atmosphere

MDPI

Article

Towards On-Site Implementation of Multi-Step Air Pollutant Index Prediction in Malaysia Industrial Area: Comparing the NARX Neural Network and Support Vector Regression

Rosminah Mustakim, Mazlina Mamat * and Hoe Tung Yew

Faculty of Engineering, Universiti Malaysia Sabah, Kota Kinabalu 88400, Sabah, Malaysia
* Correspondence: mazlina@ums.edu.my; Tel.: +60-1-932-0000

Abstract: Malaysia has experienced public health issues and economic losses due to air pollution problems. As the air pollution problem keeps increasing over time, studies on air quality prediction are also advancing. The air quality prediction can help reduce air pollution's damaging impact on public health and economic activities. This study develops and evaluates the Nonlinear Autoregressive Exogenous (NARX) Neural Network and Support Vector Regression (SVR) for multi-step Malaysia's Air Pollutant Index (API) prediction, focusing on the industrial areas. The performance of NARX and SVR was evaluated on four crucial aspects of on-site implementation: Input pre-processing, parameter selection, practical predictability limit, and robustness. Results show that both predictors exhibit almost comparable performance, in which the SVR slightly outperforms the NARX. The RMSE and R^2 values for the SVR are 0.71 and 0.99 in one-step-ahead prediction, gradually changing to 6.43 and 0.68 in 24-step-ahead prediction. Both predictors can also perform multi-step prediction by using the actual (non-normalized) data, hence are simpler to be implemented on-site. Removing several insignificant parameters did not affect the prediction performance, indicating that a uniform model can be used at all air quality monitoring stations in Malaysia's industrial areas. Nevertheless, SVR shows more resilience towards outliers and is also stable. Based on the trends exhibited by the Malaysia API data, a yearly update is sufficient for SVR due to its strength and stability. In conclusion, this study proposes that the SVR predictor could be implemented at air quality monitoring stations to provide API prediction information at least nine steps in advance.

Keywords: air quality prediction; Air Pollutant Index; Nonlinear Autoregressive Exogenous Neural Network; Support Vector Regression; multi-step-ahead prediction

Citation: Mustakim, R.; Mamat, M.; Yew, H.T. Towards On-Site Implementation of Multi-Step Air Pollutant Index Prediction in Malaysia Industrial Area: Comparing the NARX Neural Network and Support Vector Regression. *Atmosphere* **2022**, *13*, 1787. https://doi.org/10.3390/atmos13111787

Academic Editors: Esther Hontañón and Bernardete Ribeiro

Received: 30 September 2022
Accepted: 26 October 2022
Published: 29 October 2022

1. Introduction

Air pollution is a global issue that threatens the public health and economic activities of the worldwide population [1–3]. Without exception, Malaysia has experienced public health issues and economic losses due to air pollution problems [4,5]. Research by Tajudin et al. [6] reported that two air pollutants, namely Nitrogen Dioxide (NO_2) and Ozone (O_3), have an immediate effect on hospital admissions related to cardiovascular disease in Kuala Lumpur. Meanwhile, Ab Manan et al. [7] stated that the haze episode in 2013 cost Malaysians approximately MYR 410 million, accumulated from the medical expenses and income opportunity losses due to medical leave. Thus, the air pollution problem must be appropriately addressed to minimize its health effects. One solution is to predict air quality in advance. Knowing the air quality in advance can help the local administration issue early warning alerts to the residents so they can plan their activities accordingly.

Malaysia uses the Air Pollutant Index (API) to determine air quality. Malaysia, through APIMS (Air Pollutant Index Management System), has yet to develop a mechanism to predict API values in advance. There are, however, several apps that can provide the forecasted air quality index (AQI) for Malaysian cities; one such is Plume Labs: Air Quality

Apps. This app uses real-time data from the Malaysia Department of Environment (MDOE) to predict future AQI, but its accuracy is questionable. A brief comparison between the actual AQI for the Kuala Lumpur region provided by IQAir with the values predicted API by Plume Labs for 24 h (from 1 a.m., 28 August 2022 to 12 a.m., 29 August 2022) is plotted in Figure 1. The plots disagree, with large differences and an R^2 value of −0.2300. The low R^2 value indicates that the prediction made by Plume Labs has an accuracy issue.

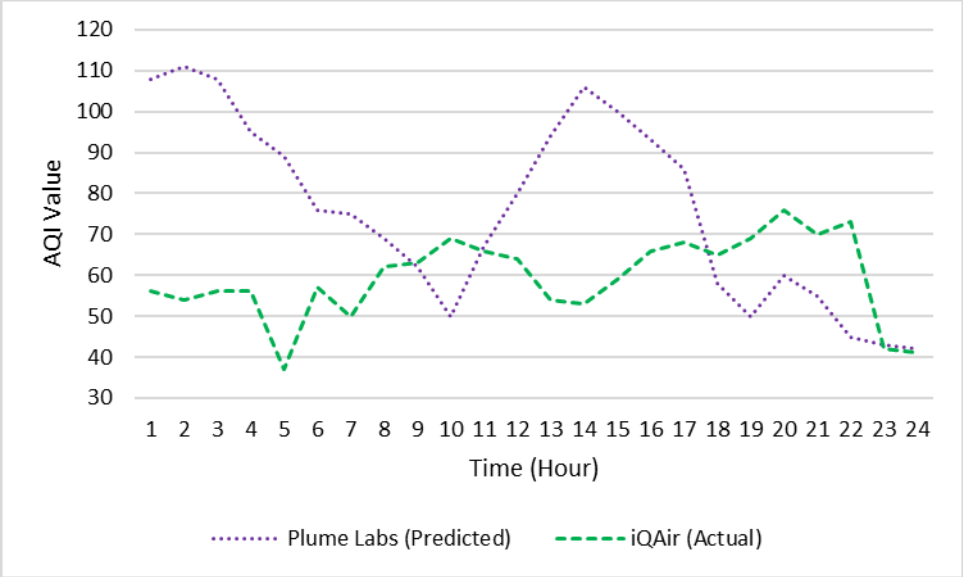

Figure 1. Predicted and actual AQI values for the Kuala Lumpur region for 24 h.

Researchers around the world have proposed many air quality prediction methods [8–12]. Among them, a technique based on the Nonlinear Autoregressive Exogenous (NARX) Neural Network was found superior in many publications. A study by Gündoğdu [13] established that NARX outperforms Multilayer Perceptron (MLP) in the one-step-ahead prediction of Particulate Matter 10 (PM_{10}) and Sulphur Dioxide (SO_2) concentrations. The RMSE values for NARX prediction of PM_{10} and SO_2 concentrations were 0.0191 and 0.0070, respectively, while MLP produced values of 0.0489 and 0.1121. Concurrently, NARX prediction of PM_{10} and SO_2 produced R^2 values of 0.9773 and 0.9984, while MLP produced values of 0.8530 and 0.6048. In another study, a popular machine learning algorithm called the Support Vector Machine (SVM) was used to predict the monthly average PM_{10} concentration seven months in advance [14]. The prediction performance was compared to MLP, Autoregressive Integrated Moving-Average (ARIMA), and Vector Autoregressive Moving-Average (VARMA). The results showed that SVM performs better than the other methods in one-step ahead and multi-step ahead predictions. The one-step-ahead prediction performances of SVM, ARIMA, MLP, and VARMA measured by RMSE were 2.061, 2.283, 3.432, and 3.451, respectively. Meanwhile, for multi-step ahead prediction, the RMSE of SVM was 1.990, followed by ARIMA (2.453), VARMA (3.121), and MLP (3.408).

A study employed NARX and SVM to predict the Air Quality Index (AQI) and concluded that NARX was better than SVM in one-step-ahead prediction [15]. The NARX gave an R^2 value of 0.9701, in contrast with SVM, which gave 0.8891. Another study compared the one-step-ahead prediction performance of NARX and SVM, amongst other methods, to predict $PM_{2.5}$ concentrations [16]. They concluded that NARX has better prediction performance than SVM, with R^2 and RMSE values of 0.99 and 0.72, respectively, while SVM gave 0.70 and 5.75.

Despite the superiority of NARX over SVM reported in the latter two publications, Kumar et al. [17] proved that SVM outperformed NARX in hourly wind speed prediction. The prediction performance measured by Mean Squared Error (MSE) was 52.32 for SVM

and 56.43 for NARX. Leong et al. [18] also achieved excellent API prediction using the SVM model. The research was conducted using the air quality data from 2009 to 2014 collected at eight monitoring stations in northern Malaysia. Prediction performance was measured in the R^2 value, and the SVM method achieved an R^2 of 0.9843 for one-step-ahead prediction. The superiority of NARX over other methods motivates this research to evaluate its performance in predicting the API in Malaysia's industrial areas. Since the SVM method was also proven to have excellent prediction performance using the Malaysia API, it will be evaluated and compared to NARX.

At present, scholars are more interested in proposing new methods to predict air quality [19–22]. Often, studies use the one-step-ahead prediction performance to evaluate the superiority of the proposed methods. We believe the evaluation should not stop at only comparing the prediction accuracy but rather extend it as if the proposed methods will be implemented on-site. Issues that might affect the prediction performance from the perspective of actual on-site implementation, such as input normalization, input parameters, practical predictability limit, and robustness, should be evaluated.

This paper addresses these four on-site implementation issues by comparing the performance of two established predictors, the NARX and SVM for regression (SVR). A careful analysis was designed and performed for each issue, providing valuable insight to researchers proposing new prediction methods. Apart from that, the outcomes of this study will make suggestions on how a multi-step-ahead API predictor for Malaysia API monitoring stations in industrial areas should be developed.

2. Materials and Methods

2.1. Study Area

Industrial activity is one of the major sources of air pollution [23,24]. Approximately 85% of air pollution in Malaysia comes from power plants emission [25]. Accordingly, this research focuses on air quality in three renowned industrial areas in Malaysia: TTDI Jaya, Larkin, and Pasir Gudang (Figure 2).

Figure 2. The location of the industrial areas.

These industrial areas are located nearby or surrounded by residential areas with a more than 1.2 million total population. The TTDI Jaya is in the Shah Alam district of Selangor. It is situated nearby Saujana Indah and the Hicom-Glenmarie industrial

park, among many other industrial areas. Food, cosmetics, and machinery are among the products manufactured in this industrial area. Larkin and Pasir Gudang are in Johor Bharu, south of peninsular Malaysia. The Larkin industrial area houses factories for plastic and metal fabrication, food products, glass manufacturing, electronic components, and mechanical machines. Most of the companies operating in the Pasir Gudang industrial area are heavy industries. This includes shipbuilding, palm oil storage and distribution, transportation and logistics, petrochemical, and construction.

2.2. Data Pre-Analysis and Treatment

The air quality data collected in 2018 and 2019 at these three industrial areas were provided by the Malaysia Department of Environment (MDOE). Each dataset contains hourly air quality parameters of Nitrogen Dioxide (NO_2), Ozone (O_3), Particulate Matter 2.5 ($PM_{2.5}$), Particulate Matter 10 (PM_{10}), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), and API. The hourly meteorological parameters, such as the ambient temperature (T), wind direction (WD), and wind speed (WS), were also provided in each dataset. A pre-analysis of the 2018 API parameter shows that the series does not exhibit seasonality for all three monitoring stations. The API values fluctuated randomly, mainly within the moderate level (50 to 100), with a maximum of 77 points and a minimum of 39 points. It can be concluded that the 2018 data represent the typical air quality in the three monitoring stations. Similar variations were observed in most parts of the 2019 data, except between September and November, when Malaysia was hit by a severe haze caused by the regional and transboundary haze from Indonesia. During the haze episode, the API reached an unhealthy level (101 to 200) and a very unhealthy level (201 to 300) for several weeks at the three monitoring stations.

Some missing values and outliers (less than 3.5%) were found in the raw air quality data provided by the MDOE. For the purposes of developing an optimized predictor, the missing values and outliers were replaced by the interpolated values using the Linear Interpolation Imputation method [26,27]. The Linear Interpolation Imputation method is explained by Equation (1), where $f(x)$ is the interpolated value of the missing value and the outlier x is the point at which the interpolation is performed. Variables x_0 and x_1 are the known values before and after the missing value, respectively.

$$f(x) = f(x_0) + (f(x_1) - f(x_0))/(x_1 - x_0)\,(x - x_0) \quad (1)$$

The outliers were determined by comparing them with the median data. The values that are more than three Median Absolute Deviations (MADs) away from the median value were replaced [28]. The scaled MAD is defined by Equation (2) where x_a is the average of the past values and x_i is the past values for each time step in the dataset.

$$\text{Scaled MAD} = (-1)/(\sqrt{2} \times erf\,cinv\,(3/2)) \times median\,(|x_i - x_a|) \quad (2)$$

Table 1 presents the data range and the correlation between each air quality parameter to the API for the three monitoring stations. The PM_{10} and $PM_{2.5}$ parameters show quite an obvious correlation with the API parameter compared to the other parameters in all three monitoring stations.

Table 1. Data range and correlation between the parameters and API.

2018											
Parameter		**NO_2**	**PM_{10}**	**$PM_{2.5}$**	**SO_2**	**CO**	**O_3**	**WD.**	**WS**	**T**	**API**
TTDI Jaya	Correlation	0.185	0.549	0.552	0.050	0.192	0.111	−0.045	−0.052	0.154	1.000
	Min	0.000	1.009	0.089	0.000	0.104	0.000	0.000	0.000	20.217	27.525
	Max	0.070	168.490	153.800	0.030	3.510	0.140	359.870	7.920	36.220	154.000
Larkin	Correlation	0.304	0.586	0.601	0.158	0.180	0.115	0.026	−0.021	0.060	1.000
	Min	0.000	0.718	0.120	0.000	0.050	0.000	0.000	0.000	21.423	12.456
	Max	0.070	172.660	163.550	0.020	2.760	0.150	359.970	22.790	35.390	91.760
Pasir Gudang	Correlation	0.323	0.579	0.576	0.319	0.189	0.093	0.084	−0.075	0.224	1.000
	Min	0.000	1.340	0.112	0.000	0.283	0.000	0.000	0.000	22.505	16.000
	Max	0.077	199.950	181.450	0.020	4.230	0.120	359.890	5.580	35.920	101.000
2019											
Parameter		**NO_2**	**PM_{10}**	**$PM_{2.5}$**	**SO_2**	**CO**	**O_3**	**WD.**	**WS**	**T**	**API**
TTDI Jaya	Correlation	0.084	0.511	0.533	−0.010	0.249	0.036	−0.018	0.026	0.068	1.000
	Min	0.000	2.315	0.046	0.000	0.044	0.000	0.000	0.000	21.282	33.000
	Max	0.060	264.640	252.310	0.020	3.200	0.160	359.930	5.950	36.270	221.000
Larkin	Correlation	0.354	0.674	0.686	0.167	0.280	0.148	−0.080	−0.187	0.136	1.000
	Min	0.000	3.728	0.090	0.000	0.052	0.000	0.000	0.000	21.843	26.000
	Max	0.070	294.900	264.720	0.020	3.250	0.120	359.950	4.750	35.280	171.000
Pasir Gudang	Correlation	0.399	0.679	0.695	0.458	0.261	0.137	−0.037	−0.017	0.186	1.000
	Min	0.000	1.1940	0.0850	0.000	0.077	0.000	0.000	0.000	23.897	18.000
	Max	0.070	173.90	161.630	0.010	2.900	0.110	359.940	6.250	37.550	143.00

2.3. Multi-Step Ahead Predictor

Three common strategies can be adapted in machine learning to perform multi-step-ahead prediction: Recursive, direct, and multiple outputs. The recursive strategy is the simplest and requires a single model with a single output. In the recursive approach, the predicted output at (t) is fed back as input to predict the output at (t + 1). Then the predicted output at (t + 1) is fed back as input to predict the output at (t + 2). The process continues until the desired step is achieved. The direct strategy requires n models to predict the outputs at (t + 1) to (t + n). Each model has a single output and is trained to predict a specific number of steps ahead of the output. Hence, ten models will be developed if the system wants to predict one to ten steps ahead. In many studies, the direct strategy produced more accurate multi-step ahead predictions [29,30]. On the other hand, a single model with n outputs is utilized in the multiple-outputs strategy to predict the (t + 1) to (t + n) values.

This paper employed the direct strategy to obtain the multi-step ahead prediction. In this study, 24 optimized models were used to obtain the hourly 1- to 24-step-ahead predictions, equivalent to a day-ahead prediction.

2.3.1. The Nonlinear Autoregressive Exogenous (NARX) Neural Network Model

NARX is a dynamic neural network with recurrent input fed by the feedback connection encircling the network layers [31]. A two-layer feed-forward NARX network that consists of a hidden layer and an output layer was used in this research. The sigmoidal transfer function is used as the hidden layer's transfer function, and the linear function was employed in the output layer. The NARX feedback connection was removed, making it a complete open-loop feed-forward network.

The inputs of the NARX model consist of the currently available air quality and meteorological parameters (NO_2, O_3, $PM_{2.5}$, PM_{10}, SO_2, CO, API, T, WD, and WS), while the output is the predicted future API values. Two hidden neurons were used in the hidden layer, determined by analysis in a preliminary study [32]. The NARX model employed the Levenberg–Marquardt algorithm for training. A total of 24 NARX models were developed and trained to obtain 1- to 24-step ahead prediction. Each unit in the 24 models was built from the s-step predictor depicted in Figure 3.

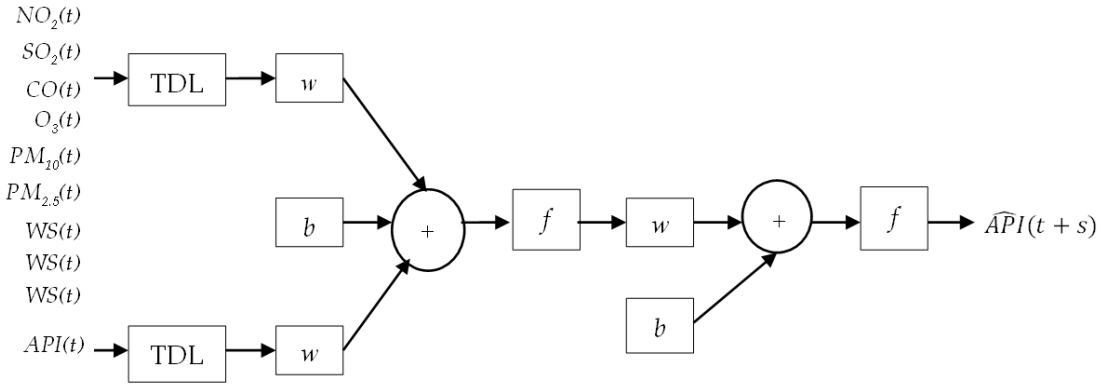

Figure 3. The NARX model for s-multi-step predictions.

2.3.2. The Support Vector Regression (SVR) Model

The Support Vector Machine (SVM) is a supervised machine learning approach widely used to solve classification problems [33]. The SVM can also be used to solve regression problems to predict discrete values and is usually referred to as Support Vector Regression (SVR). In SVR, a margin of tolerance known as epsilon is introduced to solve regression problems, which is the tolerated error for the SVR [34]. Similar to the classification problem, a kernel function was applied in SVR to solve the dimensional problem of nonlinear data. The well-tested kernel functions are Medium Gaussian, Coarse Gaussian, Fine Gaussian, Cubic, Quadratic, and Linear.

Figure 4 shows the SVR model used to perform the multi-step ahead prediction. The SVR inputs were fed with the currently available air quality and meteorological parameters, and the output was set to the s-step-ahead API value. The C and epsilon parameters were set to a default value during the training and testing stages. The default value of the C is set to the estimated value of the standard deviation using the interquartile range of the response variable y (the real API), while the default value of the epsilon is set to one-tenth of the C value. Twenty-four SVR models with the Linear kernel were employed using the direct approach to obtain the 24-step-ahead API prediction.

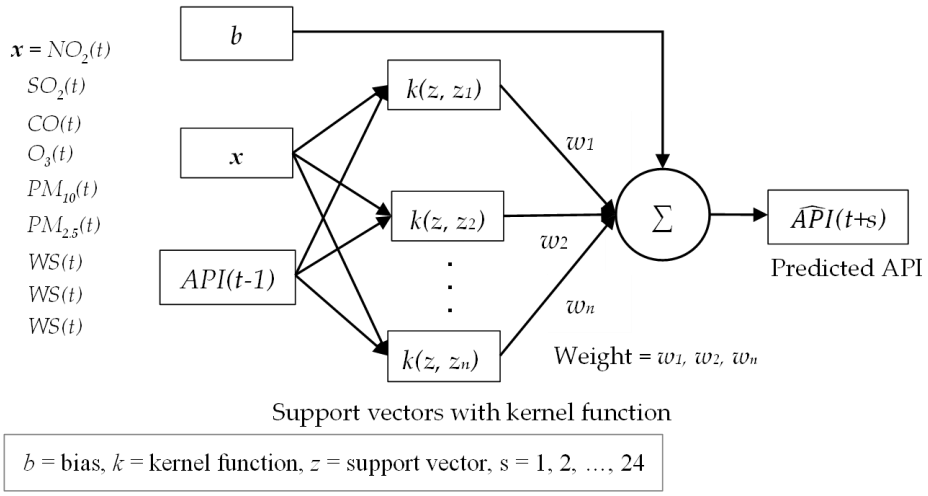

Figure 4. The SVR model for s-multi-step predictions.

2.4. Performance Indicator

RMSE and R^2 were used to assess the prediction performance of the NARX and SVM models. RMSE explains the prediction error or the difference between the predicted and

the actual value of API. The R^2 value represents the ratio of the variation in the predicted API value that can be explained by the linear association between the actual and predicted API values and the total variation of the predicted API value. Equations (3) and (4) define the RMSE and R^2, respectively.

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

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

Based on the equations, P_t is the predicted API while $\overline{P}$ is its mean, T_t is the actual value of API while $\overline{T}$ is its mean, N is the number of data points used in the measurement, σ_P is the standard deviation of the predicted API, and σ_T is the standard deviation for the actual value of API.

3. Results and Discussion

This study embarked on the following research questions:

1. Is input normalization required?
2. Which input parameters are important, and how do they affect the prediction performance?
3. How far can reasonable prediction be performed?
4. Which model is more robust?

The analyses were performed using 175,200 (10 parameters × 2 years × 365 days × 24 h) data, divided into training and testing in a ratio of 80 to 20. A large training dataset will reduce the risk of overfitting. However, during the model optimization process, the RMSE and R^2 for the training data were compared with the testing data to avoid overfitting or underfitting the model. The presented RMSE and R^2 values in the following subsections were obtained from the testing data.

3.1. Input Normalization

Each air quality and meteorological data collected at the three monitoring stations have values with differing scales, which may affect the prediction performance [35]. Applying normalization is suggested when dealing with such data [22,36]. However, the normalization approach depends on the machine learning architecture and the specific application [37]. Input normalization, if required, will introduce additional computational burdens and must be estimated correctly [38,39], and is tricky in real applications. In addition, the prediction values must be converted back to the original scale for reporting.

Considering those, an analysis was conducted to verify the need for input normalization. Here, the prediction performance of both predictors using normalized and raw data (non-normalization) was observed. Z-score normalization, or standardization, was performed on the data [40]. The z-score is calculated using Equation (5), where x is a data point in a feature with the mean $\overline{x}$ and standard deviation S.

$$z = (x - \overline{x})/S \tag{5}$$

The RMSE and R^2 values obtained by both predictors in all three monitoring stations are tabulated in Table 2. As expected, the RMSE value for normalized data is much smaller due to data rescaling and is not an accurate performance indicator. However, the smaller RMSE values obtained by SVR indicate that it is a better predictor than NARX. Further, it can be observed that the R^2 values scored by SVR and NARX are almost identical, implying that both predictors can predict aptly using raw data (non-normalized data). Performing

input normalization/standardization seems unnecessary as it does not affect the prediction performance.

Table 2. The RMSE and R^2 values using non-normalized and normalized data.

Monitoring Station	NARX				SVR			
	Non-Normalized		Normalized		Non-Normalized		Normalized	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Pasir Gudang	1.1800	0.9922	0.1282	0.9923	0.7106	0.9960	0.0787	0.9959
Larkin	1.2322	0.9883	0.1565	0.9886	0.7135	0.9948	0.0913	0.9948
TTDI Jaya	1.5797	0.9877	0.1842	0.9888	0.8914	0.9938	0.1025	0.9938

3.2. Input Parameters

We study the possibility that the API prediction can be performed using fewer parameters to reduce the computational burden. For this purpose, two prediction performance analyses were conducted using two different combinations of input parameters. The first analysis used all ten air quality and meteorological parameters, while the second analysis used selected parameters only. Parameter selection was performed using the Neighborhood Component Analysis (NCA). The NCA detects the relevant and irrelevant parameters in the data by learning the feature weights in an objective function that measures the training data regression loss [41]. The NCA results for the Pasir Gudang showed five relevant parameters: $PM_{2.5}$, CO, WD, WS, and T. Three parameters, namely PM_{10}, CO, and WD, were found to be relevant for the TTDI Jaya data. Meanwhile, all parameters were found to be relevant to the Larkin data. Figure 5 shows the NCA results for the three stations.

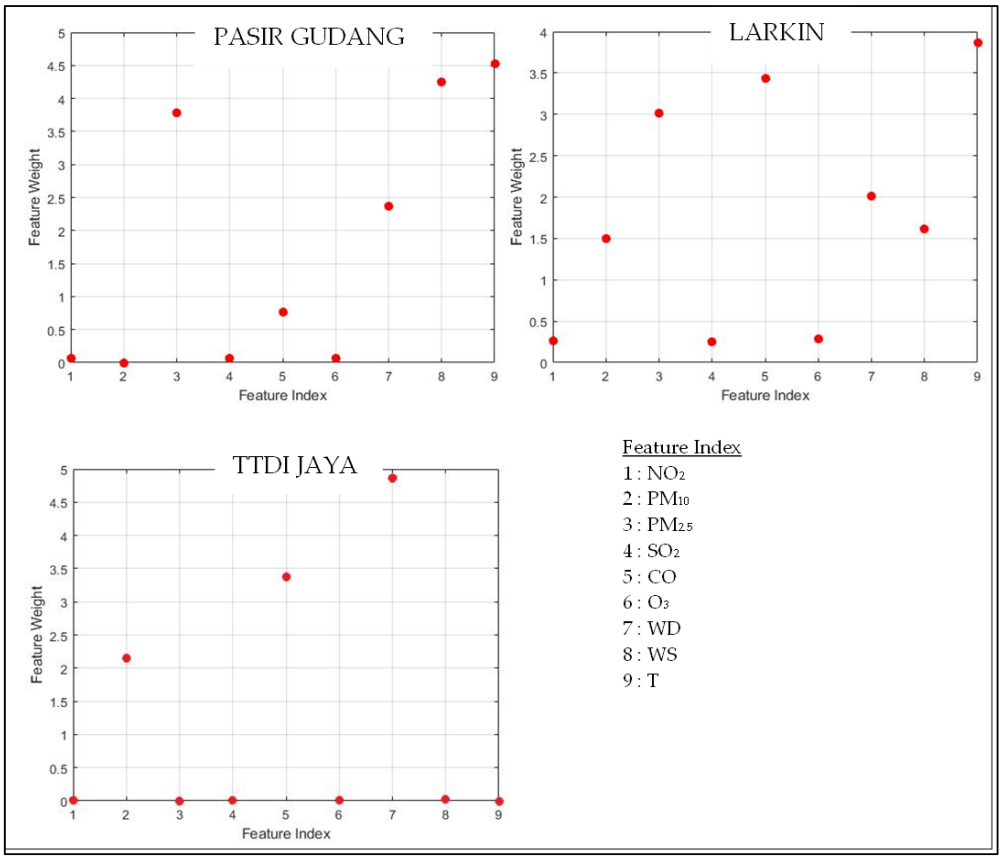

Figure 5. The parameter weights for each monitoring station based on the NCA for the parameter selection process.

Table 3 lists the RMSE and R^2 for one-step-ahead predictions of NARX and SVR for Pasir Gudang and TTDI Jaya stations using all parameters and only the relevant parameters, respectively. Using the relevant parameters seems to reduce the prediction error for Pasir Gudang but not for TTDI Jaya, using both NARX and SVR models. The negligible difference in the results proves that including all parameters, although unnecessary, will not hinder the prediction performance. This finding indicates that a universal predictor with a uniform structure can be built at every monitoring station in Malaysia without having to perform a preliminary analysis to obtain the relevant input parameters. A universal predictor with a uniform structure is preferred for easy installation at all stations.

Table 3. The RMSE and R^2 for all and relevant parameters.

Predictor	Monitoring Station	All Parameters		Relevant Parameters	
		RMSE	R^2	RMSE	R^2
NARX	Pasir Gudang	1.1800	0.9922	1.1926	0.9922
	TTDI Jaya	1.5797	0.9877	1.3186	0.9898
SVR	Pasir Gudang	0.7106	0.9960	0.7237	0.9959
	TTDI Jaya	0.8914	0.9938	0.8864	0.9939

3.3. Practical Predictibality Limit

The multi-step prediction performance in R^2 values for NARX and SVR predictors is tabulated in Table 4 and plotted in Figure 6. This analysis was performed using all ten non-normalized air quality parameters. From the plot, the prediction performance of both predictors decreases as the prediction steps progress, where the R^2 values gradually fall from 0.99 in one-step-ahead prediction to 0.68 in 24-step-ahead prediction. From the R^2 values, the NARX and SVR recorded a comparable performance for one- to six-step-ahead prediction. Beyond six-step-ahead prediction, NARX performs better prediction in all three stations. The SVR shows a more stable prediction for all 24-step-ahead predictions, whereby NARX's performance fluctuates. However, SVR recorded smaller RMSE values for all 24-step-ahead predictions for all three monitoring stations, compared to the NARX predictor (Table 5). This finding proves that the SVR is a better predictor.

Table 4. NARX and SVR multi-step-ahead prediction performance (in R^2 values).

Step Ahead Prediction	Pasir Gudang		Larkin		TTDI Jaya	
	NARX	SVR	NARX	SVR	NARX	SVR
1	0.9923	0.9959	0.9886	0.9948	0.9888	0.9938
2	0.9878	0.9909	0.9815	0.9883	0.9829	0.9866
3	0.9808	0.9841	0.9737	0.9794	0.9645	0.9771
4	0.9738	0.9758	0.9664	0.9689	0.9661	0.9676
5	0.9656	0.9671	0.9556	0.9570	0.9560	0.9564
6	0.9578	0.9573	0.9469	0.9443	0.9433	0.9446
7	0.9508	0.9474	0.9383	0.9308	0.9376	0.9323
8	0.9444	0.9366	0.9248	0.9171	0.9301	0.9196
9	0.9341	0.9260	0.9184	0.9024	0.9214	0.9067
10	0.9236	0.9151	0.9085	0.8877	0.9125	0.8934
11	0.9132	0.9042	0.8945	0.8731	0.9016	0.8789
12	0.9052	0.8924	0.8878	0.8577	0.8922	0.8664
13	0.8919	0.8812	0.8595	0.8439	0.8811	0.8526
14	0.8899	0.8699	0.8689	0.8278	0.8690	0.8407
15	0.8848	0.8586	0.8556	0.8141	0.8658	0.8269
16	0.8707	0.8474	0.8407	0.7994	0.8646	0.8142
17	0.8630	0.8371	0.8106	0.7831	0.8549	0.8002
18	0.8471	0.8262	0.8215	0.7707	0.8356	0.7887
19	0.8097	0.8159	0.8090	0.7562	0.7989	0.7757
20	0.8266	0.8050	0.7969	0.7429	0.8068	0.7630
21	0.8194	0.7938	0.7585	0.7306	0.7908	0.7529
22	0.7968	0.7835	0.7276	0.7164	0.7766	0.7393
23	0.7940	0.7721	0.6857	0.7030	0.7537	0.7258
24	0.7745	0.7600	0.7332	0.6881	0.7381	0.7120

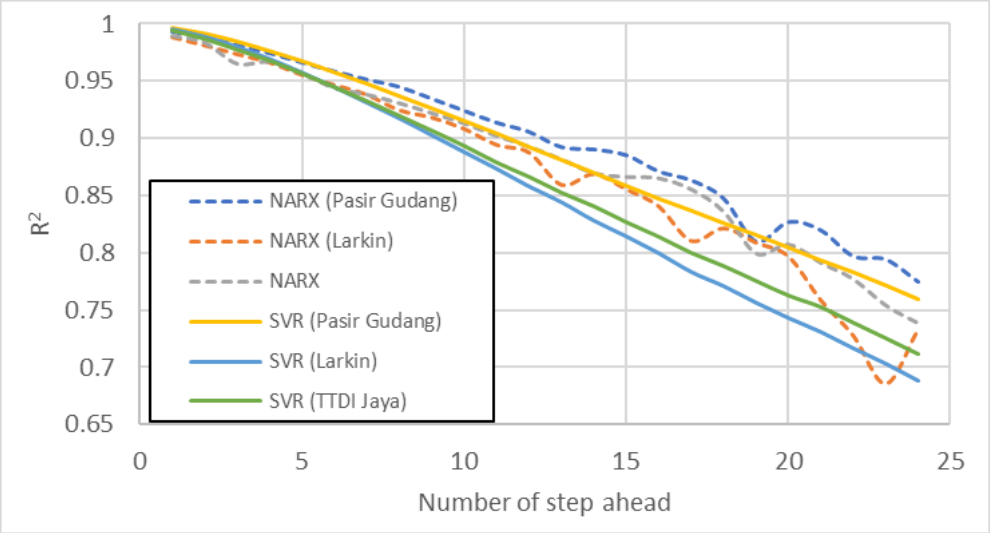

Figure 6. Multi-step prediction performance of NARX and SVR.

Table 5. NARX and SVR multi-step-ahead prediction performance (in RMSE values).

Step Ahead Prediction	Pasir GUDANG		Larkin		TTDI Jaya	
	NARX	SVR	NARX	SVR	NARX	SVR
1	1.1800	0.7106	1.2322	0.7135	1.5797	0.8914
2	1.5155	1.0731	1.5780	1.0718	2.2651	1.3052
3	2.0879	1.4193	1.8970	1.4207	2.3282	1.6941
4	2.2622	1.7528	2.2183	1.7512	2.6239	2.0248
5	2.6014	2.0468	2.5332	2.0363	2.8704	2.3275
6	2.8877	2.3319	2.7644	2.3160	2.9995	2.6360
7	3.3092	2.5902	2.9673	2.5807	3.5465	2.9113
8	3.4751	2.8384	3.1963	2.8220	3.4047	3.1732
9	3.9305	3.0722	3.3331	3.0566	3.7174	3.4352
10	4.4989	3.2913	3.6986	3.2731	3.7511	3.6636
11	4.7080	3.5036	3.8684	3.4839	4.0356	3.9097
12	4.8628	3.7186	4.0434	3.6825	4.2135	4.1243
13	5.0391	3.9023	4.3225	3.8701	4.5509	4.3245
14	5.3038	4.0881	4.6690	4.0409	4.4811	4.5119
15	5.6529	4.2832	4.9659	4.2230	4.7691	4.7013
16	5.9413	4.4577	5.1250	4.3928	4.7334	4.8971
17	5.7786	4.6249	5.5615	4.5542	5.3125	5.0703
18	6.2582	4.8255	5.5849	4.7254	5.1081	5.2186
19	6.5032	4.9927	5.5537	4.8777	5.4991	5.4078
20	7.1241	5.1766	6.2972	5.0167	5.6499	5.5586
21	7.0589	5.3722	6.0056	5.2104	5.8457	5.7227
22	7.0047	5.5242	6.6550	5.3780	6.4397	5.9267
23	7.5341	5.7317	7.5295	5.5276	6.2488	6.1325
24	8.0917	5.9580	7.2495	5.7433	6.5816	6.4285

The real and predicted API values for different R^2 values were plotted and observed to determine how far both predictors could reliably predict API. Findings from the three monitoring stations show that an R^2 of at least 0.90 can be considered sensible. As shown in Figure 7 for TTDI Jaya as an example, the real and predicted API values are closer with an R^2 of 0.90 or higher and deviate with an R^2 below 0.90. From Table 4, by using an R^2 of 0.90 as the lower limit, it can be derived that the NARX model can predict up to ten steps ahead while the SVR model can predict up to nine steps ahead.

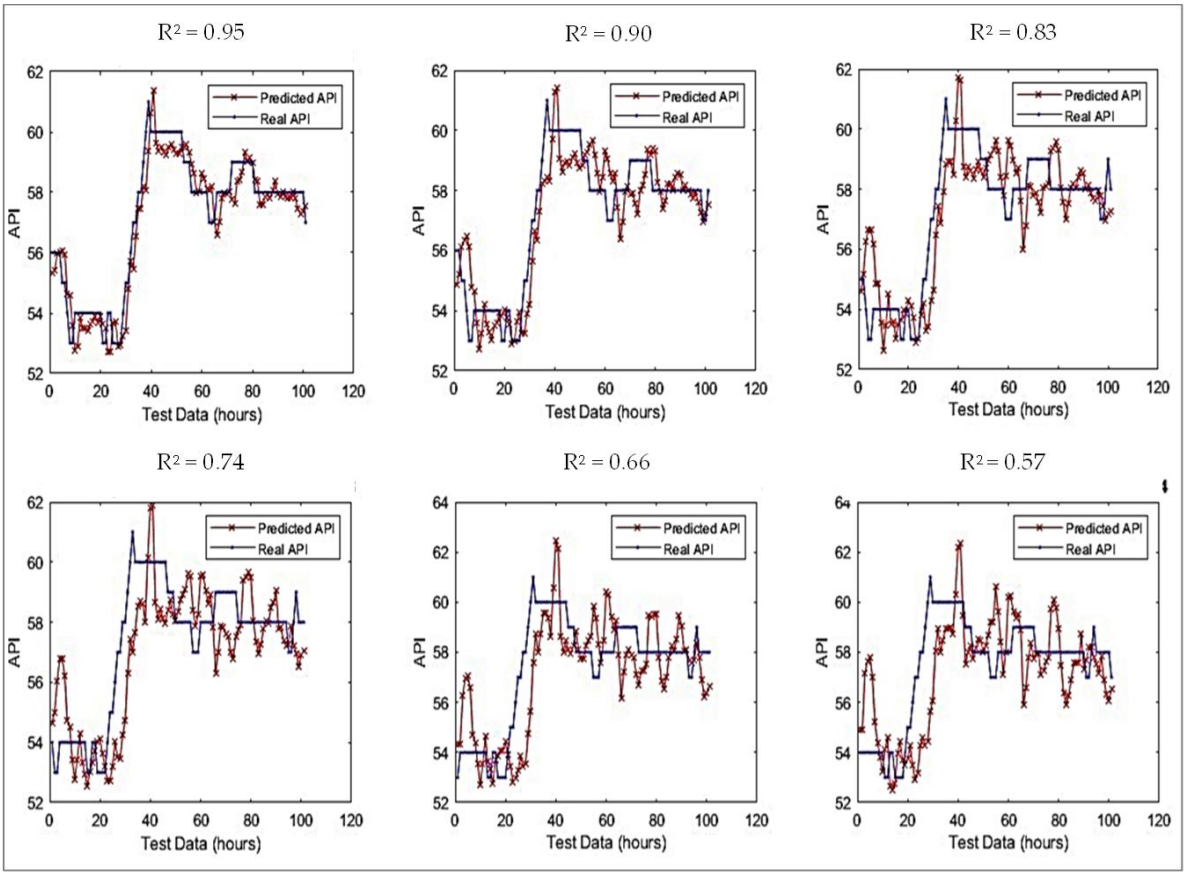

Figure 7. The actual and predicted API for different R^2 values.

3.4. Robustness

We also analyze the NARX and SVR predictors for their ability to perform reliable multi-step-ahead prediction for an extended duration to find which model requires frequent retraining and which is less susceptible to outliers. For this analysis, the NARX and SVR predictors were trained and validated using the 2018 data in the ratio of 80 to 20. The optimized predictors were tested using the 2019 data, and the RMSE values for one- to eight-step-ahead prediction, according to month, were computed. Results show that the SVR predictor produces more accurate and stable multi-step-ahead predictions than the NARX predictor in all three monitoring stations. SVR also seems more robust and makes better predictions during the haze episodes than NARX, hence is less susceptible to outliers. The same observation was also seen in all three monitoring stations.

Using Pasir Gudang data as an example (Figure 8), the RMSE values for one-step-ahead to eight-step-ahead predictions, calculated according to the respective month, are given in Table 6 and visualized in Figure 9. The results show that the SVR predictor recorded smaller RMSE values every month. Both predictors produced the worst performance for September; however, the SVR performed at least twice as good the NARX model. The worst prediction performance was due to the irregular trend of API values in September, where the API values rapidly increased and reached unhealthy status due to a haze episode that occurred during that time. Both predictors did not learn this trend during training resulting in lower prediction accuracy. The Larkin and TTDI Jaya data also exhibited a similar trend.

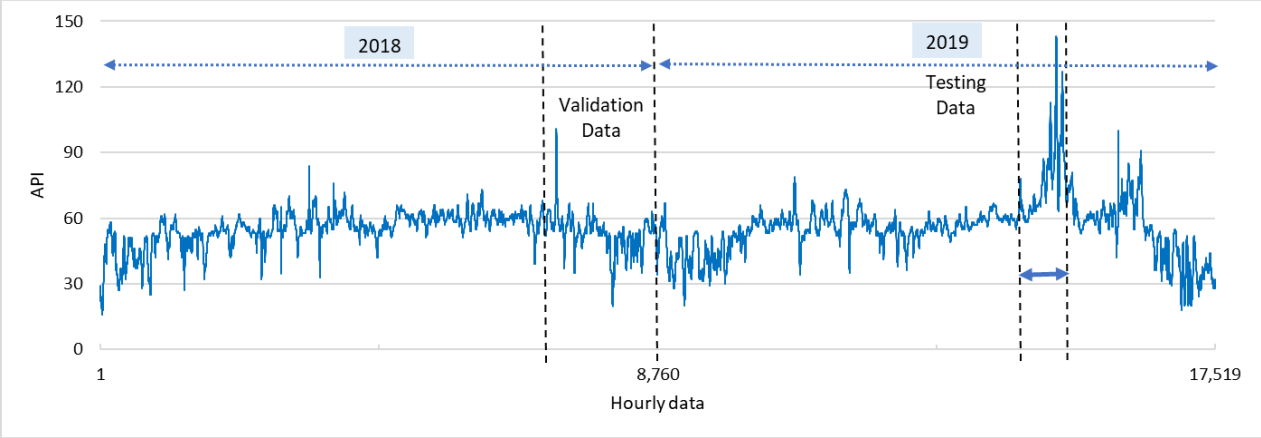

Figure 8. API values for Pasir Gudang in the years 2018 and 2019.

Table 6. RMSE values for one- to eight-step-ahead predictions, according to month.

Month	Predictor	Step Ahead							
		1	2	3	4	5	6	7	8
JAN	NARX	1.5012	2.0192	2.5765	3.0516	3.6440	3.9949	4.3751	4.863
	SVR	1.0083	1.5452	2.0231	2.4793	2.9155	3.3522	3.7866	4.1557
FEB	NARX	1.6690	2.1319	2.5442	2.9640	3.3068	3.5686	3.9424	4.1698
	SVR	1.0428	1.6669	2.1590	2.5811	2.9831	3.3427	3.6541	3.9496
MAR	NARX	1.0146	1.2100	1.4121	1.5875	1.7096	1.8502	2.1070	2.1575
	SVR	0.5847	0.8097	1.0146	1.2127	1.4194	1.5663	1.6991	1.8356
APR	NARX	1.6789	1.9057	2.0765	2.2994	2.4587	2.6509	2.8165	2.9065
	SVR	0.6792	1.0120	1.3578	1.6724	2.0073	2.2806	2.5263	2.7356
MAY	NARX	1.1885	1.4758	1.8475	2.1461	2.4324	2.5533	2.9494	3.1601
	SVR	0.7183	1.0866	1.4484	1.7961	2.1204	2.4282	2.7230	2.9844
JUNE	NARX	1.2823	1.6012	1.8989	2.1201	2.3598	2.6435	2.7535	2.9114
	SVR	0.7864	1.1608	1.5175	1.8294	2.0734	2.3118	2.5293	2.7184
JULY	NARX	0.8256	0.9408	1.0235	1.1289	1.1987	1.3069	1.4257	1.5252
	SVR	0.4587	0.5814	0.7260	0.8476	0.9831	1.0990	1.2257	1.3482
AUG	NARX	0.9711	1.0320	1.1474	1.2873	1.4461	1.5430	1.6986	1.8528
	SVR	0.4687	0.6090	0.7696	0.9259	1.0837	1.2512	1.4106	1.5599
SEPT	NARX	5.7433	9.0577	7.8669	10.487	11.1451	13.6487	14.762	15.1999
	SVR	1.0099	1.9108	2.8232	3.7294	4.6175	5.4374	6.2017	6.9276
OCT	NARX	2.2683	2.8434	3.3073	3.3148	3.5509	3.6516	3.7713	3.9953
	SVR	1.3716	2.3373	3.0473	3.4808	3.7054	3.8332	3.9658	4.1094
NOV	NARX	2.0890	2.2234	3.280	3.4581	3.8042	3.6561	4.9604	4.9131
	SVR	0.7519	1.2351	1.7034	2.1482	2.5528	2.9345	3.2891	3.6206
DEC	NARX	1.8162	4.4456	4.2650	3.7252	4.7170	5.2585	6.6709	6.5395
	SVR	1.1349	1.8912	2.5899	3.2491	3.8709	4.4628	5.0284	5.6076

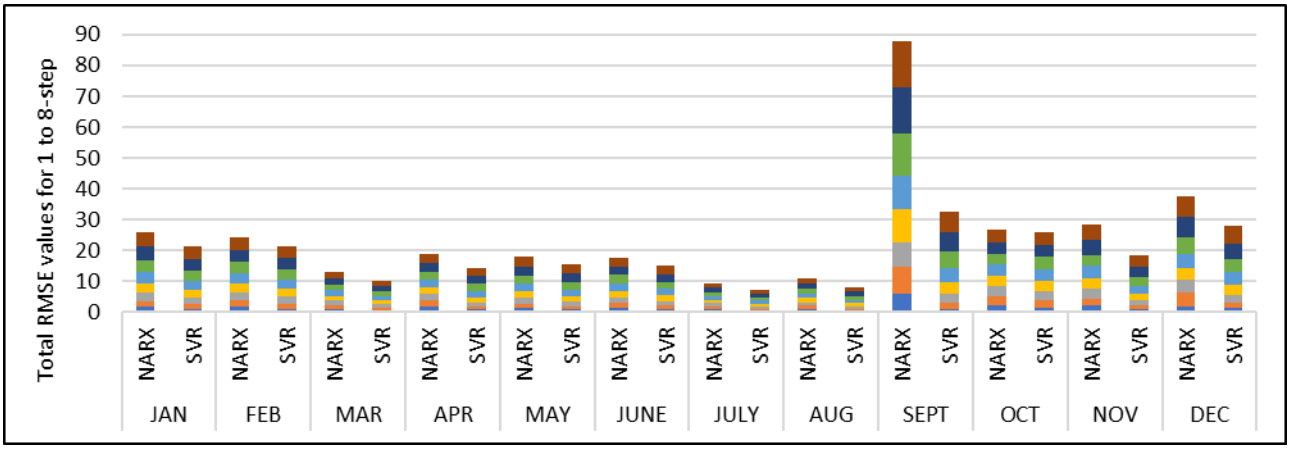

Figure 9. The performance of NARX and SVR, calculated monthly for the 2019 data.

4. Conclusions

The present study developed two multi-step-ahead API predictors based on NARX and SVR using Malaysia air quality data collected at three renowned industrial areas. Both predictors were evaluated for their ability to perform multi-step-ahead API prediction using the air quality parameters NO_2, O_3, $PM_{2.5}$, PM_{10}, SO_2, CO, and API and meteorological parameters T, WD, and WS. The analyses reveal that both predictors show comparable performance in multi-step API prediction, with the SVR slightly outperforming the NARX.

The SVR predictor can also perform multi-step prediction by using the actual (non-normalized) data, hence it is simpler to implement in actual applications. For uniformity, all air quality and meteorological parameters can be included as the predictor's inputs, as removing some parameters did not affect prediction performance. This finding indicates that a uniform SVR predictor can be installed in all air quality monitoring stations in Malaysia's industrial areas. Regarding robustness and the need for frequent retraining, SVR is also better than NARX as it shows more resilience towards outliers and is also stable. As Wang and Han [42] recommended, a predictor developed offline must be updated periodically to match the latest trends. However, based on the trends exhibited by the Malaysia API data, a yearly update is sufficient for SVR due to its resilience and stability. Based on the results, this study proposes that the SVR predictor could be applied practically to enhance MDOE service quality by providing API prediction information in advance.

As we advance, the SVR predictor should be immune to missing or false data for the API prediction to be reliable and without interruption. Thus, future research should focus on finding a supporting mechanism to provide continuous and valid data in case such a problem happens on-site. On the other hand, adaptive machine learning could be explored and adopted to deal with outliers.

Author Contributions: Conceptualization, R.M. and M.M.; methodology, R.M. and M.M.; software, R.M.; validation, R.M., M.M. and H.T.Y.; formal analysis, R.M. and M.M.; investigation, R.M.; resources, R.M.; data curation, R.M. and M.M.; writing—original draft preparation, R.M.; writing—review and editing, R.M., M.M. and H.T.Y.; visualization, R.M.; supervision, M.M. and H.T.Y.; project administration, M.M.; funding acquisition, M.M. and H.T.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Malaysia Sabah, grant number SBK0466-2021 and the Ministry of Higher Education, Fundamental Research Grant Scheme, FRGS/1/2020/TK0/UMS/02/2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Malaysian Department of Environmental (MDOE) for the air quality data provided for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Landrigan, P.J. Air Pollution and Health. *Lancet Public Health* **2017**, *2*, e4–e5. [CrossRef]
2. Shaddick, G.; Thomas, M.L.; Mudu, P.; Ruggeri, G.; Gumy, S. Half the World's Population Are Exposed to Increasing Air Pollution. *Npj Clim. Atmos. Sci.* **2020**, *3*, 23. [CrossRef]
3. Taghizadeh-Hesary, F.; Taghizadeh-Hesary, F. The Impacts of Air Pollution on Health and Economy in Southeast Asia. *Energies* **2020**, *13*, 1812. [CrossRef]
4. Hanafi, N.H.; Hassim, M.H.; Noor, Z.Z.; Ten, J.Y.; Aris, N.M.; Jalil, A.A. Economic Losses Due to Health Hazards Caused by Haze Event in Johor Bahru, Malaysia. In Proceedings of the 7th Conference on Emerging Energy and Process Technology, Johor Bahru, Malaysia, 27–28 November 2018. *E3S Web Conf.* **2019**, *90*, 01009. [CrossRef]
5. Usmani, R.S.A.; Saeed, A.; Abdullahi, A.M.; Pillai, T.R.; Jhanjhi, N.Z.; Hashem, I.A.T. Air Pollution and Its Health Impacts in Malaysia: A Review. *Air Qual. Atmos. Health* **2020**, *13*, 1093–1118. [CrossRef]
6. Tajudin, M.A.B.A.; Khan, M.F.; Mahiyuddin, W.R.W.; Hod, R.; Latif, M.T.; Hamid, A.H.; Rahman, S.A.; Sahani, M. Risk of Concentrations of Major Air Pollutants on the Prevalence of Cardiovascular and Respiratory Diseases in Urbanized Area of Kuala Lumpur, Malaysia. *Ecotoxicol. Environ. Saf.* **2019**, *171*, 290–300. [CrossRef]
7. Ab Manan, N.; Abdul Manaf, M.R.; Hod, R. The Malaysia Haze and Its Health Economic Impact: A Literature Review. *Malays. J. Public Health Med.* **2018**, *18*, 38–45.
8. Shaban, K.B.; Kadri, A.; Rezk, E. Urban Air Pollution Monitoring System with Forecasting Models. *IEEE Sens. J.* **2016**, *16*, 2598–2606. [CrossRef]
9. Lin, K.; Jing, L.; Wang, M.; Qiu, M.; Ji, Z. A Novel Long-Term Air Quality Forecasting Algorithm Based on KNN and NARX. In Proceedings of the ICCSE 2017—12th International Conference on Computer Science and Education, Houston, TX, USA, 22–25 August 2017; pp. 343–348. [CrossRef]
10. Mohebbi, M.R.; Jashni, A.K.; Jashni, K.; Dehghani, M.; Hadad, K. Short-Term Prediction of Carbon Monoxide Concentration Using Artificial Neural Network (NARX) without Traffic Data: Case Study: Shiraz City. *Iran. J. Sci. Technol. Trans. Civ. Eng.* **2018**, *43*, 533–540. [CrossRef]
11. Kang, G.K.; Gao, J.Z.; Chiao, S.; Lu, S.; Xie, G. Air Quality Prediction: Big Data and Machine Learning Approaches. *Int. J. Environ. Sci. Dev.* **2018**, *9*, 8–16. [CrossRef]
12. Zhou, Y.; Chang, F.J.; Chang, L.C.; Kao, I.F.; Wang, Y.S.; Kang, C.C. Multi-Output Support Vector Machine for Regional Multi-Step-Ahead $PM_{2.5}$ Forecasting. *Sci. Total Environ.* **2019**, *651*, 230–240. [CrossRef] [PubMed]
13. Gündoğdu, S. Comparison of Static MLP and Dynamic NARX Neural Networks for Forecasting of Atmospheric PM_{10} and SO_2 Concentrations in an Industrial Site of Turkey. *Environ. Forensics* **2020**, *21*, 363–374. [CrossRef]
14. García Nieto, P.J.; Sánchez Lasheras, F.; García-Gonzalo, E.; de Cos Juez, F.J. PM_{10} Concentration Forecasting in the Metropolitan Area of Oviedo (Northern Spain) Using Models Based on SVM, MLP, VARMA and ARIMA: A Case Study. *Sci. Total Environ.* **2018**, *621*, 753–761. [CrossRef] [PubMed]
15. Wang, L.; Bai, Y. Research on Prediction of Air Quality Index Based on NARX and SVM. *Appl. Mech. Mater.* **2014**, *602–605*, 3580–3584. [CrossRef]
16. Delavar, M.; Gholami, A.; Shiran, G.; Rashidi, Y.; Nakhaeizadeh, G.; Fedra, K.; Hatefi Afshar, S. A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 99. [CrossRef]
17. Kumar, V.; Pal, Y.; Tripathi, M.M. SVM Tuned NARX Method for Wind Speed Power Prediction in Electricity Generation. In Proceedings of the 8th IEEE Power India International Conference (PIICON 2018), Kurukshetra, India, 10–12 December 2018; pp. 1–6. [CrossRef]
18. Leong, W.C.; Kelani, R.O.; Ahmad, Z. Prediction of Air Pollution Index (API) Using Support Vector Machine (SVM). *J. Environ. Chem. Eng.* **2019**, *8*, 103208. [CrossRef]
19. Jiang, X.; Wei, P.; Luo, Y.; Li, Y. Air Pollutant Concentration Prediction Based on a CEEMDAN-FE-BiLSTM Model. *Atmosphere* **2021**, *12*, 1452. [CrossRef]
20. Muthukumar, P.; Nagrecha, K.; Comer, D.; Calvert, C.F.; Amini, N.; Holm, J.; Pourhomayoun, M. $PM_{2.5}$ Air Pollution Prediction through Deep Learning Using Multisource Meteorological, Wildfire, and Heat Data. *Atmosphere* **2022**, *13*, 822. [CrossRef]
21. He, Z.; Guo, Q.; Wang, Z.; Li, X. Prediction of Monthly $PM_{2.5}$ Concentration in Liaocheng in China Employing Artificial Neural Network. *Atmosphere* **2022**, *13*, 1221. [CrossRef]
22. Wei, F.; Zhu, R.; Jerry Chun, W.L. An air quality prediction model based on improved Vanilla LSTM with multichannel input and multiroute output. *Expert Syst. Appl.* **2022**, *211*, 118422. [CrossRef]
23. Raffee, A.F.; Rahmat, S.N.; Hamid, H.A.; Jaffar, M.I. The Behavior of Particulate Matter (PM_{10}) Concentrations at Industrial Sites in Malaysia. *Int. J. Integr. Eng.* **2019**, *11*, 214–222. [CrossRef]

24. Azid, A.; Juahir, H.; Toriman, M.E.; Endut, A.; Kamarudin, M.K.A.; Rahman, M.N.A.; Hasnam, C.N.C.; Saudi, A.S.M.; Yunus, K. Source Apportionment of Air Pollution: A Case Study in Malaysia. *J. Teknol.* **2015**, *72*, 83–88. [CrossRef]
25. Sentian, J.; Herman, F.; Yih, C.Y.; Hian Wui, J.C. Long-Term Air Pollution Trend Analysis in Malaysia. *Int. J. Environ. Impacts Manag. Mitig. Recover.* **2019**, *2*, 309–324. [CrossRef]
26. Mohamed Noor, N.; Al Bakri Abdullah, M.M.; Yahaya, A.S.; Ramli, N.A. Filling Missing Data Using Interpolation Methods: Study on the Effect of Fitting Distribution. *Key Eng. Mater.* **2014**, *594–595*, 889–895. [CrossRef]
27. Mohamed Noor, N.; Al Bakri Abdullah, M.M.; Yahaya, A.S.; Ramli, N.A. Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. *Mater. Sci. Forum* **2015**, *803*, 278–281. [CrossRef]
28. Fitriyah, H.; Budi, A.S. Outlier Detection in Object Counting Based on Hue and Distance Transform Using Median Absolute Deviation (MAD). In Proceedings of the 2019 4th International Conference on Sustainable Information Engineering and Technology (SIET 2019), Lombok, Indonesia, 28–30 September 2019; pp. 217–222. [CrossRef]
29. Mamat, M.; Samad, S.A. Comparison of Iterative and Direct Approaches for Multi-Steps Ahead Time Series Forecasting Using Adaptive Hybrid-RBF Neural Network. In Proceedings of the IEEE Region 10 Annual International Conference, Fukuoka, Japan, 21–24 November 2010; pp. 2332–2337. [CrossRef]
30. López Pouso, Ó.; Jumaniyazov, N. Direct versus iterative methods for forward-backward diffusion equations. Numerical comparisons. *SeMA* **2021**, *78*, 271–286. [CrossRef]
31. Boussaada, Z.; Curea, O.; Remaci, A.; Camblong, H.; Bellaaj, N.M. A Nonlinear Autoregressive Exogenous (NARX) Neural Network Model for the Prediction of the Daily Direct Solar Radiation. *Energies* **2018**, *11*, 620. [CrossRef]
32. Mustakim, R.; Mamat, M. The Nonlinear Autoregressive Exogenous Neural Network Performance in Predicting Malaysia Air Pollutant Index. *Trans. Sci. Technol.* **2021**, *8*, 305–310.
33. Cortes, C.; Vapnik, V. Support-Vector Network. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
34. Drucker, H.; Surges, C.J.C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. *Adv. Neural Inf. Process. Syst.* **1997**, *1*, 155–161.
35. Falocchi, M.; Zardi, D.; Giovannini, L. Meteorological Normalization of NO_2 Concentrations in the Province of Bolzano (Italian Alps). *Atmos. Environ.* **2021**, *246*, 118048. [CrossRef]
36. Bengio, Y.; Goodfellow, I.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
37. Platt, J.A.; Penny, S.G.; Smith, T.A.; Chen, T.-C.; Abarbanel, H.D.I. A Systematic Exploration of Reservoir Computing for Forecasting Complex Spatiotemporal Dynamics. *Neural Netw.* **2022**, *153*, 530–552. [CrossRef] [PubMed]
38. Passalis, N.; Tefas, A.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A. Deep Adaptive Input Normalization for Time Series Forecasting. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 3760–3765. [CrossRef]
39. Passalis, N.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A.; Tefas, A. Forecasting Financial Time Series Using Robust Deep Adaptive Input Normalization. *Signal Process. Syst.* **2021**, *93*, 1235–1251. [CrossRef]
40. Gupta, M.; Wadhvani, R.; Rasool, A. Real-Time Change-Point Detection: A Deep Neural Network-Based Adaptive Approach for Detecting Changes in Multivariate Time Series Data. *Expert Syst. Appl.* **2022**, *209*, 118260. [CrossRef]
41. Djerioui, M.; Brik, Y.; Ladjal, M.; Attallah, B. Neighborhood Component Analysis and Support Vector Machines for Heart Disease Prediction. *Ing. Yst. d'Inform.* **2019**, *24*, 591–595. [CrossRef]
42. Wang, Y.; Han, L. Adaptive Time Series Prediction and Recommendation. *Inf. Process. Manag.* **2021**, *58*, 102494. [CrossRef]