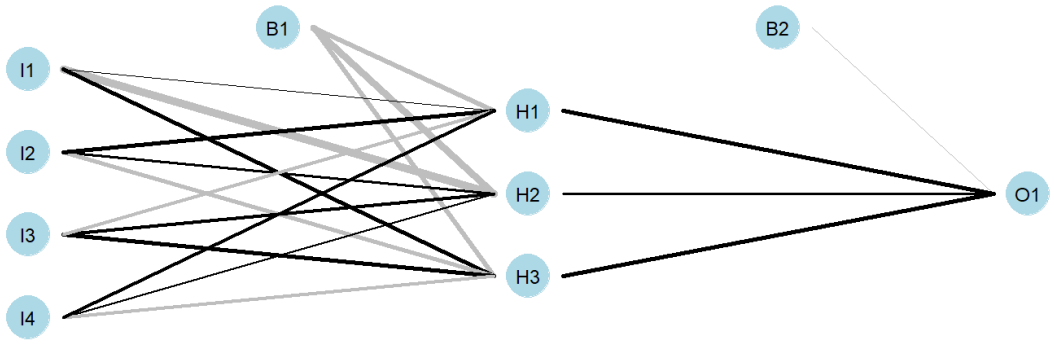


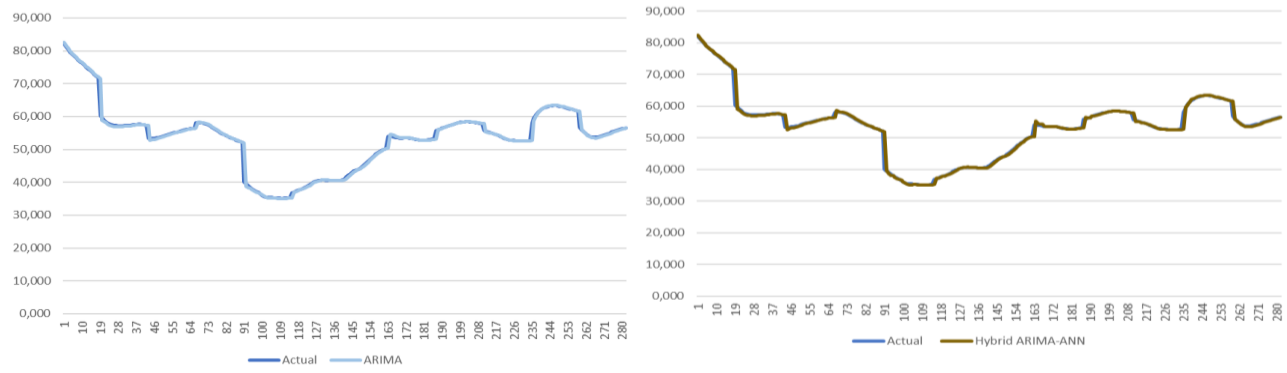
Figure 4. Architecture ANN (4-3-1)

The bias and weight values of the ANN 4-3-1 network are given as follows:

Table 8. Weights and Biases of ARIMA-ANN 4-3-1

Bias	Weight	Bias	Weight	Bias	Weight	Bias	Weight
$b \rightarrow h_1$	-2.98	$b \rightarrow h_2$	-0.26	$b \rightarrow h_3$	-0.99	$b \rightarrow o$	-0.77
$i_1 \rightarrow h_1$	0.09	$i_1 \rightarrow h_2$	0.50	$i_1 \rightarrow h_3$	2.60	$h_1 \rightarrow o$	1.53
$i_2 \rightarrow h_1$	0.34	$i_2 \rightarrow h_2$	-0.30	$i_2 \rightarrow h_3$	-0.54	$h_2 \rightarrow o$	-0.15
$i_3 \rightarrow h_1$	0.38	$i_3 \rightarrow h_2$	-0.04	$i_3 \rightarrow h_3$	-0.19	$h_3 \rightarrow o$	1.59
$i_4 \rightarrow h_1$	-0.18	$i_4 \rightarrow h_2$	-0.24	$i_4 \rightarrow h_3$	-1.30		

From the results of the formation of the ANN network and the determination of optimal weights and biases, the fit value for $PM_{2,5}$ data can then be obtained using a hybrid ARIMA-ANN. For the R^2 value, the results of the hybrid ARIMA-ANN modeling and the ARIMA model alone provide results that tend to be the same, which is around 98.62%. It is just that the SAE value of the hybrid ARIMA-ANN model provides a smaller value compared to the ARIMA model, which is 88.612. This shows that the hybrid ARIMA-ANN model can increase the accuracy of the forecasting model performance. Then, from the hybrid ARIMA-ANN equation, we can predict $PM_{2.5}$. The result plot of ARIMA and the hybrid ARIMA-ANN model given on the comparative visualization of the prediction results from the two models is shown in **Figure 5**.



(a) **(b)**

Figure 5. Comparison of Modeling Results With Actual Data : (a) ARIMA and (b) ARIMA-ANN

In **Figure 5**, it can be seen that the ARIMA and hybrid ARIMA-ANN models have almost the same plot as the actual data. This indicates that both models can capture the movement of PM data well. Therefore, to select the best model based on calculating SAE and MSE based on **Equation (3)**. From the error calculation results, especially the SAE value, it was found that the SAE value of the hybrid ARIMA-ANN model was smaller than the ARIMA model. However, from the error calculation results, especially the SAE value, it was found that the SAE value of the hybrid ARIMA-ANN model was smaller than the ARIMA model. This shows that the hybrid ARIMA-ANN model is better than the ARIMA model. Furthermore, the selected model is used to predict several periods. The hybrid ARIMA-ANN prediction results and criteria are shown in **Table 9** as follows:

Table 9. Hybrid ARIMA-ANN Forecasting Results

Periode	$PM_{2.5}$	Category
287	56.158	Moderate
288	56.147	Moderate
289	56.136	Moderate
290	56.125	Moderate
291	56.114	Moderate

From the prediction results of $PM_{2.5}$ for the next five periods are in the moderate criteria. This category of air levels does not have a significant impact on some healthy people, but the impact on health is more pronounced in sensitive groups, such as people with respiratory diseases or the elderly. However, air quality in the moderate category still requires attention, mitigation, and prevention measures to reduce exposure to

air pollution. With these prediction results, it can provide input to the Government and health agencies to guide to protection of sensitive groups in air quality conditions that are in the moderate category.

4. CONCLUSION

From the research that has been carried out, it is found that both the Classic ARIMA Statistical model and the hybrid ARIMA-ANN model are very good at describing the $PM_{2.5}$ data pattern. However, the hybrid ARIMA-ANN model provides lower SAE and MSE values compared to the ARIMA model. This shows that the hybrid model can increase model accuracy as well as being a better model compared to the ARIMA model.

The suggestions for further research are to forecast using other hybrid methods, such as ARIMA-RNN. The RNN (Recurrent Neural Network) method is specifically designed to process data that has a sequence or time, so that it is expected to increase prediction accuracy. In addition, in this study, the determination of parameters in the ANN architecture was carried out by trial and error, so there is still a possibility to optimize the determination of ANN parameters using the Genetic Algorithm or PSO.

AUTHOR CONTRIBUTIONS

Wahyuni Windasari: Conceptualization, Formal analysis, Investigation, Methodology, Writing-Review and Editing. Augistri Putri Pradani: Data Curation, Project Administration, Visualization, Writing-Original Draft. All authors discussed the results and contributed to the final manuscript.

FUNDING STATEMENT

This research has been funded by Universitas Putra Bangsa

ACKNOWLEDGMENT

The authors would like to express their gratitude and appreciation to Universitas Putra Bangsa for the support and facilities provided during this research.

CONFLICT OF INTEREST

The authors declare no conflicts of interest to report study.

REFERENCES

[1] World Health Organization, *WHO GLOBAL AIR QUALITY GUIDELINES.* Geneva, 2021.
[2] B. Doğan, O. M. Driha, D. B. Lorente, and U. Shahzad, "THE MITIGATING EFFECTS OF ECONOMIC COMPLEXITY AND RENEWABLE ENERGY ON CARBON EMISSIONS IN DEVELOPED COUNTRIES," *Sustain. Dev.*, vol. 29, no. 1, pp. 1–12, 2021, doi: https://doi.org/10.1002/sd.2125.
[3] A. Talaiekhozani and M. R. Talaie, "EVALUATION OF EMISSION INVENTORY OF AIR POLLUTANTS FROM RAILROAD AND AIR TRANSPORTATION IN ISFAHAN METROPOLITAN IN 2016," *Artic. J. Air Pollut. Heal.*, vol. 2, no. 1, pp. 1–18, 2017, [Online]. Available: http://japh.tums.ac.ir
[4] H. B. Wang F, Li Z, Zhang K, Di B, "AN OVERVIEW OF NON-ROAD EQUIPMENT EMISSIONS IN CHINA," *Atmos. Environ.*, vol. 132, pp. 283–289, 2016, doi: https://doi.org/10.1016/j.atmosenv.2016.02.046.
[5] IQAir, "WORLD'S MOST POLLUTED CITIES," 2021. https://www.iqair.com/world-most-polluted-cities?continent=59af92b13e70001c1bd78e53&country=&state=&sort=-rank&page=1&perPage=50&cities= (accessed Nov. 05, 2024).
[6] M. Hadei, S. Saeed, H. Nazari, E. Yarahmadi, M. Kermani, and M. Yarah-, "ESTIMATION OF LUNG CANCER MORTALITY ATTRIBUTED TO LONG-TERM EXPOSURE TO PM2.5 IN 15 IRANIAN CITIES DURING 2015 - 2016;

AN AIRQ+ MODELING," *J. Air Pollut. Heal.*, vol. 2, no. 1, pp. 19–26, 2017.
[7] G. Anchan, A., Shedthi, B.S., Manasa, "MODELS PREDICTING PM 2.5 CONCENTRATIONS—A REVIEW," in *Recent Advances in Artificial Intelligence and Data Engineering. Select Proceedings of AIDE 2020.*, 2021, pp. 65–83. doi: https://doi.org/10.1007/978-981-16-3342-3_6.
[8] J. J. Liaw, Y. F. Huang, C. H. Hsieh, D. C. Lin, and C. H. Luo, "PM2.5 CONCENTRATION ESTIMATION BASED ON IMAGE PROCESSING SCHEMES AND SIMPLE LINEAR REGRESSION," *Sensors (Switzerland)*, vol. 20, no. 8, pp. 1–13, 2020, doi: https://doi.org/10.3390/s20082423.
[9] H. H. Hamed *et al.*, "PREDICTING PM2.5 LEVELS OVER THE NORTH OF IRAQ USING REGRESSION ANALYSIS AND GEOGRAPHICAL INFORMATION SYSTEM (GIS) TECHNIQUES," *Geomatics, Nat. Hazards Risk*, vol. 12, no. 1, pp. 1778–1796, 2021, doi: https://doi.org/10.1080/19475705.2021.1946602.
[10] Z. Kapic, "MULTIPLE LINEAR REGRESSION MODEL FOR PREDICTING PM2.5 CONCENTRATION IN ZENICA," in *Advanced Technologies, Systems, and Applications V*, Springer, Cham, 2021, pp. 335–341. doi: https://doi.org/10.1007/978-3-030-54765-3_23.
[11] S. Gulati, A. Bansal, A. Pal, N. Mittal, A. Sharma, and F. Gared, "ESTIMATING PM2.5 UTILIZING MULTIPLE LINEAR REGRESSION AND ANN TECHNIQUES," *Sci. Rep.*, vol. 13, no. 1, pp. 1–12, 2023, doi: https://doi.org/10.1038/s41598-023-49717-7.
[12] A. P. Desvina, "PERAMALAN PENCEMARAN UDARA OLEH PARTICULATE MATTER (PM10) DI PEKANBARU DENGAN METODE BOX-JENKINS (FORECASTING OF AIR POLLUTION BY PARTICULATE MATTER (PM10) IN PEKANBARU WITH BOX-JENKINS METHOD)," pp. 63–73, 2015.
[13] B. Chrisdayanti and A. Suharsono, "PERAMALAN KANDUNGAN PARTICULATE MATTER (PM10) DALAM UDARA AMBIEN KOTA SURABAYA MENGGUNAKAN DOUBLE SEASONAL ARIMA (DSARIMA)," *J. Sains Dan Seni ITS*, vol. 4, no. 2, pp. 242–247, 2015.
[14] L. Zhang *et al.*, "TREND ANALYSIS AND FORECAST OF PM2.5 IN FUZHOU, CHINA USING THE ARIMA MODEL," *Ecol. Indic.*, vol. 95, no. 1, pp. 702–710, 2018, doi: https://doi.org/10.1016/j.ecolind.2018.08.032.
[15] B. N. Ruchjana, A. T. Arianto, K. Parmikanti, and B. Suhandi, "PERAMALAN KONSENTRASI PARTICULATE MATTER 2.5 (PM2.5) MENGGUNAKAN MODEL VECTOR AUTOREGRESSIVE DENGAN METODE MAXIMUM LIKELIHOOD ESTIMATION," *KUBIK J. Publ. Ilm. Mat.*, vol. 6, no. 1, pp. 1–12, 2021, doi: https://doi.org/10.15575/kubik.v6i1.8046.
[16] J. D. Kurniawan, "PREDIKSI KUALITAS UDARA BERBASIS MODEL LSTM DAN ARIMA," UNIVERSITAS KRISTEN SATYA WACANA, 2023.
[17] A. Agarwal and M. Sahu, "FORECASTING PM2.5 CONCENTRATIONS USING STATISTICAL MODELING FOR BENGALURU AND DELHI REGIONS," *Environ. Monit. Assess.*, vol. 195, no. 4, p. 502, 2023, doi: https://doi.org/10.1007/s10661-023-11045-8.
[18] M. Hadiyan Amaly, R. Haiban Hirzi, and P. Studi Statistika, "PERBANDINGAN METODE ANN BACKPROPAGATION DAN ARMA UNTUK PERAMALAN INFLASI DI INDONESIA," *Jambura J. Probab. Stat.*, vol. 3, no. 2, pp. 61–70, 2022.doi: https://doi.org/10.34312/jjps.v3i2.15440
[19] R. Wongsathan and I. Seedadan, "A HYBRID ARIMA AND NEURAL NETWORKS MODEL FOR PM-10 POLLUTION ESTIMATION : THE CASE OF CHIANG MAI CITY MOOR AREA," *Procedia - Procedia Comput. Sci.*, vol. 86, no. March, pp. 273–276, 2016, doi: https://doi.org/10.1016/j.procs.2016.05.057.
[20] S. et al. Zuo, X., Guo, H., Shi, "COMPARISON OF SIX MACHINE LEARNING METHODS FOR ESTIMATING PM2.5 CONCENTRATION USING THE HIMAWARI-8 AEROSOL OPTICAL DEPTH," *J. Indian Soc. Remote Sens.*, vol. 48, no. 9, pp. 1277–1287, 2020, doi: ttps://doi.org/10.1007/s12524-020-01154-z.
[21] L. Zhao, L., Li, Z., & Qu, "Forecasting of Beijing PM2.5 with a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition.," *Heliyon*, vol. 8, no. 12, 2022, doi: d https://doi.org/10.1016/j.heliyon.2022.e12239.
[22] E. Alshawarbeh, A. T. Abdulrahman, and E. Hussam, "STATISTICAL MODELING OF HIGH FREQUENCY DATASETS USING THE ARIMA-ANN HYBRID," *Mathematics*, vol. 11, no. 22, 2023, doi: https://doi.org/10.3390/math11224594
[23] O. A. Ejohwomu *et al.*, "MODELLING AND FORECASTING TEMPORAL PM2.5 CONCENTRATION USING ENSEMBLE MACHINE LEARNING METHODS," *Buildings*, vol. 12, no. 1, 2022, doi: https://doi.org/10.3390/buildings12010046.
[24] T. Liu, S. Liu, and L. Shi, "ARIMA MODELLING AND FORECASTING," in *Time Series Analysis Using SAS Enterprise Guide. SpringerBriefs in Statistics*, Singapore: Springer Singapore, 2020, pp. 61–85.doi: ttps://doi.org/10.1007/978-981-15-0321-4_4
[25] N. B. Tsae, T. Adachi, and Y. Kawamura, "APPLICATION OF ARTIFICIAL NEURAL NETWORK FOR THE PREDICTION OF COPPER ORE GRADE," *Minerals*, vol. 13, no. 5, pp. 1–18, 2023, doi: https://doi.org/10.3390/min13050658
[26] V. Ramasubramanian and M. Ray, "Package ' ARIMAANN .'" pp. 1–3, 2022. doi: 10.1016/S0925-2312(01)00702-0>.Encoding.
[27] A. Juliana, Hamidatun, and R. Muslima, *Modern Forecasting*. Yogyakarta: Deepublish, 2019.
[28] P. Wibowo, "PENGARUH PERBEDAAN JUMLAH HIDDEN LAYER DALAM JARINGAN SYARAF TIRUAN TERHADAP PREDIKSI KEBUTUHAN CAPTOPRIL DAN PARACETAMOL PADA RUMAH SAKIT," *Media Apl*, vol. 11, no. 2, pp. 118–131, 2019, doi: 10.33488/2.ma.2019.2.207.
[29] A. I. Taloba, "AN ARTIFICIAL NEURAL NETWORK MECHANISM FOR OPTIMIZING THE WATER TREATMENT PROCESS AND DESALINATION PROCESS," *Alexandria Eng. J.*, vol. 61, no. 12, pp. 9287–9295, 2022, doi: https://doi.org/10.1016/j.aej.2022.03.029.