




# Toward forecasting future day air pollutant index in Malaysia

Kok-Seng Wong<sup>1</sup> · Yee Jian Chew<sup>2</sup> · Shih Yin Ooi<sup>2</sup>  · Ying Han Pang<sup>2</sup>

Accepted: 14 October 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

The association of air pollution and the magnitude of adverse health effects are receiving close attention from the world. The effects of air pollution were found to be most significant for children, elderly, and patients with preexisting respiratory problems. The existing API forecast system is capable of predicting the air quality based on the pollutant concentrations before critical levels of air pollution are exceeded. However, there is no API forecasting system available in Malaysia that can predict the coming day API readings. This paper aims to propose an API forecast system that utilizes the hourly API in Malaysia to predict the next day API. The proposed solution allows sensitive populations to plan ahead of their daily activities and provide governments with information for public health alerts. We also propose strategies for aggregated-level predictions within the region. Nevertheless, it can be extended across the region, especially in the less economically developed regions across the world. We conduct experiments on the public API dataset to demonstrate the viability of the proposed solution.

**Keywords** Air pollution · Air quality forecasting · Machine learning · Public awareness

## 1 Introduction

Anthropogenic air pollution is a global and ongoing environmental problem faces by a majority of cities in the world. As reported by the World Health Organization (WHO), air pollution kills an estimated seven million people worldwide every year [1]. The rapid population growth accompanied by increased air pollution is now a significant challenge in developing countries, especially China and Southeast Asia

---

✉ Shih Yin Ooi  
syooi@mmu.edu.my

<sup>1</sup> College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam

<sup>2</sup> Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia

because urban areas in these countries are emerging as pollution hot spots. For instance, several cities in China have been identified as the most polluted city in the world [2].

In general, there are two leading causes of air pollution: naturally occurring phenomena and the human factor. The first cause is subject to natural sources during catastrophes such as forest fire and volcanic eruptions. In the second cause, pollutants released from transportation, fuel combustion, and industries that involved human activities. The effects of air pollution result from human activity can have a long-term impact on the environment, including climate change, global warming, and acid rain. The mixture of particles and gases can reach harmful concentrations on human health and the planet as a whole. Some of the common air pollutants are carbon dioxide ( $\text{CO}_2$ ), particulate matter ( $\text{PM}_{2.5}$  and  $\text{PM}_{10}$ ), nitrogen dioxide ( $\text{NO}_2$ ), carbon monoxide ( $\text{CO}$ ), and sulfur dioxide ( $\text{SO}_2$ ). Air pollution in the form of  $\text{CO}_2$  and methane contributes to climate change, e.g., rising temperatures. This climate change then causes warmer weather and the production of allergenic air pollutants such as mold and pollen.

The urban air quality database (covering 3000 cities in 103 countries) shows that more than 80% of people living in urban areas are affected by air pollution [3]. Those living in these areas are facing severe health impacts, including the increase of respiratory symptoms, mortality from stroke, heart disease, and aggravated asthma. Notably, they are exposed to air quality levels that exceed WHO guideline limits. Undoubtedly, air pollution is a major environmental health threat to the world. In a recent public awareness survey report released by Mid-America Regional Council [4], 67% of the residents surveyed were concerned (very or somewhat concerned) about the health consequences of poor air quality in the Kansas City area. The report also shows that 33% of residents felt air pollution in the Kansas City area was getting worse.

Air quality statistics are essential information used to show how polluted the air has become in a targeted area. Generally, the statistics reflect the levels of pollutants, i.e., the maximum allowable limits for each pollutant. The air quality statistics are used by different countries to set their air quality indices that correspond to different national air quality standards. For example, in the USA, the Environmental Protection Agency uses the Air Quality Index (AQI) for reporting daily air quality, whereas in Canada, air quality health index (AQHI) is designed to help the local authority to understand the impact air pollution on health [5].

## 1.1 Effects of air pollution

The state of air pollution is often expressed as air quality and measured by the concentrations of air pollutants. However, it is not valid to discuss air quality without identifying the impacts of air pollution [6]. The main effect of air pollution is human health, and we can foresee possible short-term and long-term health effects on the people living in a heavily polluted area. On the one hand, short-term health effects may lead to acute conditions such as respiratory infections and conjunctivitis. On the other hand, the long-term effects dominate the public health impacts of air

pollution, and there is growing evidence of a close relationship between long-term health effects and cardiovascular disease [7]. For example, heart disease, lung cancer, and respiratory diseases are long-term effects that can last for years or the entire lifetime of the patient.

The effects of air pollution were found to be most significant for children, the elderly, and patients with preexisting respiratory problems. The association of air pollution and the magnitude of adverse health effects are receiving close attention from the world. The recent statistics from WHO show that 93% of all children live in heavily polluted environments, and more than one in every four deaths of children under five years of age is linked to environmental risks [8]. In 2016, a total of 5,43,000 deaths in children under the age of five years was recorded due to respiratory tract infections. Whether air pollution affects female infertility is another hot issue under debate [9]. Besides the human health impacts, air pollution has further consequences on the social, economic, and lifestyle habits. In developing countries, the lack of awareness regarding the air pollution problem leads to a more severe environmental crisis. Despite the difficulty of sustainable management of the environment, avoidance behavior among the high-risk population is achievable; however, it requires much effort from the local authorities and possible coordination and collaboration across regions.

## 1.2 Air pollution issue in Malaysia

Due to the industrial growth and rapid urbanization in Malaysian cities, the country faces the problems of urban air pollution, just like other developing countries in the Southeast Asia region. The primary types of pollution sources are emissions from the industrial, energy generation, and transportation sectors, as well as transboundary haze pollution due to forest fires and open burning. The forest fires have been an annual occurrence in Indonesia for generations [10]. Fires have been used for land clearing because burning is the cheapest method of clearing vast areas. Every year during the southwest monsoon, the forest fires in Indonesia will cause the air pollution level in Malaysia to increase significantly [11]. The worst haze pollution in Malaysia caused by the forest fires from Indonesia was recorded in the second half of 1997. It was a large-scale air quality disaster causing widespread atmospheric visibility and health problems within Southeast Asia [12]. During the peak period of smoke haze, the number of outpatient visits in Kuching, Sarawak, increased two to three times, while those visited at Kuala Lumpur General Hospital increased from 250 to 800 per day [13]. In conclusion, the smoke haze from the forest fires in Indonesia had a deleterious effect on the health of the population in surrounding countries, including Malaysia.

There are already efforts taken by the Malaysian government to reduce the air pollution problem in the country. To monitor the air quality, the Department of Environment (DoE) of Malaysia installed the air quality monitoring networks in various locations, including roadside, commercial, industrial and residential areas [14]. Then the Air Pollutant Index (API) is released hourly to communicate with the public on the changes in air quality. Also, the DoE set a limit on the maximum

sulfur and lead contents of petroleum fuels and coal to ensure the energy produced by the vehicles and industries will not increase the burden of air pollution. On an everyday basis, many use API forecasts to determine if they should wear a mask on a given day. However, they are unable to learn the air quality level of the next day (or next few days). This difficulty has a significant impact on high-risk people, especially those who are sensitive to the changes in air pollutant concentrations. Sensitive populations are not able to plan for proper outdoor activities, e.g., traveling to new places when the air quality may sudden change. This motivates us to propose a solution to utilize the hourly historical API to predict the API readings for the coming day.

### 1.3 Our contributions

Our contributions can be summarized as follows:

1. In this paper, we propose a method to forecast API readings in Malaysia. We utilize the hourly released API in Malaysia to predict the API readings in the coming day. At present, such forecasting is not available in Malaysia.
2. We identified the benefits of the proposed solution for air pollution avoidance behavior among the high-risk population living in different areas in Malaysia.
3. We implement 11 machine learning algorithms on the public API dataset (across seven states from Malaysia) to demonstrate the viability of the proposed solution.

### 1.4 Paper organization

The rest of this paper is organized as follows. We discuss the background and related work of this research in Sect. 2, followed by our proposed solution and assumptions used in Sect. 3. The details of our methodology are presented in Sect. 4. In Sect. 5, we present our experimental results and analysis. We then discuss future work and give the conclusions in Sect. 7.

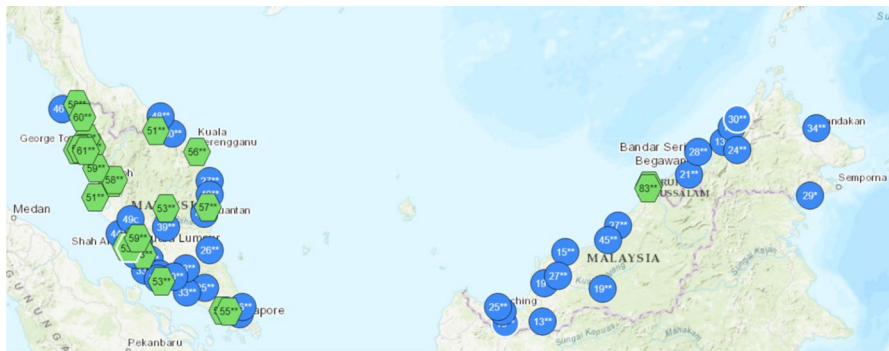
## 2 Background and related work

### 2.1 Air pollutant index

The Air Pollutant Index (API) is the ambient air quality measurement used in Malaysia, where it is a generalized and straightforward way to measure air quality [15]. The API value is calculated based on the average concentration of air pollutants, including SO<sub>2</sub>, NO<sub>2</sub>, CO, O<sub>3</sub>, PM<sub>2.5</sub> and PM<sub>10</sub>. The highest concentration pollutant is used to determine the API value, e.g., the concentration of particulate matter PM<sub>2.5</sub> is usually the highest among other pollutants and will be used to determine the API value. The DoE in Malaysia will release these API values hourly to communicate to the public how polluted the air in an area. The indication of air quality (based on API values) is summarized in Table 1.

**Table 1** Indication of air quality based on API values

API	Air pollution level	Description
Below 50	Good	Low pollution—no harmful effect on health
51–100	Moderate	Moderate pollution—does not pose any harmful effect on health
101–200	Unhealthy	Worsen the health condition of high-risk people with heart and lung complications
201–300	Very Unhealthy	Worsen the health condition and low tolerance of physical exercises to people with heart and lung complications. Affect public health
More than 300	Hazardous	Hazardous to high-risk people and public health



**Fig. 1** Hourly API readings in Malaysia (accessed on June 22, 2020, at 10:00 pm from <https://apims.doe.gov.my/>). From a total of 68 stations, 42 stations indicate a “good” level of air quality while 26 stations are at a “moderate” level

API in Malaysia is calculated based on 24 h of data retrieved from the air quality-monitoring network throughout the country and updated hourly. The data retrieval process requires a complete cycle of one hour before API readings can be obtained. As shown in Fig. 1, the hourly published API readings allow the public to be aware of the changes in air quality. However, there have been occasions in Malaysia where the API level published by DoE does not update despite worsening haze conditions [16].

## 2.2 API forecasting

Air quality forecasting is a prevention method that computes the pollutant concentrations before critical levels of air pollution are exceeded [17]. It is commonly performed with two approaches: empirical and deterministic. The empirical approaches employ statistical models to find associations among pollutants from the time series of past measurements. For example, in Athens, the  $PM_{10}$  concentrations at four different places are forecast hourly [18]. The main advantage of the empirical approaches is that they are easy to implement and only requires limited computing resources [19]. However, they require a broad observation to train the statistical

models and can be challenging for long-term air quality forecasts. The deterministic approach adopts chemical transport models (CTMs) to overcome the inabilities of statistical models. CTM is a complex system that requires large datasets and high computational power in order to produce long-term forecasts with comparable accuracy.

Although the API in Malaysia is already being used as an indicator to show the air quality and levels of the pollutant concentrations in a given period, it was not utilized for the coming day forecasting purposes. To the best of our knowledge, there is no API forecasting system available in Malaysia that can predict the coming day API readings. Hence, the high-risk people with heart and lung complications cannot prepare themselves when the air quality becomes bad, e.g., limiting outdoor activities as the API rises to unhealthy.

### 2.3 Related work

Air pollution-related research in Malaysia mainly focuses on the identification of possible sources of air pollutants in air quality around the study area based on the data (i.e., hourly API readings) obtained from the DoE [11, 20–22] particularly, with a focus on quantifying the level of  $PM_{10}$  and  $PM_{2.5}$  on the air pollutant concentrations collected from different hot spots set up by the DoE or independent party. A widely used analysis tool to estimate the concentrations of air pollutants is the Extreme Value Distribution (EVD) [23]. It is a statistical technique that can be used to model extreme events such as the maximum or minimum concentrations of the data. Another group of researchers concentrates on the study related to the association of air pollution and human health impacts in Malaysia [13, 14]. Recently, a study has been conducted to explore public perception of current air pollution in Malaysia, environmental awareness, and attitudes toward environmental protection [24], also some works utilizing the API readings for applications such as the traffic congestion modeling [25] and haze monitoring system [16, 26].

Traditional air pollution evaluation requires expensive equipment, significant infrastructure, and trained personnel to operate [27–30]. With the rapid growth in computational advancement, machine learning is offering new opportunities to overcome the limitations in traditional air pollution prediction. Mainly, machine learning techniques are capable of analyst vast amounts of data to establish a reasonable and accurate prediction model. Such a prediction model is useful for forecasting purposes because it can estimate the pollutant concentration at a future date [31]. The effectiveness of machine learning algorithms has been continuously improved and makes predictions more efficiently. With the improvement of machine learning algorithms, many air quality predictions in Malaysia were deemed accurate. The proliferation of good air quality predictions contributes to the possibility of performing future day API forecasting. Even though it has many advantages, there are also some disadvantages when applying machine learning for air pollution prediction. For instance, identifying the best one from many machine learning techniques for the air pollution problem can be challenging [32]. Furthermore, the feasibility of these sophisticated approaches for environmental problems is hard to be verified [33]. In

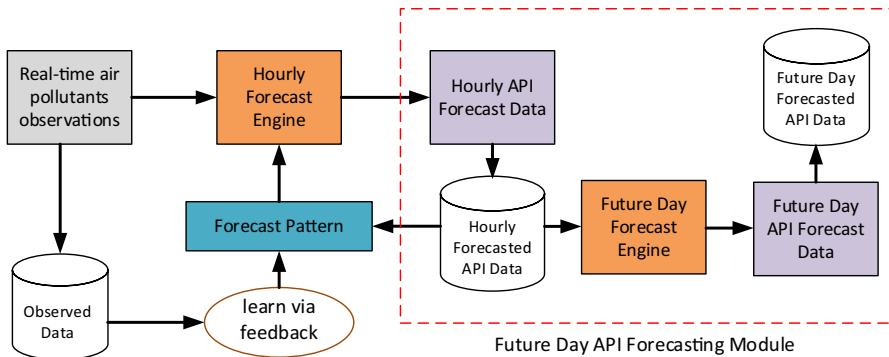


Fig. 2 Overview of the proposed solution

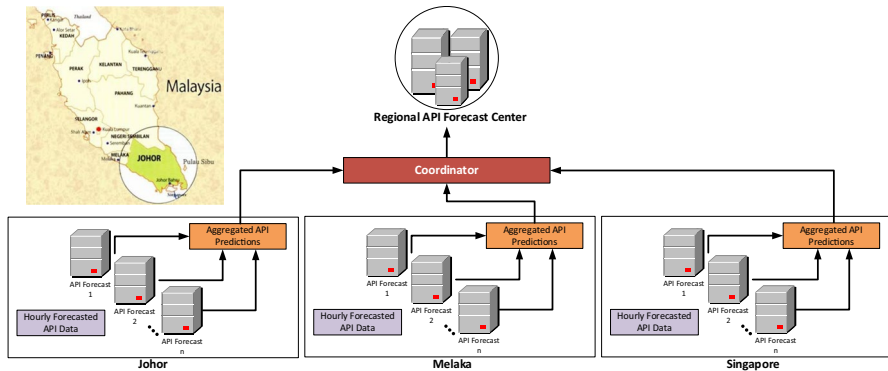
reality, the predicted results are seldom used for daily forecasting in many developing countries due to a lack of public awareness or interest in areas that suffer from heavy air pollution. As a consequence, the air quality in the pollution hot spots cannot be improved, or it could be getting worst.

There are some studies on API forecasting that have been conducted in Malaysia. In an early work [34], the Integrated Autoregressive Moving Average (ARIMA) and the Integrated Long Memory Model (ARFIMA) models were used to forecast API in Selangor, Malaysia. They utilize 70 monthly observations of API published from 1998 to 2003 in their analysis. Another study focuses on the prediction of monthly mean API using the time series model in Johor, Malaysia [35]. They employ the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Fuzzy Time Series (FTS) on data collected between 2000 and 2009. The work presented in [36] provides essential insights on how to improve air quality forecasting problems in Sarawak, Malaysia. Recently, a daily API prediction based on fuzzy time series Markov chain model has been proposed in [37]. In this paper, we focus on the API forecast application that utilizes the API released by the authority in Malaysia for future day prediction. Notably, our solution can utilize the historical API readings to predict both next-hour and next-day API readings.

### 3 Future day API forecasting system

To support avoidance behavior among the high-risk population, we design a future day API forecasting system that benefits from the existing air quality predictions in Malaysia. As shown in Fig. 2, our proposed solution consists of two forecast engines, namely an hourly forecast engine and a future day forecast engine. The former is responsible for hourly API prediction based on the observed real-time air pollutants. It could be maintained independently by the DoE or other solution providers for air pollution in Malaysia. The later uses the hourly API as the input to predict the coming day API.

To stay focus, we employ the following strategies in the proposed solution:



**Fig. 3** An example of a regional API forecasting system

1. Our direct strategy is to utilize the hourly API readings on a specific day to predict a future day API. This strategy is useful for a prediction where the pollutant concentration in an area is estimated to be similar, e.g., haze concentration during the southeast monsoon in Malaysia.
2. In the second strategy, we propose to combine different API prediction results for aggregated-level prediction, regardless of the techniques used. This approach is essential to overcome the common problem for machine learning algorithms, i.e., lack of pollutant concentration data and lack of useful pollutant concentration data.
3. Also, the proposed solution can be shared across borders, especially in the less economically developed regions across the world. For instance, the next-day API forecast in Singapore can be used as a reference in Johor Bharu, capital of the Malaysian state of Johor. Note that Johor Bahru and Singapore are geographically in the same region.

In Fig. 3, we show an example of a regional API forecasting system for three closed geographical cities: Johor, Melaka, and Singapore. The coordinator can be a local authority (e.g., DoE) or an independent solution provider.

We consider several assumptions in the design of our proposed solution, in particular, from technological development, environmental awareness to regional fragmentation. Since the first step to an accurate air quality forecast is an excellent weather forecast, we assume that the past weather forecast in Malaysia is a good indicator of the future. With the advancement of technologies, the API forecasting techniques will become more accurate. The accuracy and accessibility of air pollution forecasts will raise public awareness. Also, this enables sensitive populations to plan ahead of their daily activities and provide governments with information for public health alerts. The air pollution controls should be implemented stringently and continuously in each city within a region.



### 4 Methodology

As discussed earlier, machine learning is a booming field with numerous practical weather forecast applications. There are a number of notable machine learners worth for exploring in this study, i.e., tree-based model, ensemble learning model, neural network-based model, support vector machine (SVM), etc. All of them can perform very well in ordinary classification and prediction. However, applying them directly to process temporal- or sequential-based datasets may yield poor prediction accuracy because the underlying sequential-values and time-values are ignored from learning and thus limiting their direct usage in API forecast application. Therefore, a process termed as “data flattening” is proposed in this paper so that it can be embedded to the ordinary classifiers and extending their usages in tackling temporal- and sequential-based data (i.e., API data in this case).

Data flattening can be viewed as a process to transform the API record into a temporalized API record. In order to predict the next-day readings of API, we postulated that the historical API values will be useful. Thus, we adopted the method of data flattening to capture the data in a way that the historical API values will be merged and used to predict the next-day API value. The consecutive API records in a dataset can be merged based on the desired time window,  $w$ .

Let  $A$  represents a set of attributes consisting of some hourly API values, where  $A = \{a_1, a_2, a_3, \dots, a_M\}$ ,  $a$  indicates each hourly API value of that day and  $M$  indicates the total number of API values of that day. To merge several consecutive API records in a dataset with a time window of  $w$ , it can be done through the string theory (concatenation) of computer science. However, it is also important to record the time label for human interpretation later. For an instance, an attribute set ( $A$ ) occurred in time window ( $w$ ) of 1 will be labeled as  $A^1$  and can be represented as  $A^1 = \{a_1, a_2, a_3, \dots, a_M\}^{w=1}$ , and so forth.

Thereafter, if one would like to observe the data pattern occurred across a time window of  $j$ th value ( $w = j$ ), the new set of temporal attribute structures ( $\tilde{A}$ ) can be cascaded in the manner of  $\tilde{A}^{w=j} = A^1 A^2 A^3 \dots A^j$ , where  $\{(a_1, \dots, a_M)^{w=1} \in A^1; (a_1, \dots, a_M)^{w=2} \in A^2; (a_1, \dots, a_M)^{w=3} \in A^3; \dots, (a_1, \dots, a_M)^{w=j} \in A^j\}$ . The symbol  $\bullet||\bullet$  represents the concatenation operator among the attributes’ sets based on the desired window size ( $w$ ).

For instance, consider an observation record avails with 3 attributes ( $x$ ,  $y$ , and  $z$ ) as depicted in Fig. 4. To consider a set of its historical values before its next prediction, data flattening process can be performed on the window size of 2 ( $w = 2$ ). By doing so, the original record can be transformed into a new set of temporalized record by merging all of the attributes across  $w$  instances (2 in this case). The respective sequence (time) will be labeled for all attributes, i.e., a superscript of “1” indicates its past record (1 timestep before the current one), and the current one will be labeled with a superscript of “2”.

The consequences of record merging will also yield a new set of temporalized ( $T$ ) records from the original ( $N$ ) records, where  $T$  records =  $N$ records – ( $w - 1$ ). The algorithm of data flattening can be summarized as below:

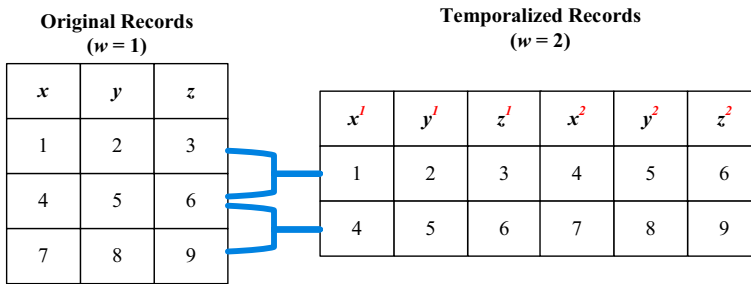


Fig. 4 An example of flattening process on a window size of 2 ( $w=2$ )

---

### Algorithm 1. Data Flattening Algorithm

---

**Input:**

$D$ , an API dataset with  $d$  records.  
Window size,  $w$ .

**Output:** temporalized records.

**Procedure:**

1. **for**  $i = 1$  to  $i = d-w+1$  **do**
  2.   **merge** row  $i$  to  $i+w-1$
  3. **endfor**
- 

It is interesting to know that how many time windows are needed to achieve the optimal prediction. However, it is unpractical to statistically compute the optimum window value, because it is very subjective to the manner of collecting datasets. Therefore, in this paper, this framework will be embedded to a number of machine learning techniques and run recursively on different window sizes until the error rates stop increasing.

## 5 Experimental results and analysis

### 5.1 Description of datasets

In this study, 11 datasets across 7 states from Malaysia are adopted, as in Table 2. All of them are publicly accessible from Malaysia's Open Data Portal, contributed by Department of Environment (DOE).<sup>1</sup>

<sup>1</sup> [https://www.data.gov.my/data/ms\\_MY/organization/department-of-environment-doe?page=2](https://www.data.gov.my/data/ms_MY/organization/department-of-environment-doe?page=2).

**Table 2** Summary of datasets description

	Location	State	Description
D1	Larkin	Johor	Hourly API values recorded throughout year 2017–2019
D2	Pasir Gudang		Hourly API values recorded throughout year 2017–2019
D3	Kota Tinggi		Hourly API values recorded throughout year 2017–2019
D4	Segamat		Hourly API values recorded throughout year 2017–2019
D5	Kulim	Kedah	Hourly API values recorded throughout year 2017–2019
D6	Batu Muda	Kuala Lumpur	Hourly API values recorded on year 2017 and 2019
D7	Cheras		Hourly API values recorded on year 2017 and 2019
D8	Kuala Terengganu	Terengganu	Hourly API values recorded throughout year 2017–2019
D9	Kuching	Sarawak	Hourly API values recorded throughout year 2017–2019
D10	Kota Kinabalu	Sabah	Hourly API values recorded throughout year 2017–2019
D11	Minden	Penang	Hourly API values recorded throughout year 2017–2019

## 5.2 Experimental settings and evaluation metrics

To avoid the biasness from the cherry-picked validation set, the experimental evaluation in this paper is based on a tenfold cross-validation (for Type III error). The data flattening methodology is integrated into 11 machine learners, including J48 decision tree, random forest, Elman network, decision stump, logistic model trees, random tree, reptime, hidden Markov model, hoeffding tree, support vector machines, and naïve Bayes to explore their adaptability and the strength of prediction. Empirical results are reported based on five standard performance evaluation metrics, which are: (i) prediction accuracy, (ii) true positive rate (TPR), (iii) false positive rate (FPR), (iv) precision, and (v) recall.

## 6 Result discussions

In order to utilize the historical API readings to predict the next-hour or even the next-day API readings, all API values from the D1–D11 are sorted by day and hourly basis in a chronological order. This setting is used across for all experiments. To observe how many historical values are needed to predict the next-day API, the experiments are tested on 10 window sizes (from  $w = 1$  to 10), depicting 1–10 days will be observed and monitored before their values are used to predict the current event or next event. The best value of  $w$  is reported when it hits the optimum classification accuracy. Refer to Table 3, almost all machine learning techniques except hidden Markov model are able to achieve the prediction accuracy of more than 97%. This is very encouraging because it depicts that the prediction of next-hour or next-day API values is possible with the emerging trend of machine learning which may bring benefits to certain group of users. The best result is reported by random forest with the window size,  $w = 1$ , indicating that 1-day historical API values should be used to predict the next-day API level.

**Table 3** Experimental exploration of 11 machine learning techniques on Dataset D1: Larkin across 10 observing time-step, from  $w = 1$  to  $w = 10$ 

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines and (11) naïve Bayes on dataset D1: Larkin												
	1	2	3	4	5	6	7	8	9	10	11	
Best window size, $w$	8	1	1	1	5	2	9	1	1	2	1	
Prediction accuracy (%)	98.7	99.1*	98.5	97.5	98.9	98.0	98.6	43.1	97.6	98.4	97.5	
TPR (%)	98.7	99.1	98.5	97.5	98.9	98.0	98.6	43.1	97.6	98.4	97.5	
FPR (%)		1.3	1.1	1.4	2.3	1.2	2.3	1.5	43.1	2.9	1.7	3.1
Precision (%)	98.7	99.1	98.5	97.5	98.9	98.0	98.6	18.6	97.7	98.4	97.6	
Recall (%)	98.7	99.1	98.5	97.5	98.9	98.0	98.6	43.1	97.6	98.4	97.5	

\*Indicates the best classification result

**Table 4** Experimental exploration of 11 machine learning techniques on Dataset D2: Pasir Gudang across 10 observing time-step, from  $w = 1$  to  $w = 10$ 

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D2: Pasir Gudang												
	1	2	3	4	5	6	7	8	9	10	11	
Best window size, $w$	5	2	1	1	2	1	3	1	1	1	1	
Prediction accuracy (%)	99.1	99.4*	99.3	98.2	99.3	98.3	98.4	51.1	97.4	99.3	97.4	
TPR (%)	99.1	99.4	99.3	98.2	99.3	98.3	98.4	51.1	97.4	99.3	97.4	
FPR (%)		0.8	0.5	0.6	1.7	0.6	1.6	1.5	51.1	2.4	0.6	2.4
Precision (%)	99.1	99.4	99.3	98.2	99.3	98.3	98.4	26.1	97.4	99.3	97.4	
Recall (%)	99.1	99.4	99.3	98.2	99.3	98.3	98.4	51.1	97.4	99.3	97.4	

\*Indicates the best classification result

The same trend as in D1: Larkin can be observed in D2: Pasir Gudang as well. Refer to Table 4, random forest is outperformed the rest, with the observed window size,  $w$ , of 2 indicating the fact that to predict today's API, we need the historical API values of yesterday and the day before yesterday. It is interesting to observe that different machine learning techniques might need to use different window size to boost their predictive performance. This substantiate the usefulness of historical API values in the prediction task.

In the monitoring area of Kota Tinggi and Segamat, the best-performed technique is Elman network as shown in Tables 5 and 6, respectively. It is important to note that the Elman network is a simple recurrent network. It has the ability to observe the sequential values of API in the past, and thus, it does not need more

**Table 5** Experimental exploration of 11 machine learning techniques on Dataset D3: Kota Tinggi across 10 observing time-step, from  $w = 1$  to  $w = 10$

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D3: Kota Tinggi											
	1	2	3	4	5	6	7	8	9	10	11
Best window size,	1	3	1	6	9	3	2	1	1	1	1
Prediction accuracy (%)	98.7	99.3	99.7*	95.9	98.7	98.2	98.1	72.9	96.3	99.0	96.3
TPR (%)	98.7	99.3	99.7	95.9	98.7	98.2	98.1	72.9	96.3	99.0	96.3
FPR (%)	2.6	0.7	0.6	6.4	1.8	3.6	3.1	72.9	1.9	1.1	1.9
Precision (%)	98.7	99.3	99.7	95.9	98.7	98.2	98.1	53.2	96.7	99.0	96.7
Recall (%)	98.7	99.3	99.7	95.9	98.7	98.2	98.1	72.9	96.3	99.0	96.3

\*Indicates the best classification result

**Table 6** Experimental exploration of 11 machine learning techniques on Dataset D4: segamat across 10 observing time-step, from  $w = 1$  to  $w = 10$

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D4: segamat											
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	2	2	1	1	1	3	6	1	1	1	1
Prediction accuracy (%)	98.1	98.9	99.2*	97.7	98.9	97.9	98.3	52.9	96.0	98.7	96.0
TPR (%)	98.1	98.9	99.2	97.7	98.9	97.9	98.3	52.9	96.0	98.7	96.0
FPR (%)	2.0	0.9	0.7	2.2	1.0	2.0	1.6	52.9	3.5	1.2	3.5
Precision (%)	98.1	98.9	99.2	97.7	98.9	97.9	98.3	28.0	96.0	98.7	96.2
Recall (%)	98.1	98.9	99.2	97.7	98.9	97.9	98.3	52.9	96.0	98.7	96.0

\*Indicates the best classification result

than 1-day historical values in order to improve its predictive performance. This observation is constant by looking into Table 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13.

Refer to Table 7, the best-performed technique is random forest with the prediction accuracy of 99.3% by utilizing 1-day historical value. Hidden Markov model is slightly poorer compared to others, most probably hampered by the unbalanced dataset. This phenomenon can be observed throughout the 11 datasets used in the experiments. In this 11 datasets, most of the API levels are skewed toward “Good” to “Moderate” in the observation period of year 2017 to 2019.

In the monitoring area of Batu Muda, almost all techniques except logistic model trees can reach their optimum prediction by using only 1-day historical value as shown in Table 8. This is logic in the sense that API values does not

**Table 7** Experimental exploration of 11 machine learning techniques on Dataset D5: Kulim across 10 observing time-step, from  $w = 1$  to  $w = 10$ 

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D5: Kulim											
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	1	1	1	4	8	4	5	1	1	1	1
Prediction accuracy (%)	98.6	99.3*	98.6	96.9	97.9	98.4	98.1	53.1	96.1	97.7	96.0
TPR (%)	98.6	99.3	98.6	96.9	97.9	98.4	98.1	53.1	96.1	97.7	96.0
FPR (%)	1.4	0.6	1.3	3.1	2.1	1.7	1.9	53.1	3.4	2.1	3.5
Precision (%)	98.6	99.3	98.6	96.9	97.9	98.4	98.1	28.2	96.1	97.7	96.2
Recall (%)	98.6	99.3	98.6	96.9	97.9	98.4	98.1	53.1	96.1	97.7	96.0

\*Indicates the best classification result

**Table 8** Experimental exploration of 11 machine learning techniques on dataset D6: Batu Muda across 10 observing time-step, from  $w = 1$  to  $w = 10$ 

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines and (11) naïve Bayes on dataset D6: Batu Muda											
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	1	1	1	1	2	1	1	1	1	1	1
Prediction accuracy (%)	99.8*	99.6	99.4	99.4	99.6	99.4	99.1	66.5	98.3	99.8*	98.2
TPR (%)	99.8	99.6	99.4	99.4	99.6	99.4	99.1	66.5	98.3	99.8	98.2
FPR (%)	0.09	0.4	0.5	1.0	0.1	0.8	1.2	66.5	0.8	0.09	1.1
Precision (%)	99.8	99.6	99.4	99.4	99.6	99.4	99.1	44.2	98.4	99.8	98.2
Recall (%)	99.8	99.6	99.4	99.4	99.6	99.4	99.1	66.5	98.3	99.82	98.2

\*Indicates the best classification result

change rapidly or immediately within few hours in natural, unless hampering by the third party factors, i.e., fire, leak of gases, etc.

The same trend of prediction can be seen in Cheras area as well, where almost all machine learning techniques except Hidden Markov Model are able to achieve more than 96% prediction accuracy. Most of them are leveraging window size of 1–3, except the decision stump which needs 10 window sizes to achieve its optimum prediction as shown in Table 9. This could be due to the data distribution in D7 which is unable to split well by using a single attribute as practiced by decision stump.

The stability of random forest is substantiated again in the monitoring area of Kuala Terengganu and Kuching, as shown in Tables 10 and 11, respectively. In both of these datasets, the window size used by all techniques is less than 4, and mostly retain with

**Table 9** Experimental exploration of 11 machine learning techniques on dataset D7: Cheras across 10 observing time-step, from  $w = 1$  to  $w = 10$

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D7: Cheras											
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	2	2	1	10	3	1	3	1	1	1	1
Prediction accuracy (%)	97.0	97.4	98.9*	97.1	98.5	96.3	96.8	62.8	96.3	97.7	96.3
TPR (%)	97.0	97.4	98.9	97.1	98.5	96.3	96.8	62.8	96.3	97.7	96.3
FPR (%)	4.6	3.9	1.1	4.0	1.9	4.8	4.9	62.8	2.7	2.5	2.7
Precision (%)	97.1	97.4	98.9	97.2	98.5	96.3	96.9	39.5	96.4	97.7	96.4
Recall (%)	97.0	97.4	98.9	97.1	98.5	96.3	96.8	62.8	96.3	97.7	96.3

\*Indicates the best classification result

**Table 10** Experimental exploration of 11 machine learning techniques on dataset D8: Kuala Terengganu across 10 observing time-step, from  $w = 1$  to  $w = 10$

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D8: Kuala Terengganu											
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	1	1	1	1	1	2	1	1	1	1	1
Prediction accuracy (%)	98.7	99.4*	98.5	97.5	98.7	98.5	98.1	53.4	96.6	98.3	96.5
TPR (%)	98.7	99.4	98.5	97.5	98.7	98.5	98.1	53.4	96.6	98.3	96.5
FPR (%)	1.2	0.5	1.5	2.4	1.2	1.3	2.0	53.4	2.9	1.6	3.0
Precision (%)	98.7	99.4	98.5	97.5	98.7	98.5	98.1	28.6	96.8	98.3	96.7
Recall (%)	98.7	99.4	98.5	97.5	98.7	98.5	98.1	53.4	96.6	98.3	96.5

\*Indicates the best classification result

$w = 1$  to 2. This pattern holds consistent throughout these 11 datasets, demonstrating that we need at least 1 or 2-day past historical API values to predict the next-hour and next-day API.

The logistic model trees outperformed random forest in both of these areas: Kota Kinabalu and Minden, as shown in Tables 12 and 13, respectively. Likewise, most of the used window size is less than 4. One exceptional case is the reptime usage in D11: Minden, where it uses  $w = 9$  most probably due to its sensitivity toward overfitting.

**Table 11** Experimental exploration of 11 machine learning techniques on dataset D9: Kuching across 10 observing time-step, from  $w = 1$  to  $w = 10$ 

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D9: Kuching											
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	2	2	1	1	2	1	4	1	1	1	1
Prediction accuracy (%)	98.4	98.8*	97.6	96.8	97.8	97.6	97.5	72.7	90.3	97.3	90.3
TPR (%)	98.4	98.8	97.6	96.8	97.8	97.6	97.5	72.7	90.3	97.3	90.3
FPR (%)	2.6	2.2	2.2	3.2	2.8	3.1	3.5	72.7	4.1	2.8	3.6
Precision (%)	98.2	98.7	97.5	96.7	97.8	97.7	97.4	52.9	92.5	97.3	92.7
Recall (%)	98.4	98.8	97.6	96.8	97.8	97.6	97.5	72.7	90.3	97.3	90.3

\*Indicates the best classification result

**Table 12** Experimental exploration of 11 machine learning techniques on dataset D10: Kota Kinabalu across 10 observing time-step, from  $w = 1$  to  $w = 10$ 

Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D10: Kota Kinabalu											
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	1	1	1	2	2	1	1	1	2	1	1
Prediction accuracy (%)	97.5	98.2	95.3	95.5	98.8*	97.8	97.3	73.5	93.6	97.1	92.9
TPR (%)	97.5	98.2	95.3	95.5	98.8	97.8	97.3	73.5	93.6	97.1	92.9
FPR (%)	5.0	3.3	3.0	4.6	1.2	4.1	5.3	73.5	3.0	2.1	3.3
Precision (%)	97.4	98.1	95.6	95.6	98.7	97.7	97.2	54.1	94.6	97.1	94.0
Recall (%)	97.5	98.2	95.3	95.5	98.8	97.8	97.3	73.5	93.6	97.1	92.9

\*Indicates the best classification result

## 7 Conclusion

In this paper, we have proposed a solution for the next day API forecasting in Malaysia. The proposed solution ensures the sensitive population can plan their activities to avoid possible sudden changes in the air pollutant concentrations. To the best of our knowledge, this is the first future day forecasting system based on the hourly API readings in Malaysia. Extensive experimental results on public datasets verified the viability of our solution for the next day prediction. As future work, we plan to study the correlation between air pollution forecasts and other environmental forecasts. Such a study will allow the local health authorities to make informed decisions on mitigation measures to reduce public exposure risk in heavily polluted hot spots.



**Table 13** Experimental exploration of 11 machine learning techniques on dataset D11: Minden across 10 observing time-step, from  $w = 1$  to  $w = 10$ 

	Performance comparisons of (1) J48 decision tree, (2) random forest, (3) Elman network, (4) decision stump, (5) logistic model trees, (6) random tree, (7) reptime, (8) hidden Markov model, (9) hoeffding tree, (10) support vector machines, and (11) naïve Bayes on dataset D11: Minden										
	1	2	3	4	5	6	7	8	9	10	11
Best window size, $w$	4	1	1	2	1	2	9	1	1	1	1
Prediction accuracy (%)	97.7	98.9	98.5	96.5	99.0*	97.4	97.1	28.5	98.9	98.1	98.7
TPR (%)	97.7	98.9	98.5	96.5	99.0	97.4	97.1	28.5	98.9	98.1	98.7
FPR (%)	2.9	2.0	3.2	4.3	2.0	3.6	4.9	28.5	98.9	4.7	98.7
Precision (%)	97.8	98.9	98.5	96.5	99.0	97.4	97.1	8.1	2.4	98.1	2.8
Recall (%)	97.7	98.9	98.5	96.5	99.0	97.4	97.1	28.5	98.9	98.1	98.7

\*Indicates the best classification result

## References

- World Health Organization (2016) Ambient air pollution: a global assessment of exposure and burden of disease. WHO. <http://apps.who.int/iris/bitstream/10665/250141/1/9789241511353-eng.pdf>. Accessed 1 June 2020
- Ma D, Zhao T (2014) Talking about the current situation of air pollution in China and its governance recommendations. *J Hebei Inst Arch Eng* 32(2):53–54
- Organization WH (2016) Air pollution levels rising in many of the world's poorest cities. Switzerland, Geneva
- Council M-AR (2019) 2019 Annual air quality awareness survey. ETC Institute, Olathe
- Szyszkowicz M, Kousha T (2014) Emergency department visits for asthma in relation to the air quality health index: a case-crossover study in Windsor, Canada. *Can J Public Health* 105(5):e336–e341
- Dalsøren SB, Jonson JE (2016) Socio-economic impacts—air quality, Chap 16. In: Quante M, Colijn F (Hrsg) (ed) North Sea region climate change assessment. Springer, Berlin, Heidelberg, pp 431–446
- Beverland IJ, Cohen GR, Heal MR, Carder M, Yap C, Robertson C, Hart CL, Agius RM (2012) A comparison of short-term and long-term air pollution exposure associations with mortality in two cohorts in Scotland. *Environ Health Perspect* 120(9):1280–1285
- Organization WH (2018) Air pollution and child health: prescribing clean air: summary. World Health Organization, Geneva
- Conforti A, Mascia M, Cioffi G, De Angelis C, Coppola G, De Rosa P, Pivonello R, Alviggi C, De Placido G (2018) Air pollution and female fertility: a systematic review of literature. *Reprod Biol Endocrinol* 16(1):117
- Sastry N (2002) Forest fires, air pollution, and mortality in Southeast Asia. *Demography* 39(1):1–23
- Dominick D, Juahir H, Latif MT, Zain SM, Aris AZ (2012) Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmos Environ* 60:172–181
- Ogawa H (1998) The Haze Episode of 1997 in Countries of South-East Asia, pp 416–428
- Afroz R, Hassan MN, Ibrahim NA (2003) Review of air pollution and health impacts in Malaysia. *Environ Res* 92(2):71–77
- Murad MW, Pereira JJ (2011) Malaysia: environmental health issues. In: Nriagu JO (ed) Encyclopedia of environmental health. Elsevier, Burlington, pp 577–594. <http://www.sciencedirect.com/science/article/pii/B9780444522726005390>. Accessed 5 May 2020
- Malaysia DoE (2000) A guide to air pollutant index (API) in Malaysia. [https://issuu.com/universiti\\_teknologimalaysia/docs/a\\_guide\\_to\\_pollutant\\_index\\_\\_api\\_\\_in](https://issuu.com/universiti_teknologimalaysia/docs/a_guide_to_pollutant_index__api__in)

16. Ya'acob N, Zainuddin A, Abidin IFZ, Yusof AL, Idris A (2016) Web-based real time haze monitoring system. In: 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE). IEEE, pp 51–55
17. Carnevale C, De Angelis E, Finzi G, Turrini E, Volta M (2019) An integrated forecasting system for air quality control. In: 2019 18th European Control Conference (ECC). IEEE, pp 830–835
18. Carnevale C, Finzi G, Pisoni E, Singh V, Volta M (2011) An integrated air quality forecast system for a metropolitan area. *J Environ Monit* 13(12):3437–3447
19. Carnevale C, Finzi G, Pederzoli A, Turrini E, Volta M (2016) Lazy Learning based surrogate models for air quality planning. *Environ Model Softw* 83:47–57
20. Azmi SZ, Latif MT, Ismail AS, Juneng L, Jemain AA (2010) Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Qual Atmos Health* 3(1):53–64
21. Azid A, Juahir H, Toriman ME, Endut A, Kamarudin MKA, Rahman MNA, Hasnam CNC, Saudi ASM, Yunus K (2015) Source apportionment of air pollution: a case study in Malaysia. *J Teknol* 72(1):83–88
22. Suparta W, Alhasa KM, Singh MSJ, Latif MT (2015) The development of PWV index for air pollution concentration detection in Banting, Malaysia. In: 2015 International Conference on Space Science and Communication (IconSpace). IEEE, pp 498–502
23. Coles S, Bawa J, Trenner L, Dorazio P (2001) An introduction to statistical modeling of extreme values, 208th edn. Springer, London
24. Chin YSJ, De Pretto L, Thuppil V, Ashfold MJ (2019) Public awareness and support for environmental protection—a focus on air pollution in peninsular Malaysia. *PLoS ONE* 14(3):e0212206
25. Anjum SS, Noor RM, Aghamohammadi N, Ahmedy I, Kiah LM, Hussin N, Anisi MH, Qureshi MA (2019) Modeling traffic congestion based on air quality for greener environment: an empirical study. *IEEE Access* 7:57100–57119
26. Ya'acob N, Azize A, Adnan NM, Yusof AL, Sarnin SS (2016) Haze monitoring based on air pollution index (API) and geographic information system (GIS). In: 2016 IEEE Conference on Systems, Process and Control (ICSPC). IEEE, pp 7–11
27. EPA U (1997) Ambient air monitoring reference and equivalent methods. Federal Register 40 CFR Parts, p 50. <https://www.federalregister.gov/documents/2020/05/07/2020-09704/ambient-air-monitoring-reference-and-equivalent-methods-designation-of-one-new-equivalent-method>
28. Hesketh HD (1996) Air Pollution Control: Traditional Hazardous Pollutants. CRC Press, Boston
29. Martín MJ, Parada M, Doallo R (2004) High performance air pollution simulation using OpenMP. *J Supercomput* 28(3):311–321
30. Yang C-T, Chan Y-W, Liu J-C, Lou B-S (2020) An implementation of cloud-based platform with r packages for spatiotemporal analysis of air pollution. *J Supercomput* 76(3):1416–1437
31. Bai L, Wang J, Ma X, Lu H (2018) Air pollution forecasts: an overview. *Int J Environ Res Public Health* 15(4):780
32. Sinnott RO, Guan Z (2018) Prediction of air pollution through machine learning approaches on the cloud. In: 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT). IEEE, pp 51–60
33. Karimian H, Li Q, Wu C, Qi Y, Mo Y, Chen G, Zhang X, Sachdeva S (2019) Evaluation of different machine learning approaches to forecasting PM<sub>2.5</sub> mass concentrations. *Aerosol Air Qual Res* 19(6):1400–1410
34. Siew LY, Chin LY, Wee PMJ (2008) ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor. *Malays J Anal Sci* 12(1):257–263
35. ABD Rahman NH, Lee MH, Suhartono LM (2016) Evaluation performance of time series approach for forecasting air pollution index in Johor, Malaysia. *Sains Malays* 45(11):1625–1633
36. Zakaria NN, Othman M, Sokkalingam R, Daud H, Abdullah L, Abdul Kadir E (2019) Markov chain model development for forecasting air pollution index of Miri, Sarawak. *Sustainability* 11(19):5190
37. Alyousifi Y, Othman M, Sokkalingam R, Faye I, Silva PC (2020) Predicting daily air pollution index based on fuzzy time series Markov chain model. *Symmetry* 12(2):293