ASM Sc. J., 16, 2021

https://doi.org/10.32802/asmscj.2021.756

Prediction of Carbon Monoxide (CO) Atmospheric Pollution Concentrations with Machine Learning and Time Series Analysis in Langkawi, Malaysia

Wah Chyang Choy[1], Azleena Mohd Kassim[1*] and Ahmad Zia Ul-Saufie[2]

[1]*School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia*

[2]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia*

Carbon monoxide (CO) is a non-irritant toxic and odourless gas produced from the incomplete combustion of fossil fuels. Long-term exposures to lower levels of carbon monoxide have wide implications for human health. Thus, an early warning system for CO atmospheric concentration with an accurate and reliable forecasting method is crucial. Studies for predicting CO atmospheric concentration are still limited in Malaysia especially using data science approaches. This study aims to develop and predict future CO concentration for the next few hours by using the statistical time series approach and machine learning approach. The data used for the project is the air quality data of the monitoring station in Langkawi, Malaysia. The data mining tool used for this project is RapidMiner Studio. Based on the results, it showed that Time Series analysis with deep learning gave a reasonably good CO concentration prediction for the next 3 hours with a relative error of approximate 10%. The model developed in this project can be used by authorities as public health's protection measure to provide an early alarm for alerting the Malaysian populations on the air pollution issue.

Keywords: machine learning; Time series analysis; carbon monoxide; prediction

I. INTRODUCTION

Clean air is a basic requirement to sustain life. Human beings need an essential continuous supply of air at 10 m3–20 m3 per day. "All people should have free access to clean air as one of the fundamental human rights", as said by the World Health Organisation (WHO) Director-General during the closing of a three-day global conference on air pollution and health (Clean air is a human right: WHO, 2018). Clean air in nature may have more components difference than pure air from a scientific perspective, and thus it is quite complicated to precisely define clean air. Figure 1 shows a list of gaseous components of natural pure air (Baumbach, 1996).

Air pollution is caused by pollutants in the atmosphere in a certain concentration and period, which can cause an unwanted effect on humans, plants, animal life or property. Human activities or even certain natural phenomena can become the root cause of the unfavourable concentrations of air pollutants. Some examples of traditional pollutants are carbon monoxide, hydrogen sulphide, nitrogen oxides, sulphur dioxide and haze.

		Volume content in % related to dry air
Oxygen	(O_2)	20.93
Nitrogen	(N_2)	78.10
Argon	(Ar)	0.9325
Carbon dioxide	(CO_2)	0.03-0.04
Hydrogen	(H_2)	0.01
Neon	(Ne)	0.0018
Helium	(He)	0.0005
Krypton	(Kr)	0.0001
Xenon	(Xe)	0.000009

Figure 1. Natural Composition of Air (Baumbach, 1996)

*Corresponding author's e-mail: azleena.mk@usm.my

WHO has provided guidelines for public health with regard to risks that can occur from several chemicals that are commonly present in the air (Penney *et. al.,* 2010; Ambient (outdoor) air pollution, 2018). Carbon monoxide (CO) is one of the pollutants that need to be given attention. Carbon monoxide (CO) is a non-irritant toxic gas that is odourless. It is produced from fossil fuels, such as diesel and kerosene in the combustion of Otto or Diesel engines, which is mainly from automobiles in the street traffic.

Baumbach (1996) claimed that it is quite impossible to eliminate CO in exhaust gas emissions because the complete combustion process of carbon to CO2 (carbon dioxide) requires an ignition temperature of at least 717°C for a certain period. Most of the time this condition cannot be achieved because the automobile engines will not be operated with a constant number of revolutions and constant load. Therefore, automobiles are the main contributors of carbon monoxide in the atmosphere.

Berg *et al.* (2002) explained that Oxygen (O_2) appears as oxyhaemoglobin (HbO_2) in blood transportation and is attached weakly to Fe^{2+} in haemoglobin. Blood oxygen-carrying capacity is reduced by CO as it forms a stable carboxyhaemoglobin (COHb) by merging with haemoglobin.

The affinity of haemoglobin for CO is around 200–250 times stronger than for oxygen (Higgins, 2005). CO can poison the haemoglobin oxygen transport system as it cannot regenerate the COHb, and thus making haemoglobin unavailable for oxygen transport (Vesilind *et al.,* 2013). COHb level is determined by variables, such as CO in the inhaled air and the exposure period (Penney *et al.,* 2010). In Malaysia, the number of studies to predict CO atmospheric concentration is still limited, especially by using data science approaches.

This study aims to develop and predict CO concentration in the future. The study objectives are: (1) to determine the characteristics of CO and its relationship with other meteorological parameters by using descriptive statistics and data visualisation, (2) to develop a model for CO prediction concentration by using time series approach (ARIMA), machine learning techniques and deep learning techniques, and (3) to determine the best model for predicting CO concentration in Langkawi Island, Malaysia.

The model developed in this project will be a localised model that suits Malaysia's topography and could be implemented for public health protection in providing an early alarm to alert the Malaysian population on the air pollution issue.

In the next section of this paper, some related works are presented. This is followed by the methods used, wherein the analytical techniques and data selection are presented in the third section. The fourth section discusses further about data preparation, which is followed by evaluation in the fifth section. Results are further discussed with the help of some visualisation in the sixth section. In the seventh section, subsequent discussion is presented, focusing on the two different approaches used for prediction and the prediction with the best model built. In the eighth section of the paper, the conclusion is presented.

II. RELATED WORKS

UK AIR (2021) presents a frequently updated air pollution forecast and the latest measured air quality, but it is limited to the United Kingdom only. READY (2019) is a system which displays meteorological data products by using a dispersion model on their Air Resources Laboratory's web server. The prediction of air pollutants can be considered as a key component in environmental monitoring, for instance, to help identify possible trends and as a guideline for environmental policies. This can be seen in a report presented by Tonellato (2001) on the Italian law which required short-term forecasts by public authorities at some locations of monitoring stations.

Hamid *et al.* (2017) carried out a study to predict the CO concentration at two locations in Malaysia, which are Kuala Terengganu (in Terengganu) and Bachang (in Melaka). Statistical time series models were used. Findings showed that in Bachang, the most suitable model was ARIMA (1,0,1) whereas, in Kuala Terengganu, the ARIMA (1,0,2) was found to be the most appropriate model.

A study carried out by Shaadan (2019) showed that several industrial sites in peninsular Malaysia had different temporal behaviour of CO levels. Each industrial site had a different best-appropriate model. Therefore, a specified model is needed to be developed for a specific location, such as Langkawi Island. This model can be used by authorities

as a public health protection measure to provide early alerts to Malaysia's population concerning the air pollution issue. For instance, the model can be used to alert the tourists on the CO forecast in Langkawi Island and help them in their planning, i.e., suitability alert for outdoor activities.

In 2005, Venkatasubramanian (2019) reported on three significant ideas that emerged in Data Science, which were reinforcement learning, deep or convolutional neural networks (CNNs), and statistical machine learning (ML). Exploring further in Data Science and air quality prediction, techniques like the support vector machine (SVM), mixture model and artificial neural network (ANN), had grown to be favourable (Heo & Kim, 2004; Lu & Wang, 2008).

This project used a systematic project life cycle approach known as Cross-Industry Standard Process for Data Mining (CRISP-DM). It is broadly used in industry (James, 2019) (Figure 2). CRISP-DM comprises six phases, which are initialised by the phase of (1) business understanding (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation and, (6) deployment. It iterated as a cyclical process. In each phase of the CRISP-DM process, there are a few second-level generic activities with specialised operations. CRSIP-DM methodology phase is not a one-direction flow. Some phases are two-way and iterative.

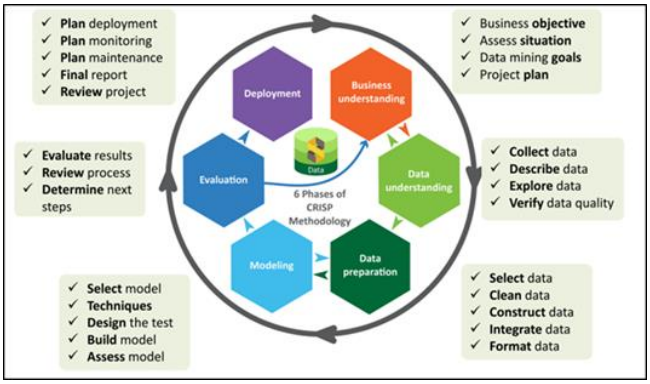

Figure 2. The CRISP-DM process Model (James, 2019)

III. DATASET DESCRIPTIONS AND ANALYTICAL TECHNIQUES

The data used for the project is the air quality data from a monitoring station in Langkawi Island, which is an island in Kedah, Malaysia. Langkawi Island was selected as the case study area because it is a tourist area with very few industrial activities. Therefore, very few air pollution studies were carried out. The data were acquired from the Department of Environment (DOE), Malaysia. The data were presented in excel worksheet format. The dataset contained data collection for the year 2004 taken at every hour interval.

The dataset comprised 8 attributes: 1) date, 2) time, 3) carbon monoxide concentration (CO Conc), 4) air pollution index (API), 5) wind speed (WS), 6) wind direction (WD), 7) relative humidity (Humidity) and 8) temperature (Temp). All data were in a numerical format and chronological order. There were 8,784 records in 2004, which were taken from 1 January at 1:00 am to 31 December at 24:00 am. The 2004 data was selected to train the model because it contained fewer missing values as compared to the data of other years, which had a large amount of missing values and outliers (for instance temperature with negative value).

The data selection approach was also aligned with the recommendation by Pyle (2003), who stated fundamental principles for the right selection of data which select relevant and redundant attributes and ensure that the records cover the complete range of between-attributes and within attributes behaviours. Figure 3 shows the partial dataset for 24 hours on 1 January 2004.

Besides dataset selection, the selection of a suitable learning algorithm is crucial, but this cannot be achieved without understanding the data characteristic and advantages and limitations of each learning algorithm. For this project, the dataset contained the labelled output variable for the research project goals - CO concentration. In this project, the output was focused on numerical attributes with the goal to discover changes in each of the predictor meteorological variables that will trigger a change in the output variable (CO concentration), and thus a prediction that will estimate a numerical value.

One of the key characteristics of the dataset is that the CO concentration is a set of quantitative observations arranged in chronological order in the same regularity of the observation frequency (every hour). Therefore, the analytical techniques used in the project were based on the time series approach.

RapidMiner was used for data visualisation and data analysis. RapidMiner was selected because of its ease of use and did not require any coding. Therefore, the work can be

focused on the analytical methodology rather than on programming.

Alongside with RapidMiner, MATLAB was used for plotting a chart for data visualisation. Time series forecasting in RapidMiner was based on the Windowing concept. Windowing is terminology for RapidMiner time series data which is converted into a generic cross-sectional dataset, whereby the target variable and the next time steps are predicted by applying ARIMA, Machine Learning or Deep Learning algorithms.

Row No.	Date	Time	WS	WD	API	COConc	Temp	Humidity
1	20040101	100	7.900	115	41	0.341	26.400	69.500
2	20040101	200	8.300	112	39	0.306	26.100	69.600
3	20040101	300	8.700	104	38	0.264	25.600	71.300
4	20040101	400	8.300	113	38	0.233	25.300	71.700
5	20040101	500	7.800	109	37	0.199	25.100	73.100
6	20040101	600	6.200	98	38	0.169	24.400	74.800
7	20040101	700	5.700	100	38	0.153	24.200	75.100
8	20040101	800	7.800	102	39	0.169	24.600	74.200
9	20040101	900	7.100	96	38	0.210	25.800	72.200
10	20040101	1000	7.800	128	37	0.236	28.100	67.900
11	20040101	1100	8	127	37	0.263	31.200	63
12	20040101	1200	8.400	142	37	0.290	33.400	59.100
13	20040101	1300	9.300	147	37	0.317	34.500	56
14	20040101	1400	10.100	138	37	0.353	35.700	52.700
15	20040101	1500	10.700	113	38	0.352	35.900	51
16	20040101	1600	9.500	123	38	0.349	36.100	49
17	20040101	1700	7.300	107	39	0.339	36.700	48
18	20040101	1800	7.100	116	38	0.350	35	50.200
19	20040101	1900	7.400	98	38	0.376	32.500	55.800
20	20040101	2000	7.800	101	39	0.391	29.800	63.800
21	20040101	2100	9.300	115	40	0.415	29	66.600
22	20040101	2200	8.200	114	41	0.431	28.100	69
23	20040101	2300	8.800	104	42	0.438	27.100	70.800
24	20040101	2400	7.600	93	43	0.432	26.500	71.400

Figure 3. Partial of the Dataset for the year 2004

IV. DATA PREPARATION

A. Treating Missing Value

The dataset contained records with attributes that have no measured values, which is often termed as "missing value" in data mining terminology. The possible reasons for missing value can be from errors during the gathering process, measuring sensor malfunction, and data corruption in the way data is processed. When data are missing in a variable of a particular case, it is very important to fill this attribute with some intuitive data for those algorithms that require one, especially for time series forecasting. Although the best way to eliminate missing values is to fill them through own further research, it is most time-consuming and it is not possible for the historical data in this context.

Therefore, a reasonable estimate of a suitable data value for missing data is required rather than leaving it blank. The methodology of replacing missing values depends on criteria without adding or removing any information from the data set and depends on the assumption about the dataset pattern.

One common replacement method is by choosing the variable's mean value as a replacement. Kolehmainen *et al.* (2001) presented a solution for missing data items, whereby the data were filled by using the weighted nearest-neighbour method for applications of neural networks in the NO2 time-series, whereby the average of the neighbouring values in the series were used as a replacement. For time series analysis in this project, forward or backward filled missing value by nearest neighbours' data was more appropriate, it may be closer to the true value as compared to mean substitution.

B. Features Selection

Features selection is aimed to identify important features in the dataset and discard any other features which are irrelevant and redundant. The process of feature selection is a very important strategy, especially for algorithms that are computationally intensive when dealing with large datasets. Although additional attributes are added to a model, it may be able to predict a number better, but it will lead to the problem of slow convergence on those solutions either during the iterative learning process or the error minimisation process. Therefore, before the data set is used for modelling, those attributes to be used as predictors need to be selected.

The features selection for this project was based on domain knowledge rather than an analytical approach. From the research dataset, there were five independent attributes to be selected for modelling, which are wind speed (WS), wind direction (WD), relative humidity (Humidity), temperature (Temp) and air pollution index (API). The air pollutant index (API) is an indicator used to represent air quality status in the area under study. It is determined by the sub-index values computed based on the average concentration (for air pollutants, namely SO_2, NO_2, CO, O_3,

$PM_{2.5}$ and PM_{10}). The maximum sub-index of all six pollutants will be chosen as the API, and thus, it is not independent and related to the target variable to be studied – CO concentration. Therefore, the API variable was excluded as a predictor variable.

V. EVALUATION

After building a model from the dataset, the quality of the model needs to be examined. In the CRISP-DM process model, the evaluation phase is one of the major steps. For this research project, the historical data of past CO concentration experience with its corresponding meteorological parameters were given, and the objective was to predict the CO concentration when only other meteorological parameters were known. A few questions need to be answered before the model can be used. For example, "Are its predictions sufficiently accurate that makes its future application worthwhile?" If the predicted CO concentration is not accurate, the data is useless to the public and will affect the authorities reputation, which is the DOE of Malaysia. There are several different performance benchmarks to assess the relative merits of models, such as goodness-of-fit, robustness, forecasting accuracy and others.

Several model metrics can be used to evaluate the "goodness" of a model. Yeganeh *et al.* (2012) used root mean squared errors, relative mean errors and mean absolute relative error to evaluate model performance in the hourly CO concentrations prediction by using SVM (support vector machine) regression. Kolehmainen *et al.* (2001) applied root mean squared error (RMSE) to produce the numerical description of the goodness of the model estimates. In this paper, the selected statistical indicators which produced the numerical description of the goodness of estimates, are presented as follows: 1) Root mean squared error (RMSE) 2) Absolute error 3) Relative Error 4) Akaikes Information Criterion.

A. Root Mean Squared Error

The Mean square error (MSE) is often used as an error metric. In MSE, the difference in predicted and expected values in the records are observed and the value, which is then squared to retain the numerical quantity as well as eliminate the negative signs. RMSE on the other hand, will convert back the mean-squared error to the original data scale. RMSE is seen to have a more practical comparison because it appears in the same unit as the data.

B. Absolute Error

The absolute error sums up positive and negative errors to quantify the accuracy of the overall model, but without a clear indication of how the error varies. For example, it may seem that the error is almost balanced when both the positive and negative values are quite large. For accuracy, the performance of the model is reflected for the whole scored population. Therefore, this measure can be beneficial when one dimension of the error is considered.

C. Relative Error

By using a simple predictor, the relative error correlates the total error to the error. A simple model is utilised to be the baseline, taking the average value of all the expected values. The relative error will then show the difference between the model and the simple model.

D. Akaikes Information Criterion

To determine how a trained ARIMA model defines a time series, indicators such as Akaikes Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can be used.

These performance measures are calculated with the help of the ArimaTrainer operator in RapidMiner, and the calculated values are generated as a performance vector. The Akaike Information Criterion (AIC) measure is used extensively for the statistical model, whereby it measures the "goodness" of a model. In a comparison carried out between two models, the lower value of AIC showed that it was better than the one with a higher value.

VI. RESULTS AND DATA VISUALISATION

The data for the year 2004 was selected for model training. This was because the year 2004 data contained less missing value as compared to data of other years (Figure 4). There was no outlier detected as the minimum and maximum values were observed within a common range. There were missing data of 26 datapoints of wind speed (WS), 26

datapoints of wind direction (WD), 32 datapoints of API, 31 datapoints of Temperature and 26 datapoints of Relative Humidity (Humidity) were imputed. The missing values were imputed by filling the missing value with the nearest neighbour's data (Figure 5).

In Figure 6, it is shown that CO concentration was approximately normally distributed. From the scatter plot matrix in Figure 7, it was shown that there was no linear correlation between CO concentration and meteorological variables. In Figure 8, the polar plot also showed that the wind direction recorded were random and did not come from a particular direction. Therefore, it could be concluded that the data quality was good and comply with the assumptions of the parametric statistical model as below.

A. Assumption of Independency

Each variable in the documented effects was independent of each other. This was in line with the variable independency that was stated in Fisher's mathematics (Nisbet *et al.,* 2018).

B. Assumption of Normality

Nisbet *et al.* (2018) explained that based on Fisher's mathematics, each variable's distribution of values in a dataset will keep on a normal distribution around the mean value. The assumptions of normality and independency can be made when a classical parametric statistical procedure is applied. False assumptions can occur when there are significant departures from a normal distribution. These significant departures can cause the results to be biased and thus become untrustworthy. When some predictor variables are firmly linked to one another, it can cause significant departures from the assumption of independency, which will trigger more issues.

C. Assumption of Stationary

Stationary is a prerequisite before times series data can be applied with most statistical forecasting methods. Before the modelling starts, it is advised that the trends and seasonality found in time series datasets be eliminated. This is because trends and seasonality identify time series as non-stationary. Eventually, the mean value in trends can be varied, whereby there can be changing variance in seasonality.

A stationary time series has constant statistical properties, for instance, mean, variance and so on. The future values in stationary time series tend to become more predictable over time, and thus making this series simpler to model. To identify whether the time series is stationary or not, the line plot of series can be observed over time. To identify non-stationary series, the signs can be observed in series, on obvious trends, seasonality, or even some other systematic structures.

Name	Type	Missing	Statistics		
			Min	Max	Average
Date	Integer	0	20040101	20041231	20040667.123
Time	Integer	0	100	2400	1250
WS	Real	26	0.900	17.700	5.393
WD	Integer	26	0	360	176.498
API	Integer	32	27	74	45.629
COConc	Real	0	0.043	1.306	0.527
Temp	Real	31	20.700	43.100	28.834
Humidity	Real	26	34.600	95.600	76.136

Figure 4. Data for the year 2004 before pre-processing

Name	Type	Missing	Statistics		
			Min	Max	Average
Date	Numeric	0	20040101	20041231	20040667.123
Time	Numeric	0	100	2400	1250
WS	Real	0	0.900	17.700	5.393
WD	Numeric	0	0	360	176.567
API	Numeric	0	27	74	45.634
COConc	Real	0	0.043	1.306	0.527
Temp	Real	0	20.700	43.100	28.841
Humidity	Real	0	34.600	95.600	76.128

Figure 5. Data for the year 2004 after pre-processing

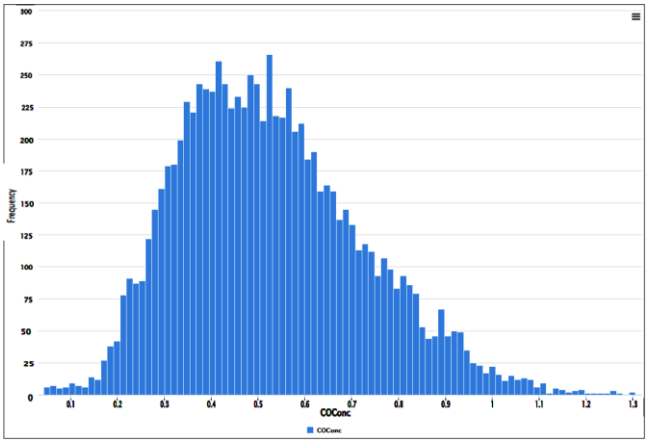

Figure 6. Histogram for CO concentration for the year 2004

From Figure 9 and Figure 10, the time series for the year 2004 CO concentration over time was stationary. This was also confirmed by the study by Hamid *et al.* (2017) on the study of CO concentration time series for Bachang, Melaka and Kuala Terengganu, in which after being tested by Augmented Dickey Fuller test, it indicated that they were stationary.

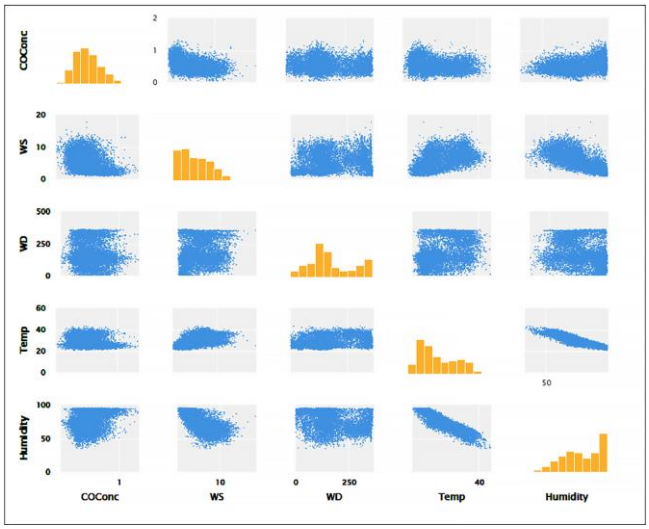

Figure 7. Scatter Plot CO concentration vs meteorological variables

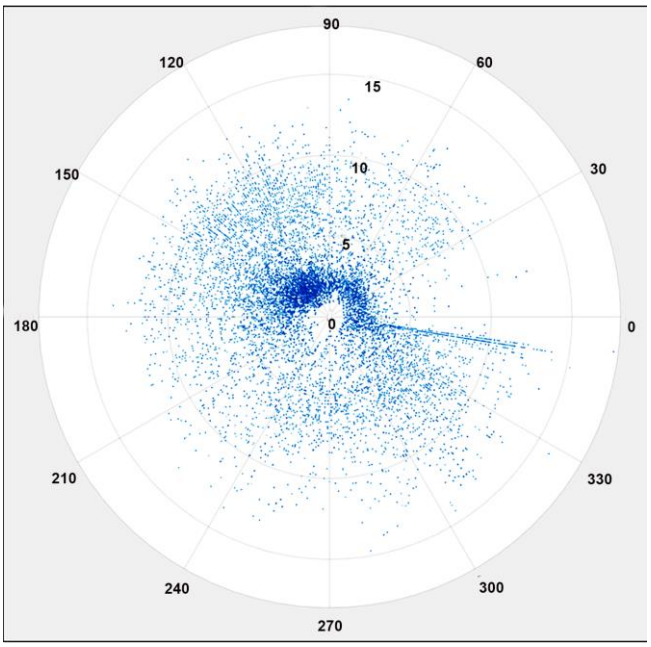

Figure 8. Polar coordinate plot wind direction vs wind speed

VII. DISCUSSION

In this section, the discussion is divided into two sections, which discuss the two different approaches used for the prediction. The first sub-section shows the time series with ARIMA approach, and the second subsection will discuss the machine learning and deep learning approach.

A. Time Series Approach (ARIMA)

The CO concentration is a set of quantitative observations arranged in chronological order and the same regularity of every hour, and thus time series analysis with ARIMA was used for the prediction. In the ARIMA model, the target variable was the CO concentration, and the predictor variable was the CO concentrations of the previous hour and time. In order to identify the best-fit parameters, various window sizes of window operator and ARIMA *p,d,q* parameters were explored, where some different combinations of terms were tried out to determine which combinations work best in the RapidMiner.

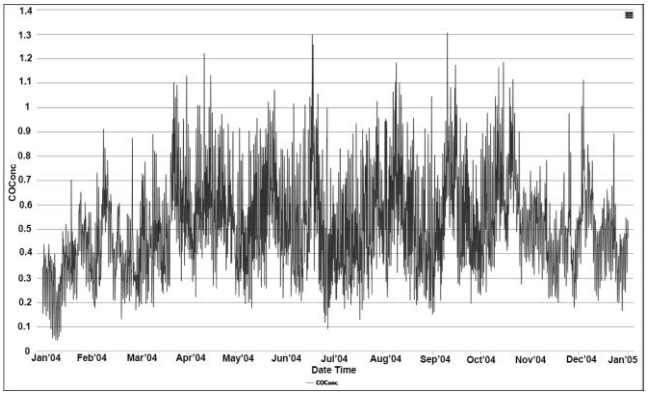

Figure 9. Time series plot for CO concentration

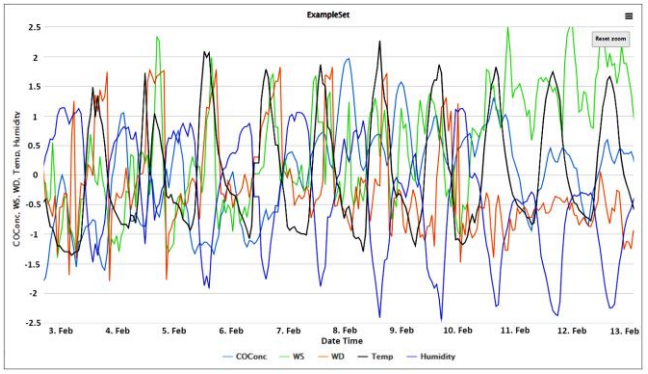

Figure 10. Time series plot for various meteorological variables (normalised & zoomed in)

ARIMA with p=2, d=0, q=1 is the most suitable model to predict the CO concentration in Langkawi Island as its AIC was lowest amongst other ARIMA (Table 1). The small AIC statistic values indicated the most appropriate model with the smallest error. ARIMA (2, 0, 1) or (AR, I, MA) means that the model described some response variable (Y). The value "0" embodies the 'I' or "Integrative' part of the model, can be ignored in the model, especially for stationary data.

Table 1 also shows that ARIMA (1,0,1) and ARIMA (1,0,2) is also a good model for prediction, these findings also aligned with the study by Hamid *et al.* (2017) in that the best model for CO concentration prediction for Bachang, Melaka was ARIMA (1,0,1) and the best model for CO concentration prediction for Kuala Terengganu was ARIMA (1,0,2). Table 2 shows the ARIMA prediction with new few hours' prediction. Figure 11 shows the time series plot of predicted CO concentration. From Figure 12, it is shown that CO concentration in Langkawi Island can be modelled as shown in Equation (1).

$$X_t=0.428+1.652X_{(t-1)}-0.744X_{(t-2)}+0.243e_{(t-1)}+e_t \quad (1)$$

The best model ARIMA (2,0,1) with a window size of 60 gives prediction ability similar or slightly better than the best model reported by Hamid *et al.* (2017), which are ARIMA (1,0,1) and ARIMA (1,0,2) from the study for Bachang and Kuala Terengganu, respectively. The results reported in Hamid *et al.* (2017) is shown in Table 3 as reference.

Table 1. ARIMA with various p,d,q settings

Window size	60	60	60	60	60
p,d,q	1,0,1	1,1,0	2,0,1	1,1,1	1,0,2
AIC	-209.74 ±51.30	-212.06 ±55.63	-228.49 ±55.33	-213.01 ±55.43	-219.17 ±53.07
BIC	-201.37 ±51.30	-205.82 ±55.63	-218.02 ±55.33	-204.70 ±55.43	-208.70 ±53.07
Root Mean Squared Error	0.035 ±0.033	0.685 ±0.277	0.030 ±0.031	0.569 ±0.257	0.033 ±0.031
Absolute Error	0.035 ±0.033	0.685 ±0.277	0.030 ±0.031	0.569 ±0.257	0.033 ±0.031
Relative Error	7. 26 ±7.6%	133. 75 ±41.5%	6. 35 ±7.1%	111. 89 ±42.8%	6. 74 ±7.0%

Table 2. ARIMA prediction with new few hours prediction

Prediction	**Next 1 hour**	**Next 2 hours**	**Next 3 hours**
AIC	-228.497 ±55.338	-228.484 ±55.329	-228.472 ±55.321
BIC	-218.025 ±55.338	-218.013 ±55.329	-218.000 ±55.321
Root Mean Squared Error	0.030 ±0.031	0.050 ±0.044	0.068 ±0.055
Absolute Error	0.030 ±0.031	0.046 ±0.042	0.061 ±0.05
Relative Error	6. 35 ±7.17 %	9. 71 ±10.11 %	12. 92 ±12.75 %

B. Time series approach (Windowing with Machine Learning and Deep Learning)

In the time series with windowing, the target variable was CO concentration. Five predictor variables used to train the time series model were CO concentration of previous hour, Wind Speed, Wind Direction, Temperature and Humidity. Amongst a few algorithms used in prediction for the data

year 2004 in the Windowing, deep learning gave the most accurate result, as shown in Table 4. RapidMiner deep learning operator was based on H2O open-source platform, a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent by using back propagation. It can contain a large number of hidden layers consisting of neurons with various activation functions.

After further optimised the parameters (window = 24 hours, Epochs = 20, with rectifier activation function), deep learning gave a relative error of 5.02 % for the next one-hour prediction, as shown in Table 5. It is shown that time series analysis with deep learning gave reasonably good CO concentration prediction for the next 3 hours with a relative error of less than or approximate 10%.

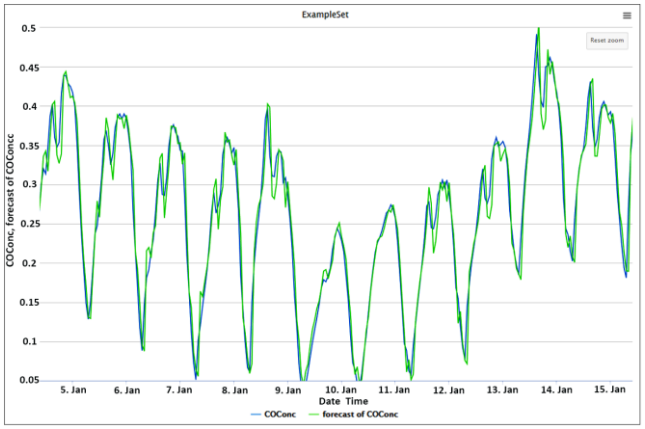

Figure 11. ARIMA (2,0,1) CO prediction for next 1 hour

VIII. PREDICTION WITH THE BEST MODEL BUILT

The result was reproducible from the model trained for the year 2004 data, applied to the data from the year 2006, giving a similar performance. After the optimum model of deep learning was built by using the year 2004 data; data from the year 2006 with a few purposely deleted values was fed into the model to test the model performance, as shown in Figure 13. The results were found to be satisfactory, as shown in Figure 14 and Figure 15, where the prediction for the next 1 hour and the next 3 hours gave a satisfactory result. However, for more the next 24 hours, the model was unable to give an accurate result, as shown in Figure 16.

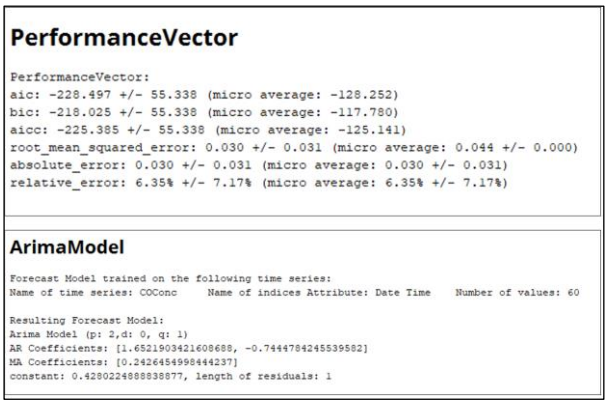

Figure 12. ARIMA (2,0,1) model performance result and parameters

Table 3. CO prediction with ARIMA for Bachang and Kuala Terengganu, Malaysia (Hamid *et al.*, 2017)

Location	Performance Indicator (Error Measure)		
	RMSE	***NAE***	***MAPE***
Bachang, Melaka	0.1119	0.0845	17.074
Kuala Terengganu	0.0773	0.0641	13.131

Table 4. Comparison of algorithm in Windowing

Algorithm	**Gradient Boosted Tree**	**Generalised Linear Model**	**Decision Tree**	**Deep Learning**	**Random Forest**	**SVM**
Root Mean Squared Error	0.037 ±0.001	0.038 ±0.001	0.054 ±0.002	0.034 ±0.001	0.084 ±0.003	0.137 ±0.003
Absolute Error	0.027 ±0.00	0.027 ±0.001	0.038 ±0.001	0.026 ±0.001	0.064 ±0.002	0.101 ±0.003
Relative Error	5. 93 ±0.24%	5. 75 ±0.23%	8. 20 ±0.28%	5.61 ±0.26%	14. 91 ±0.76%	22. 73 ±1.02%

Table 5. Deep learning prediction with new few hours

Prediction	Next 1 hour	Next 2 hours	Next 3 hours
Root Mean Squared Error	0.032 ±0.001	0.050 ±0.002	0.063 ±0.002
Absolute Error	0.023 ±0.001	0.037 ±0.001	0.047 ±0.01
Relative Error	5. 02 ±0.29 %	7. 89 ±0.20 %	10. 13 ±0.42 %

Date Time	Actual COConc	COConc	prediction(COConc + 3 (horizon))
Jan 3, 2006 3:00:00 PM SGT	0.521	0.521	0.551
Jan 3, 2006 4:00:00 PM SGT	0.459	0.459	0.500
Jan 3, 2006 5:00:00 PM SGT	0.410	0.410	0.458
Jan 3, 2006 6:00:00 PM SGT	0.384	0.384	0.485
Jan 3, 2006 7:00:00 PM SGT	0.403	0.403	0.458
Jan 3, 2006 8:00:00 PM SGT	0.420	0.420	0.437
Jan 3, 2006 9:00:00 PM SGT	0.423	0.423	0.455
Jan 3, 2006 10:00:00 PM SGT	0.417	0.417	0.458
Jan 3, 2006 11:00:00 PM SGT	0.401	0.401	0.412
Jan 4, 2006 12:00:00 AM SGT	0.411	0.411	0.416
Jan 4, 2006 1:00:00 AM SGT	0.414	0.414	0.438
Jan 4, 2006 2:00:00 AM SGT	0.393	0.393	0.381
Jan 4, 2006 3:00:00 AM SGT	0.345	0.345	0.357
Jan 4, 2006 4:00:00 AM SGT	0.325	0.325	0.366
Jan 4, 2006 5:00:00 AM SGT	0.299	0.299	0.327
Jan 4, 2006 6:00:00 AM SGT	0.273	0.273	0.289
Jan 4, 2006 7:00:00 AM SGT	0.280	0.280	0.313
Jan 4, 2006 8:00:00 AM SGT	0.385	0.385	0.355
Jan 4, 2006 9:00:00 AM SGT	0.425	0.425	0.381
Jan 4, 2006 10:00:00 AM SGT	0.454	0.454	0.435
Jan 4, 2006 11:00:00 AM SGT	0.481	0.481	0.504
Jan 4, 2006 12:00:00 PM SGT	0.489	?	0.530
Jan 4, 2006 1:00:00 PM SGT	0.514	?	0.574
Jan 4, 2006 2:00:00 PM SGT	0.556	?	0.609

attributes, 3 regular attributes)

Figure 13. Data set of year 2006 used to test model

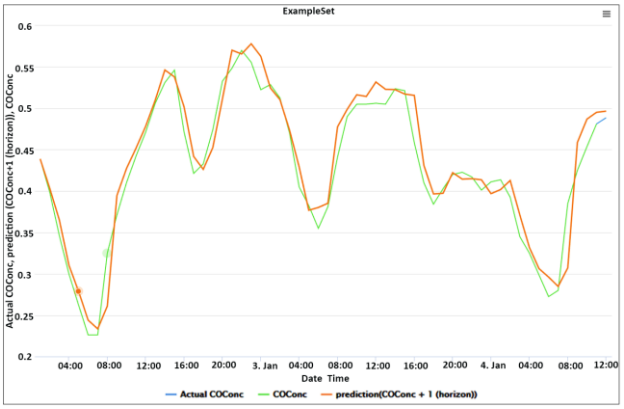

Figure 14. Deep learning CO prediction for next 1 hour

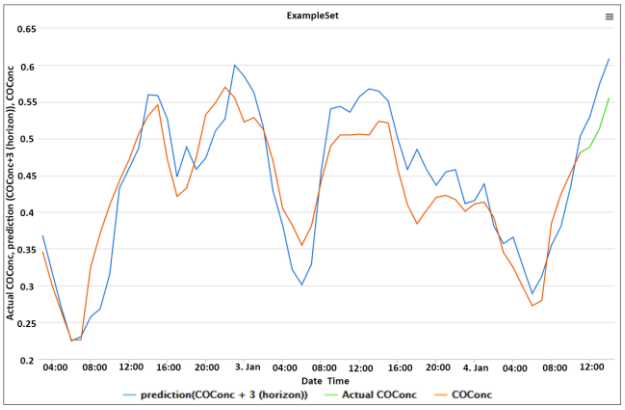

Figure 15. Deep learning CO prediction of next 3 hours

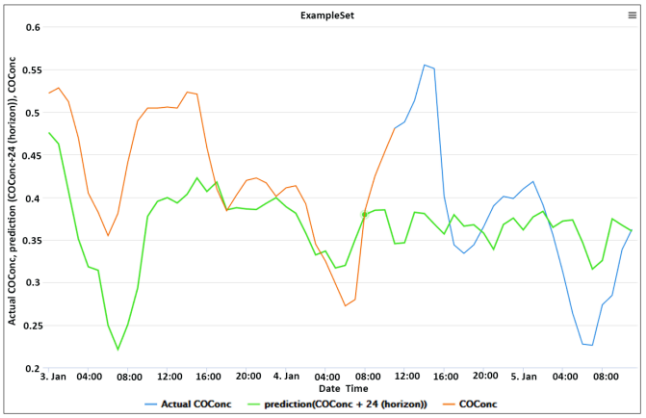

Figure 16. Deep learning CO prediction for next 24 hours

IX. CONCLUSION

It is very important to design air pollution forecasting models appropriately, as the models can help improve the management of air quality. Together with the implementation of such models, the efforts to improve the techniques of forecasting accuracy are also very significant. The temporal elements (time series) are the important variable to make an accurate CO concentration prediction.

From this study, both time series approaches in RapidMiner- (1) ARIMA and (2) Windowing with deep learning gave satisfactory results, where Windowing with deep learning was more superior. Regardless, time series with ARIMA gave more interpretable results where the model can be translated into mathematics equations.

However, Windowing with deep learning was more superior not only in terms of low relative error but also in terms of more variables that can be used in the model building. In ARIMA, only a univariate variable (CO concentration at a particular time) was used in model

building. But for Windowing with deep learning, besides CO concentration and time, other variables such as wind direction, wind speed, temperature and humidity parameters were also used as the predicator variables. This gave a more accurate and generalised model. In this work, it was demonstrated that RapidMiner Studio is a useful tool for CO prediction. Therefore, in future work, model deployment features in RapidMiner Studio can be further explored.

X. ACKNOWLEDGEMENT

The authors are grateful to the Department of Environmental Malaysia (DOE) for providing necessary data for this research.

XI. REFERENCES

Ambient (outdoor) air pollution, 2018, World Health Organization, viewed 5 November 2019, <https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health>.

Baumbach, G 1996, 'Air quality control: Formation and sources, dispersion, characteristics and impact of air pollutants — measuring methods, techniques for reduction of emissions and regulations for air quality control, Springer-Verlag Berlin Heidelberg, 1st edn, Berlin.

Berg JM, Tymoczko JL & Stryer L 2002, 'Section 10.2, Hemoglobin transports oxygen efficiently by binding oxygen cooperatively', Biochemistry, 5th edn, WH Freeman, New York.

Chris Higgins, October 2005, Causes and clinical significance of increased carboxyhemoglobin, Radiometer Medical ApS, viewed 23 November 2019, <https://acutecaretesting.org/en/articles/causes-and-clinical-significance-of-increased-carboxyhemoglobin>.

Clean air is a human right: WHO, 2018, DownToEarth, viewed 2 November 2019, <https://www.downtoearth.org.in/news/ air/clean-air-is-a-human-right-who-62023>.

Hamid, HA, Japeri, AZUM, & Ahmat, H 2017, 'Characteristic and prediction of Carbon Monoxide concentration using time series analysis in selected urban area in Malaysia', MATEC Web of Conferences, vol. 103

Heo, J & Kim, D 2004, 'A new method of ozone forecasting using fuzzy expert and neural network systems', Science of the Total Environment, vol. 325, no. 1-3, pp. 221-237.

James, T 2017, 'Four problems in using CRISP-DM and how to fix them', viewed 2 October 2019, <https://www.kdnuggets.com>.

Kolehmainen, M, Martikainen, H & Ruuskanen, J 2001, 'Neural networks and periodic components used in air quality forecasting', Atmospheric Environment, vol. 35, no.5, pp. 815- 825.

Lu, W & Wang, D 2008, 'Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme', Science of the Total Environment, vol. 395, no. 2-3, pp. 109-116.

Nisbet, R, Miner, G & Yale, K 2018, 'Handbook of statistical analysis and data mining applications', Elsevier, London.

Penney, D, Benignus, V, Kephalopoulos, S, Kotzias, D, Kleinman, M & A 2010, Carbon monoxide. In: WHO Guidelines for Indoor Air Quality: Selected Pollutants, Geneva: World Health Organization; 2 October 2019, <https://www.ncbi.nlm.nih.gov/books/NBK138710/>.

Pyle, D 2003, 'Data collection, preparation, quality, and visualization', The Handbook of Data Mining, Lawrence Erlbaum Associates, Inc., New Jersey, pp. 366-391.

READY (Real-time Environmental Applications and Display System), 2019, United States NOAA-EPA Air Quality Forecasting System, viewed 20 December 2019 < https://www.ready.noaa.gov/index.php>.

Shaadan, N, Rusdi, MS, Azmi, NNSNM, Talib, SF & Azmi, WAW 2019, 'Time series model for carbon monoxide (CO) at the several Industrial sites In Penisular Malaysia', Malaysian Journal of Computing, vol. 4, no. 1, pp. 246-260.

Tonellato, S 2001, 'A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentration', Applied Statistic, vol. 50, no. 2, pp. 187-200.

UK AIR, 2021, Department for Environment Food & Rural Affairs, viewed 26 November 2020 <https://uk-air.defra.gov.uk/>.

Venkatasubramanian, V 2019, 'The promise of artificial intelligence in chemical engineering: Is it here, finally?', AIChE Journal, vol. 65, no. 2, pp. 466-478.

Vesilind, PA, Peirce, JJ & Weiner, RF 2013, 'Environmental Pollution and Control' Technology & Engineering, 3th edn, Elsevier Science

Yeganeh, B, Motlagh, MSP, Y, Rashidi, Y & Kamalan H 2012, 'Prediction of CO concentrations based on a hybrid partial least square and support vector machine model. atmospheric environment', Atmospheric Environment, vol. 55, pp. 357-365.