Performance Comparison of Malaysian Air Pollution Index Prediction Using Nonlinear Autoregressive Exogenous Artificial Neural Network and Support Vector Machine

Rosminah Mustakim[1], and *Mazlina* Mamat[2]

[1]Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
[2]Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

Abstract. This paper compares the performance of Nonlinear Autoregressive Exogenous (NARX) Neural Network and Support Vector Machine (SVM) regression model to predict the Air Pollutant Index (API) in Malaysia. Two models namely the NARX and SVM regression were developed using the API and air quality time series data from three monitoring stations: Pasir Gudang, TTDI Jaya and Larkin. Hourly data of API and air quality parameters collected in year 2016 and 2018 were utilized to produce one step ahead API prediction. The air quality parameters consist of the NO2, SO2, CO, O3, PM2.5, PM10 concentration as well as three meteorological parameters which are wind speed, wind direction and ambient temperature. The NARX model was realized using a series-parallel feed-forward network. For the SVM regression model, different kernel functions: Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian and Coarse Gaussian were evaluated. The performance of NARX and SVM regression was measured using the Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) values. Results show that the NARX model outperformed the SVM regression model in both 2016 and 2018 data respectively.

1 Introduction

Air pollution has been a major concern amongst the developing and developed country across the globe for decades. As the population increases, so does the air pollution problem. Generally, air pollution occurs in the area when air pollutants in the atmosphere are high in concentration, exceeding the healthy air quality standard imposed. According to World Health Organization (WHO) database updated in May 2018, approximately 7 million deaths were estimated every year due to air pollution [1].

In Malaysia, the air quality level is supervised by the Department of Environment (DOE) under the Ministry of Energy, Science, Technology, Environment and Climate Change (MESTEC). The common air pollutants found in Malaysia are Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), Ozone (O_3) and Particulate Matter 10 (PM_{10}). These air pollutants are the air quality parameters used to calculate the Air Pollutant Index (API) in the country. However, starting 2017, the Particulate Matter 2.5 ($PM_{2.5}$) was added in the API calculation. The API reading indicates the air pollution level and is calculated based on the concentration of the air quality parameters. In practice, the API for each pollutant is calculated individually and the highest API will be selected as the API for the particular hour. In other word, the pollutant with the highest API will be the responsible pollutant for the published API value.

Many studies have been conducted actively in the past 20 years to find ways to reduce air pollution and minimize its impact on human health [2, 3]. Apart from that, studies were also made to predict the air quality level in advance [4,5]. The air quality prediction helps minimize the impact of air pollution to people by allowing them to take preventive actions accordingly. Among the prediction approaches to predict air quality is by using the machine learning techniques, a subfield of artificial intelligence. It generally operates by self-learning the pattern of a given historical dataset to make decision or prediction. The Artificial Neural Network (ANN) and Support Vector Machine (SVM) are among the popular methods applied for the purpose. Both have been explored in two air quality time series prediction studies as in Tehran [6] and Malaysia [7].

The study in Tehran analyzed four prediction models: Support Vector Machine (SVM) regression model, Geographically Weighted Regression (GWR), Basic Artificial Neural Network (ANN) and Nonlinear Autoregressive Exogenous (NARX) Neural Network, to predict the $PM_{2.5}$ and PM_{10} concentrations in Tehran. The results showed that NARX outperformed the other models with R^2 and RMSE of 0.99 and 0.72 respectively. SVM on the other hand was employed in a previous study conducted in Malaysia to predict the API. The research

* Corresponding author: mazlina@ums.edu.my

found that SVM with Radial Basis Function (RBF) kernel gave the best performance with R^2 value of 0.9843.

Motivated by these two studies, this research aims to develop the NARX and SVM models to predict the API in Malaysia and further on to compare their predicting performance. The result will be beneficial for DOE Malaysia to develop the API prediction system at their API monitoring stations in Malaysia.

2 Methodology

2.1 Air Quality Data

This research used the hourly air quality parameters data collected in 2016 and 2018 at the Continuous Ambient Air Quality Monitoring (CAQM) stations owned by Alam Sekitar Malaysia Sdn. Bhd. (ASMA) respectively. ASMA is the agency in charged to provide the air quality monitoring data for DOE Malaysia [8]. Data from three monitoring stations located in the industrial area: Pasir Gudang, TTDI Jaya and Larkin were used. These three monitoring stations were selected due to their location apart from consisting the most complete and continuous air quality data. The data for 2017 were excluded due to the addition of $PM_{2.5}$ concentrations as one of the air quality parameters in the middle of the year (July 4, 2017). This has contributed to conflicting API value for that year. Thus in 2016, the API was calculated without the $PM_{2.5}$ concentrations while in 2018 the API was calculated with the $PM_{2.5}$ concentrations.

Each set of the air quality data contains the concentration of pollutants: NO_2, SO_2, CO, O_3, PM_{10}, and $PM_{2.5}$ (for 2018 only); and meteorological parameters which affect the air pollution dispersion: wind speed (WS), wind direction (WD) and ambient temperature (T). In addition, the hourly API values were also included. The raw data received from the DOE were organized and preprocessed before being fed to the prediction models. The missing values were treated using mean of nearby points imputation technique by replacing them using a set of compromised values instead of using a fixed value for all the missing data. Using this technique, the mean of the two nearby data points was used to replace the missing values [9].

Outliers or values which differ significantly were identified and removed by using the Mahalanobis Distance analysis. The Mahalanobis Distance can be explained by equation 1.

$$d = \sqrt{(x - \bar{x})^T \cdot C^{-1} \cdot (x - \bar{x})} \quad (1)$$

Where d is the Mahalanobis Distance, x is the vector of the parameters or row in the datasets, $\bar{x}$ is the mean values vector of the parameters or mean for each column in the datasets and C^{-1} is the inverse covariance matrix of the parameters. The Mahalanobis Distance for each dataset was measured and compared to a chi-square distribution with the same degree of freedom to identify the outliers. The degree of freedom was corresponding to the number of the air quality parameters used to predict API. Using this method, it was found that around 1.3% to 3.3% of data were outliers and removed from the data.

2.2 The API Prediction Models

2.2.1 The Nonlinear Autoregressive Exogenous (NARX) Neural Network Model

The Artificial Neural Network (ANN) operates by imitating the human brain intelligence and learns the character or pattern of a given input data to make decision. ANN consists of neurons which positioned in a multi-layer network. The neurons in each layer are connected through weighted connections. Input data will enter ANN through the input layer and are processed through the multiple layers. The processed data then are transmitted to the output layer as the decision obtained based on the input data.

The Nonlinear Autoregressive Exogenous (NARX) Neural Network is a type of ANN [10]. It is a recurrent dynamic network which has feedback connection as its fundamental feature. Normally, the output of the NARX is fed back to the input of the network through a feedback connection. However, in this research, the NARX model was implemented using a feed-forward network with series-parallel architecture. The series-parallel architecture or open loop allowed the network to use the real output which is the past value of API as the input to the NARX instead of feeding back the estimated output which is the predicted API value. This gives the advantage of more accurate inputs for the network. The series-parallel architecture was applied while training the model to obtain the one step ahead prediction of API value. The NARX model is illustrated in Figure 1.

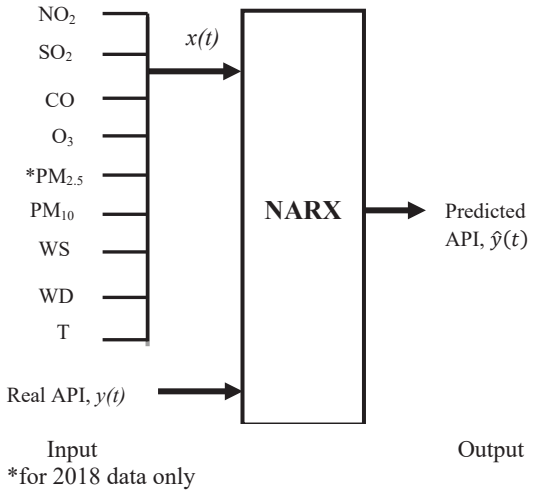

Fig. 1. Input and Output of the NARX model in a series parallel architecture.

The input and output of the NARX model architecture can be explained by equation 2.

$$\hat{y}(t) = f\left(x(t-1), \ldots x(t-d_x), y(t-1), \ldots y(t-d_y)\right) \quad (2)$$

where $\hat{y}(t)$ is the estimated output of NARX which is the predicted value of API at time *t*, f is the mapping function of the NARX, $x(t-1),\dots x(t-d_x)$ are the exogenous input of NARX which are the historical data of air quality parameters with input delay, d_x while $y(t-1),\dots y(t-d_y)$ are the real outputs i.e. the past value of API with the output delay, d_y.

All ANNs including NARX require a comprehensive amount of data for training and optimization before it can be employed as a prediction model. For this purpose, the data were divided into three sets: 75% for training and 15% each for testing and validation. The Levenberg-Marquardt was selected as the training algorithm for the NARX.

2.2.2 The Support Vector Machine (SVM) Regression Model

The Support Vector Machine (SVM) is commonly used to solve the classification and regression problems [11]. Initially, it was developed to perform linear classification but later expanded to solve the non-linear classification with the help of kernel trick. SVM has proven of having an excellent generalization ability and able to solve high dimensional problems [12].

In this research, the SVM regression was employed as the prediction model. Six kernel functions were identified and analyzed namely the Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian and Coarse Gaussian. To avoid over-fitting, the cross validation technique was employed during training where the data were partitioned into a number of folds. In this research, the number of folds was set to 10 [13]. The SVM regression model was trained using the full training dataset while the performance of the model was assessed using the data from each fold. The performance error for the SVM regression model was the average error over all folds. This technique helps to improve the prediction accuracy of the SVM regression model.

2.3 Prediction Performance

The performance of both NARX and SVM regression models was evaluated using the Root Mean Square Error (RMSE) and the coefficient of determination (R^2) analyses. The RMSE represents the standard deviation of the prediction errors while the R^2 represents the correlation between the predicted value and the actual value. This means that the model with lowest RMSE and highest R^2 values performs better prediction. Equation 3 and 4 presented the RMSE and R^2 respectively.

$$\mathbf{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(P_t - T_t)^2} \quad (3)$$

$$\mathbf{R^2} = \left(\frac{1}{N}\frac{\sum_{t=1}^{N}(P_t-\bar{P})(T_t-\bar{T})}{\sigma_P\sigma_T}\right)^2 \quad (4)$$

Where *N* is the number of datasets or hour in which the API value was taken, P_t is the predicted API value, T_t is the real API value, $\bar{P}$ is the mean for the predicted API value, $\bar{T}$ is the mean for the real API value, σ_P is the standard deviation of the predicted API value and σ_T is the standard deviation of the real API value.

3 Result and Discussion

The performance of NARX and SVM regression models was summarized in Table 1 and Table 2. Both tables show the average of RMSE and R^2 values obtained for the data collected from Pasir Gudang, TTDI Jaya and Larkin for year 2016 and 2018.

For the NARX model, different number of hidden neurons were tested starting from 2 to 30 as listed in Table 1. For year 2016, it can be observed that the RMSE values vary from 2.561 to 2.749 while for R^2 values, the range is from 0.943 to 0.951. For 2018, the RMSE values are from 0.978 to 1.085 while for R^2 values are from 0.975 to 0.981. These values show that the NARX was a steady model and its performance was not much affected by the number of hidden neuron used. However, it was suggested that the number of hidden neuron should be twelve to get the optimum performance in terms of computation and prediction accuracy.

Table 1. The performance of NARX model.

HIDDEN NEURON	2016		2018	
	RMSE	R^2	RMSE	R^2
2	2.749	0.943	1.072	0.975
4	2.683	0.946	1.046	0.977
6	2.632	0.948	1.039	0.977
8	2.657	0.947	1.085	0.974
10	2.628	0.948	1.027	0.978
12	**2.601**	**0.950**	**0.983**	**0.980**
14	2.594	0.949	1.010	0.978
16	2.572	0.951	0.978	0.980
18	2.600	0.950	1.049	0.977
20	2.639	0.948	1.007	0.978
22	2.650	0.949	1.011	0.980
24	2.621	0.949	0.979	0.979
26	2.561	0.951	0.994	0.978
28	2.561	0.951	1.001	0.981
30	2.606	0.949	1.005	0.978

For the SVM regression model, few kernel functions were tested and the performance was shown in Table 2. Based on the results, the Medium Gaussian kernel was identified to be the ideal kernel function for the SVM regression model. This kernel produced RMSE value of 3.642 and R^2 of 0.90 for 2016 data while for 2018 data, the RMSE was 5.503 and R^2 was 0.440.

Table 2. The performance of SVM regression model.

KERNAL FUNCTION	2016		2018	
	RMSE	R^2	RMSE	R^2
LINEAR	4.750	0.830	6.010	0.340
QUADRATIC	3.965	0.883	5.774	0.383
CUBIC	3.794	0.893	5.589	0.423
FINE GAUSSIAN	5.727	0.756	5.699	0.400
MEDIUM GAUSSIAN	**3.642**	**0.900**	**5.503**	**0.440**
COARSE GAUSSIAN	4.060	0.876	5.854	0.370

Table 1 and Table 2 show a significant difference in the performance of NARX and SVM regression models between 2016 and 2018 data. It can be said that the performance of both models was influenced by the pollutants used for the API prediction.

For comparison purpose, Table 3 listed the highest RMSE and R^2 values obtained by the NARX and SVM regression models. The results show that the NARX model was superior than the SVM regression model for both 2016 and 2018 data. The NARX model scored R^2 value of more than 0.950 for both years while the SVM regression model scored 0.900 for 2016 and only 0.440 for 2018. The scatter plots of the predicted versus real API for the NARX and SVM regression models are depicted by Figure 2 and Figure 3 respectively. It was evidenced that the predicted API using NARX model fall closer to the real API compared to the SVM regression model.

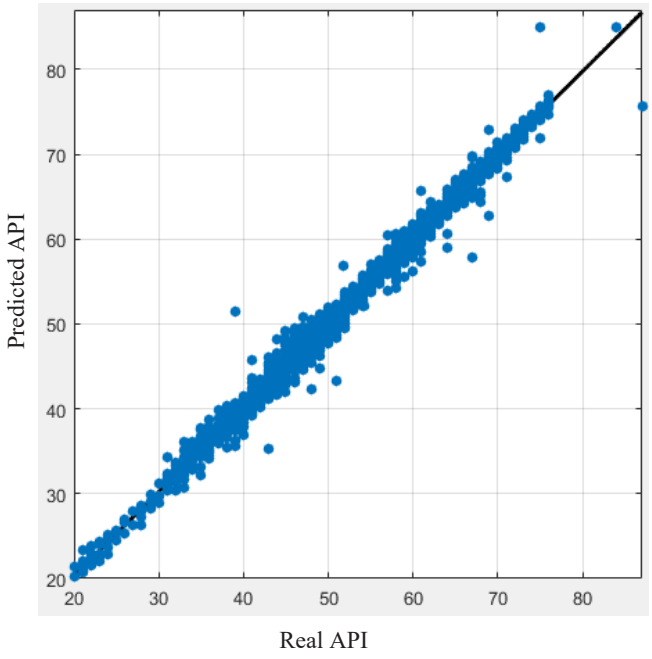

Fig. 3. The scatter plot of the Predicted versus real API of SVM regression model.

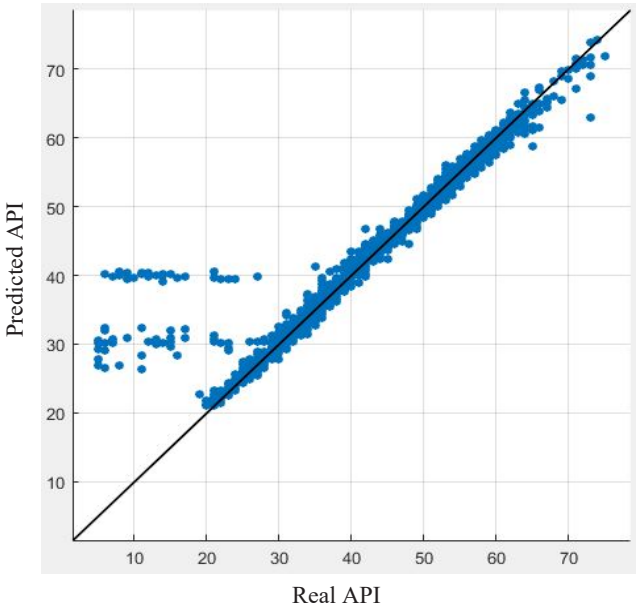

Fig. 2. The scatter plot of the Predicted versus real API for NARX Model.

Considering the RMSE value in Table 3, the addition of $PM_{2.5}$ as predictor in 2018 data has improved the performance of NARX model by approximately 62% while reducing the performance of SVM regression model by 51%. As the outcome of the missing data treatment and outlier removal, the real API value for 2016 and 2018 data were ranged between 20 to 90. The predicted value of API for both years are also fall within the same range in both NARX and SVM regression models.

Table 3. The performance of NARX and SVM regression models.

MODEL	2016		2018	
	RMSE	R^2	RMSE	R^2
NARX	**2.601**	**0.950**	**0.983**	**0.980**
SVM	3.642	0.900	5.503	0.440

4 Conclusion

This research compared the performance of two air quality prediction models namely the NARX and SVM regression models. Both models were trained and tested using the air quality data collected at three different locations: Pasir Gudang, TTDI Jaya and Larkin in 2016 and 2018. The parameters for each model were varied and the performances were recorded. Results showed that the NARX model produced more accurate and steady API prediction than the SVM regression model. This confirmed the conclusion made by the previous study conducted in Tehran where the NARX model was found superior than the SVM regression model. However, the results were for one step ahead prediction. Future studies should extend to the multiple steps ahead API prediction.

References

1. "WHO | Ambient and household air pollution and health," WHO, 2018. [Online]. Available: https://www.who.int/airpollution/data/en/.
2. D. J. Nowak, S. Hirabayashi, M. Doyle, M. McGovern, and J. Pasher, Urban For. Urban Green., **29**, 40–48 (2018)
3. M. Ben Jaber, A. Couvert, A. Amrane, F. Rouxel, P. Le Cloirec, E. Dumont, N. Biotechnol., **33**, 136–143 (2016)
4. D. Zhu, C. Cai, T. Yang, X. Zhou, Big Data and Cogn. Comput., **2**, 5 (2018)
5. A. K. Paschalidou, S. Karakitsios, S. Kleanthous, and P. A. Kassomenos, Environ. Sci. Pollut. Res., **18**, 316–327, (2011)
6. M. Delavar et al., ISPRS Int. J. Geo-Information, **8**, 99 (2019)
7. W. C. Leong, R. O. Kelani, and Z. Ahmad, J. Environ. Chem. Eng., 103208 (2019)
8. S. Shamsul, *A Study of Health Impact and Risk Assessment of Urban Air Pollution in the Klang Valley, Malaysia* (UKM Pakarunding Report, 2004)
9. S. Al, S. Dacey, J. Data Min. Knowl. Manag. Process, **7**, 75–91 (2017)
10. Z. Boussaada, O. Curea, A. Remaci, H. Camblong, N. M. Bellaaj, Energies, **11**, 620 (2018)
11. C. Cotes, V.N. Vapnik, Mach. Learn., **20**, 273-297 (1995)
12. B. Yganeh, M. S. P. Motlagh, Y. Rashidi, H. Kamalan, Atmos. Environ., **55**, 357–365 (2012)
13. G. James, D. Witten, T. Hastie, R. Tibshirani, *An introduction to Statistical Learning* (Springer, New York, NY, 2013)