Sains Malaysiana 53(11)(2024): 3831-3843
http://doi.org/10.17576/jsm-2024-5311-24

The Effect of Meteorology and Air Quality to the COVID-19 Cases in Malaysia: A Multivariate Deep Learning Approach

(Kesan Meteorologi dan Kualiti Udara kepada Kes COVID-19 di Malaysia: Suatu Pendekatan Pembelajaran Mendalam Multivariat)

PEGGY YEO[1], AZURALIZA ABU BAKAR[1,*], ZALINDA OTHMAN[1], MAZRURA SAHANI[2], SUHAILA ZAINUDIN[1] & ZAILIZA SULI[3]

[1]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[2]Center for Toxicology and Health Risk Studies (CORE), Faculty of Heath Sciences, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia
[3]Hulu Langat District Health Office, 43000 Kajang, Selangor, Malaysia

Received: 7 December 2023/Accepted: 4 October 2024

ABSTRACT

In October 2022, the World Health Organization (WHO) reported that over six hundred million people globally had been infected by the COVID-19 pandemic, leading to six million deaths. Malaysia, like many other countries, has experienced significant economic and societal impacts due to COVID-19. Previous research has identified meteorological conditions and air quality as critical factors influencing the spread of infectious diseases like influenza. In this study, we explore the impact of meteorological and air quality factors on COVID-19 case numbers in Malaysia, focusing on a case study in the Hulu Langat district of Selangor state, utilizing a deep learning approach. Our model, which employs a neural network architecture incorporating both Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), was trained using multivariate time-series data. This data included meteorological and air quality metrics from the Department of Environment, Malaysia, and COVID-19 case data collected from the Hulu Langat Health Office. We prepared three datasets for predictive modeling: one combining all features, one including only meteorological data, and another with only air quality data. Our results indicate that the CNN model outperformed the LSTM model in terms of prediction accuracy. Furthermore, the dataset incorporating all features resulted in the lowest prediction error, compared to datasets with only meteorological or air quality features. Feature importance analysis showed that air quality factors were the most significant predictors, suggesting that air quality has a greater impact on COVID-19 case numbers than meteorological factors.

Keywords: Air quality; COVID-19 cases; feature ranking; meteorology; multivariate LSTM and CNN

ABSTRAK

Pada Oktober 2022, Pertubuhan Kesihatan Sedunia (WHO) melaporkan bahawa lebih enam ratus juta orang di seluruh dunia telah dijangkiti oleh pandemik COVID-19, yang membawa kepada enam juta kematian. Malaysia, seperti kebanyakan negara lain, telah mengalami kesan ekonomi dan sosial yang ketara akibat COVID-19. Penyelidikan sebelum ini telah mengenal pasti keadaan meteorologi dan kualiti udara sebagai faktor kritikal yang mempengaruhi penyebaran penyakit berjangkit seperti influenza. Dalam kajian ini, kami meneroka kesan faktor meteorologi dan kualiti udara terhadap bilangan kes COVID-19 di Malaysia, memfokuskan kepada kajian kes di daerah Hulu Langat, Selangor menggunakan pendekatan pembelajaran mendalam. Model kami yang menggunakan seni bina rangkaian saraf yang menggabungkan Memori Jangka Pendek Panjang (LSTM) dan Rangkaian Neural Konvolusi (CNN) telah dilatih menggunakan data siri masa multivariat. Data ini termasuk metriks meteorologi dan kualiti udara daripada Jabatan Alam Sekitar, Malaysia dan data kes COVID-19 yang dikumpul daripada Pejabat Kesihatan Hulu Langat. Kami menyediakan tiga set data untuk pemodelan ramalan: satu menggabungkan semua ciri, satu hanya data meteorologi dan satu lagi dengan hanya data kualiti udara. Keputusan kami menunjukkan bahawa model CNN mengatasi model LSTM dari segi ketepatan ramalan. Tambahan pula, set data yang menggabungkan semua ciri menghasilkan ralat ramalan yang paling rendah, berbanding set data dengan hanya ciri meteorologi atau kualiti udara. Analisis kepentingan ciri mendedahkan bahawa faktor kualiti udara adalah peramal yang paling penting, menunjukkan bahawa kualiti udara mempunyai kesan yang lebih besar terhadap bilangan kes COVID-19 berbanding faktor meteorologi.

Kata kunci: Kedudukan ciri; kes COVID-19; kualiti udara; meteorologi; multivariat LSTM dan CNN

INTRODUCTION

COVID-19 is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The virus will cause an infection in the respiratory tract. The first case detected by the world was a case that occurred in Wuhan, Hubei, China, in December 2019. The dense population has been one of the catalysts for forming the COVID-19 pandemic recognized by the World Health Organization (WHO). As of October 28, 2022, the epidemic has infected more than six hundred million people worldwide, with six million deaths based on World Health Organization investigations (WHO 2022).

The activity of COVID-19 is believed to be influenced by environmental factors because the virus is spread through the air. Some studies have found that environmental factors such as weather, temperature and humidity can affect the spread of the virus (Valsamatzi-Panagiotou & Penchovsky 2022). The study has proven that the coronavirus can survive longer in lower temperatures and higher humidity. In addition, a study also states a positive correlation between the spread of COVID-19 and air pollution (Khan et al. 2021). The longer the time exposed to air pollution, the higher the risk of contracting COVID-19 (Ali & Islam 2020). Most studies state that the spread of COVID-19 will be more widespread, especially in locations with a cold climate, high humidity, and high air pollution.

Predicting COVID-19 cases related to meteorology and air quality data has attracted researchers in artificial intelligence (AI) technology to achieve accurate and fast predictions. As part of AI, deep learning technology has been used to solve complex data prediction. This technique can be used to study the health sector, especially in epidemiology, such as COVID-19, to benefit the community. Many studies employing machine learning for predicting COVID-19 cases have been conducted since 2020. Several studies have focused on meteorology and air quality data's impact on the increase in cases. The effect of meteorological and air quality parameters may vary in different countries or climatic regions. Therefore, this study aims to investigate the effect of meteorology and air quality data on COVID-19 cases in Malaysia.

In this study, we investigate the relationship between environmental factors such as weather temperature, air humidity and air quality with cases of COVID-19 in Hulu Langat Malaysia to produce accurate prediction models. We developed the prediction model using two deep learning algorithms, namely the Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) algorithm. Besides that, we estimate both methods' prediction accuracy and identify the important parameters that 1contribute to the accurate prediction. Three data sets, the COVID-19 cases, meteorological (temperature, humidity, wind speed), and air quality data (PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, CO), were used for the prediction modelling.

The contribution of this paper are as follows: 1) Integrating the meteorology, air quality, and COVID-19 cases dataset of Hulu Langat district, Selangor, is a novel approach to employing deep learning for prediction models in Malaysia, 2) We performed feature ranking to find the most affecting factors and showed that air quality factors contribute most to the accurate COVID-19 case prediction for the Hulu Langat region, and 3) Our prediction model showed that CNN deep learning outperforms LSTM in prediction accuracy and error.

This paper is organized into five sections. The related works on machine learning methods in predicting COVID-19 cases worldwide and the previous studies on finding the effect of air quality and meteorology data on COVID-19 cases are presented in the first section. Next section comprehensively describes the material and methods of predictive modelling using deep learning methods. The analysis, discussion of results and concluding remarks are presented in the last two sections, respectively.

RELATED WORK

LONG SHORT-TERM MEMORY (LSTM) AND CONVOLUTIONAL NEURAL NETWORKS (CNN)

The Long Short-Term Memory Networks (LSTM) and the Convolutional Neural Networks (CNN) are the extended variation of neural network algorithms with some specific specialties. The Long Short-Term Memory (LSTM) network is a Recurrent Neural Network (RNN) designed for sequence problems. The LSTMs have recurrent connections so that the state from previous activations of the neuron from the previous time step is used as context for formulating an output. LSTM has a unique feature that allows it to avoid the problems that prevent the training and scaling of other RNNs.

The LSTM architecture was motivated by an analysis of error flow in existing RNNs, which found that long-time lags were inaccessible to existing architectures because back propagated error either blows up or decays exponentially. An LSTM layer consists of a set of recurrently connected memory blocks. These blocks can be considered a differentiable version of the memory chips in a digital computer. Each one contains one or more recurrently connected memory cells and three multiplicative units - the input, output and forgets gates - that provide continuous analogues of write, read and reset operations for the cells (Nielsen 2015).

Convolutional Neural Networks (CNN) are a neural network designed to handle image data efficiently. They have proven effective in computer vision problems, image classification and providing a component in hybrid models for new problems such as object localization and image captioning. CNN automatically processes raw data, such as

raw pixel values, instead of domain-specific or handcrafted features derived from the raw data. The model performs representation learning and automatically extracts the features from the raw data, which is helpful for the problem being addressed.

The ability of CNNs to learn and automatically extract features from raw input data can be applied to time series forecasting problems. A sequence of observations can be treated like a one-dimensional image that a CNN model can read and distill into the most salient elements. CNNs get the benefits of Multilayer Perceptrons for time series forecasting, namely support for multivariate input, multivariate output and learning arbitrary but complex functional relationships, but do not require that the model learn directly from lag observations. Instead, the model can learn a representation from a large input sequence most relevant to the prediction problem (Goodfellow, Bengio & Courville 2016).

MACHINE LEARNING APPROACHES IN COVID-19 PREDICTION

In this section, we reviewed related works regarding Machine Learning approaches in COVID-19 prediction and the studies that investigate the effect of meteorology and air quality data to the increase of the COVID-19 cases. Ogunjo, Fuwape and Rabiu (2022) conducted the COVID-19 case prediction using the meteorological and air quality data. They utilized machine learning methods such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (K-NN) to construct the prediction models. The dataset was divided into a training set (80%) and a test set (20%) for the modelling process. The evaluation metric used was the Root Mean Square Error (RMSE), with 67, 87, 86, and 65 accuracies observed for the DT, KNN, SVM, and RF models. The findings indicate that climate temperature is the most relevant parameter for predicting COVID-19 cases, followed by air humidity.

Zhou (2022) conducted a study in the United States, analyzing the impact of meteorological factors and $PM_{2.5}$ on COVID-19 spread from April 12, 2020, to October 13, 2020. They found that rain indirectly affected cases through air humidity, while temperature and wind speed were closely related to the number of cases due to virus survival and spread. Higher $PM_{2.5}$ concentrations and poorer air quality were associated with increased COVID-19 hospitalization rates, emphasizing the importance of considering these factors in assessing the virus's spread.

Ramirez-Alcocer et al. (2022) employed the LSTM method to predict COVID-19 cases in Mexico using meteorological and air quality data. The study used the LSTM algorithm for modelling and K-means clustering for parameter selection. Data from the third wave of COVID-19 in Victoria, Mexico, along with meteorological and air quality data from May 27 to October 13, 2021, were utilized. The prediction model achieved high accuracy, with RMSE 0.0892, MAE 0.0592, and MAPE 0.2061 in the test stage and RMSE values between 0.4152 and 3.9084, and less than 4.1% for MAPE in the validation stage using three different data sets. The study found that air humidity and climate temperature were influential variables in predicting the death rate of COVID-19.

Yu et al. (2021) investigates the effect of air pollution on COVID-19 cases. The air quality data, including $PM_{2.5}$, ozone (O_3), NO_2, and air humidity, was obtained from the 'World's Air Pollution: Real-time Air Quality Index', while COVID-19 case data was obtained from the Our World in Data portal. Artificial neural networks (ANNs) were used to estimate variables' maximum and minimum values. The results showed that long-term exposure to air pollution increased the probability of death from COVID-19. Higher concentrations of air pollution were associated with an increased death rate in COVID-19 patients due to the potential impact of air pollution on the virus spread.

Adhikari and Yin (2021) conducted a study in 2021 to examine the effect of ozone lag, $PM_{2.5}$, and meteorological factors on COVID-19 cases in New York. The research investigated the impact of ozone lag, $PM_{2.5}$, and meteorological data, including wind speed, climate temperature, relative humidity, absolute humidity, and cloud percentage, on COVID-19 cases. The data set used in the study comprised confirmed COVID-19 cases from March 1 to April 11, 2020, obtained from the USA FACTS portal page, while meteorological data was collected from the United States Environmental Protection Agency (AQS). The findings indicated that lower temperatures contributed to the prolonged survival of the virus, and ozone, $PM_{2.5}$, and five meteorological factors significantly influenced new COVID-19 cases with a lag of 9 to 13 days. The review by Ghobakhloo et al. (2022) reported that most investigations in the present study showed that air pollution and meteorology could be essential in transmitting COVID-19. Exposure to air pollutants, particularly $PM_{2.5}$ and NO_2, positively affects COVID-19 patients and mortality. Temperature and humidity variables are negatively correlated with virus transmission.

In Malaysia, several studies investigating the effects of meteorology and air quality parameters on COVID-19 cases cover data analysis over a specific duration of time. Several researchers employed statistical methods, and only a few used machine learning methods (Jalaludin et al. 2023; Lloyd & Viswanathan 2022; Mohan et al. 2022; Shen, Bakar & Mohamad 2023). Lloyd and Viswanathan (2022) employed Statistical analyses such as Pearson's correlation, factor analysis, and factor score for data analysis and interpretation. Their finding showed that Temperature, Wind Speed, Pressure, and Dew Point are significant parameters of the COVID-19 outbreak in Malaysia. Different meteorological parameters influence every lockdown phase. Mohan et al. (2022) investigated the

effects of meteorological parameters and air quality on the COVID-19 pandemic spreading during three MCO phases in Malaysia. Time series analysis determined that higher relative humidity and temperature in Peninsular Malaysia and high amounts of high precipitation, relative humidity, and good air quality in East Malaysia have controlled the spreading during MCO phase 3.

Jalaludin et al. (2023) employs a machine learning technique called Boosted Trees (BT) to find the association between air pollutants, meteorological parameters, air pollutants, and the transmission of COVID-19 in Selangor and Kuala Lumpur, Malaysia. The duration of the cumulative daily cases of COVID-19 was taken from 18 March to 26 May 2020 (3 months). This study highlights the significant correlation between air pollution, meteorological parameters, and COVID-19 cases in Malaysia. The findings also indicate that COVID-19 cases were positively correlated with O_3, NO_2, RH, PM_{10}, and $PM_{2.5}$ but negatively correlated with the meteorology parameters. Hasan and Jamaludin (2023) employed LSTM and GRU to forecast verified COVID-19 fatalities in Malaysia, Egypt, and the U.S. Their findings indicated that GRU outperformed LSTM in predicting COVID-19 cases in the three countries. Madini, Mutasem and Reem (2022) employed RNN and LSTM to predict the number of confirmed cases of COVID-19 in Malaysia, Morocco and Saudi Arabia. Shen, Bakar and Mohamad (2023) employed Long Short-Term Memory deep learning to investigate the effect of meteorology factors on COVID-19 cases in Malaysia and discovered that temperature and humidity are critical factors in the increase of cases. They also reported that different states showed the effect of different factors.

There are limited studies on employing deep learning approaches in finding the effect of meteorology and air quality data on COVID-19 cases in Malaysia. Most research reports the effects of air quality during the MCO in Malaysia, not the effect of air quality on the COVID-19 case trends (Mohd Halim et al. 2022). Several researchers used statistical methods to investigate the effects of air quality and meteorology on COVID-19 cases. Due to these limitations, in this paper, we investigate the advantage of deep learning in finding the effect of air quality and meteorology factors on the COVID-19 cases in Malaysia.

MATERIALS AND METHODS

The implementation of predictive models follows the standard data analytics methodology that consists of four phases: Business understanding, data understanding and preparation, model development, and model evaluation. These phases are part of the CRISP-DM data mining methodology (Kelleher, Namee & D'Arcy 2020).

Business Understanding

Business understanding involves identifying the analytical business goal and questions. In this study, our business goal was to identify the parameters of air quality and meteorology that affect the COVID-19 case prediction. It begins with developing the prediction model and determining the factors or features contributing to the accurate models. The feature selection approach ranks the critical features that affect the COVID-19 case prediction.

Data Understanding and Preparation

The dataset used in this study was obtained from three sources. The COVID-19 patients and positive cases dataset was obtained from Selangor's Hulu Langat district Health Department. Data ethics approval: NMRR ID-22-00719-

TABLE 1. Dataset description

Dataset	Source of Data	Parameters
COVID-19 Cases	Hulu Langat Health Department, Selangor	COVID-19 positive patients profile and symptom data in Hulu Langat District Number of cases (by area) Ampang (1228), Beranang (222), Cheras (1292), Hulu Langat (306), Kajang (3013), Semenyih (803)
Meteorology Data	The Department of Environment Malaysia	Average wind direction, average wind speed, average humidity, average temperature 5848 records for year 2020 (2928) and year 2021 (2920) records, daily weather records for the state of Selangor from January 2020 to December 2021.
Air Quality Data	The Department of Environment	PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, CO 5848 records for year 2020 (2928) and year 2021 (2920) records, daily weather records for the state of Selangor from January 2020 to December 2021

IBZ (Public Health/Epidemiology). Hulu Langat is one of the largest districts in Selangor, with a vast amount of COVID-19 cases. The data records of COVID-19 were collected from December 2020 to December 2021. The data was divided into six areas in Hulu Langat: Ampang, Beranang, Cheras, Hulu Langat, Kajang and Semenyih. Table 1 shows the distribution of COVID-19 cases in each area in Hulu Langat district.

The meteorological data obtained from the Department of Environment has a total of 5848 data records, which is the year 2020 has 2928 data records while the year 2021 has 2920 data records, which are daily weather records for the state of Selangor from January 2020 to December 2021. The meteorological data consists of four parameters: Average Wind Direction, Average Wind Speed, Average Humidity, and Average Temperature. The air quality data comprises six parameters: PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, and CO. We prepared three sets of data: The all-features dataset, which uses the meteorology, air quality, and COVID-19 cases data; the meteorology dataset; and the air quality dataset. Three prediction models were developed using the deep learning methods from these datasets. Table 1 describes the datasets used in the study.

The daily active COVID-19, meteorological and air quality data were examined at the beginning of data pre-processing to acquire the appropriate dataset integration for the prediction model. The data were visualized in histograms and density plots to improve data understanding. Finally, the cleaned data set was used to construct the prediction model. The performance of the prediction model was evaluated through evaluation metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error), and R2-Score to find the most suitable model for the prediction of daily active cases of COVID-19 in Selangor.

DEEP LEARNING METHODS FOR MODEL DEVELOPMENT

This paper employed two well-known deep learning algorithms for the model development. Figure 1 shows the LSTM architecture employed in this paper. LSTM consists of memory blocks known as cells. Two states will be transferred to the front cell, namely the cell state and the hidden state, which have three layers: the input layer, the output layer, and the hidden layer. The memory block is responsible for storing and manipulating memory by performing three main mechanisms: the forget gate f_t, input gate i_t and output gate o_t. Each gate has its function and task.

Forget Gate The first step in implementing the LSTM algorithm is to remove information from the cell state made by the forget gate f_t. Equation (1) is used to calculate the importance of information; if the result is equal to 0, it means that the data needs to be deleted, while the result equal to 1 means that the data is essential and needs to be stored.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f). \tag{1}$$

Input Gate The next step is identifying the new information stored in the cell state. A sigmoid layer named input gate (i_t) identifies the value to be updated while a *tanh* layer C_t produces a vector of new values that can be added, and the combination of these two layers will be updated to the state in front as shown in Equation (2).

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ C_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \end{aligned} \tag{2}$$

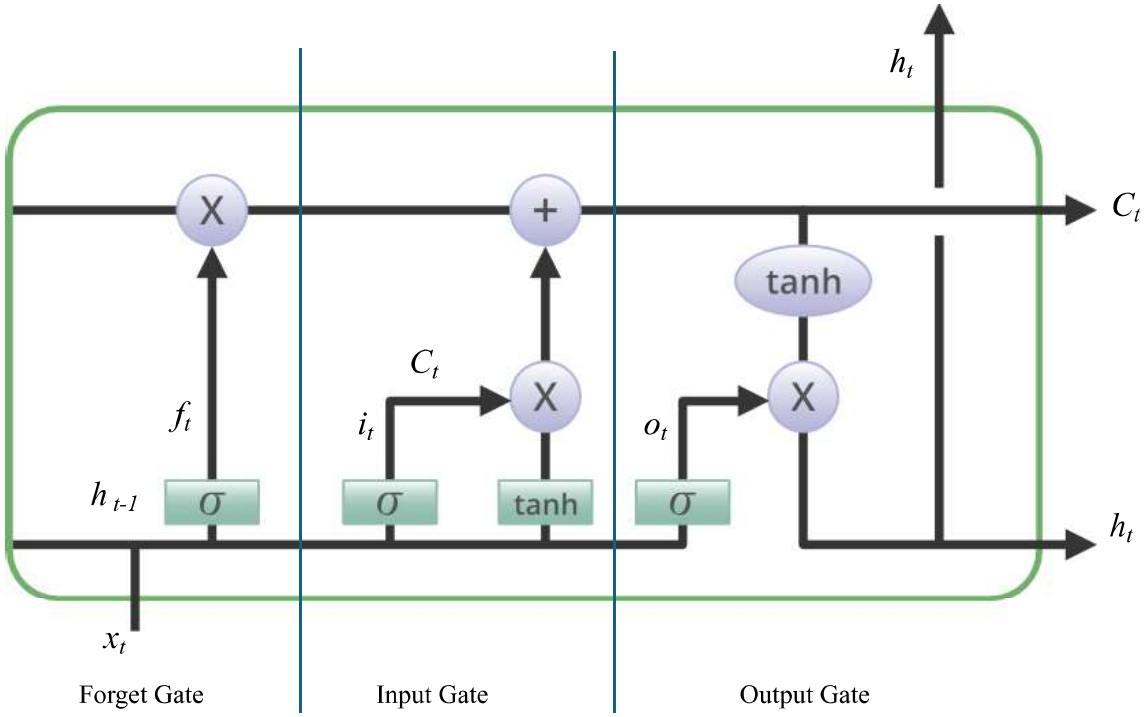
Output Gate Finally, a sigmoid layer will be run to determine the part of the cell state that will be displayed. The cell state will be calculated using tanh to get a value between -1 and 1 before multiplying with the value of the sigmoid gate so that it can display only the desired part (Felix Gers 2015). The layer that performs the task is the output gate o_t shown in Equation (3). The output data will be based on the cell state but will be filtered beforehand.

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{3}$$

Convolutional Neural Networks (CNNs) are traditionally used for image data, but they can also be effectively applied to non-image data, particularly in time series and sequence-based tasks. In time series or sequence data, CNNs can be applied using one-dimensional (1D) convolutions. Instead of sliding a 2D filter over an image, a 1D filter is slide across the sequence data. This allows the network to capture local patterns in the data over time, such as trends or seasonal effects in time series data. Figure 2 shows the CNN architecture employed in this paper.

Model Evaluation

We developed three LSTM and three CNN models based on three datasets, the all-features, air quality model and meteorology model, to predict daily COVID-19 cases for areas in the Hulu Langat district. The time series data from 1 October 2020 to 31 December 2021 was selected for the study and divided into two parts; the first thirteen months' data were selected as training sets, while the remaining month's data were made as test sets. The evaluation metrics used to evaluate the prediction model of daily active cases of COVID-19 includes Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and R2-Score (Kelleher, Namee & D'Arcy 2020).



https://www.geeksforgeeks.org/understanding-of-lstm-networks/

FIGURE 1. The LSTM architecture

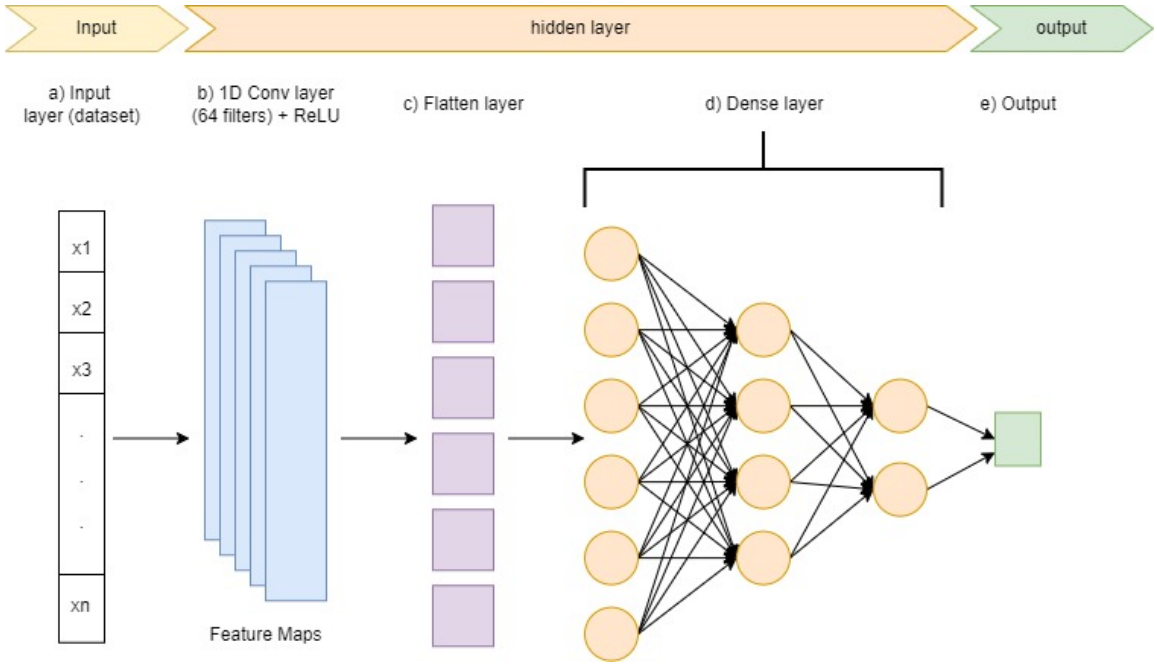


FIGURE 2. 1D convolutional layer for CNN architecture

MAE measures the average absolute difference between predicted and actual values as in Equation (4). It indicates how far, on average, the forecast is from the actual value. The MAE is calculated by taking the average absolute difference between the predicted and actual values where $t_1 \ldots t_n$ is a set of n expected values, and $M(d_1) \ldots M(d_n)$ is a set of n predictions for a set of test instances, $d_1 \ldots d_n$.

$$MAE = \frac{\sum_{i=1}^{n} abs(t_i - M(d_i))}{n} \quad (4)$$

RMSE is another widely used evaluation method for measuring residual variance. RMSE represents the average squared difference between the actual and the predicted value in the data set and gives higher weight to larger errors as in Equation (5).

$$RMSE = \frac{\sum_{i=1}^{n} (t_i - M(d_i))^2}{n} \quad (5)$$

MAPE measures the average percentage difference between the predicted and actual values. It calculates the absolute percentage difference between each forecast and the corresponding actual value to obtain an average of these differences. MAPE can give an understanding of the relative error in terms of percentage. MAPE is the most frequently used metric for forecasting because it is easier to understand. MAPE works well if there are no extremes in the data. Where n is the number of points; A_t is the actual value; and F_t is the predicted value. (Refer to Equation (6)).

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \quad (6)$$

The R^2 score is a statistical measure representing the proportion of variance in the dependent variable that can be explained by the independent variable in the regression model. It indicates how well the model fits the data. The R^2 score ranges from 0 to 1, with 1 indicating a perfect fit and 0 indicating that the model does not explain any variance. The R^2 score is usually used to evaluate the goodness-of-fit of the model as in Equation (7).

$$R^2 = 1 - \frac{sum\ of\ squared\ errors}{total\ sum\ of\ squares} \quad (7)$$

The lower value of MAE, MAPE, and RMSE implies the higher accuracy of a regression model. However, a higher value of R^2 is considered desirable.

RESULTS AND DISCUSSION

We performed the analysis of results in two forms. First, the performance of the deep learning method CNN and LSTM on the prediction error are analyzed. Secondly, we observed the performance of three datasets: 1) the all-features dataset (COVID-19 cases, meteorology, and air quality features), 2) the meteorology dataset (meteorology and COVID-19 case features), and 3) the Air Quality dataset (air quality and COVID-19 case features).

The aim is to determine the appropriate method and features for accurate prediction. LSTM and CNN multivariate models were developed to predict active COVID-19 cases for each area in the Hulu Langat district. The LSTM and CNN models will be compared regarding the prediction accuracy of the daily COVID-19 active cases in respective areas. In the study, the data set for December 2021 was used as a test set, while the other data set was used as a training set. Table 2 depicts the prediction result for LSTM and CNN methods on six districts in Hulu Langat for the three datasets. Figure 3 illustrates the predictive error of LSTM and CNN methods on three datasets for six areas in Hulu Langat. The six areas in Hulu Langat and number of cases are Ampang (1228 cases), Beranang (222 cases), Cheras (1292) cases, Hulu Langat (306 cases), Kajang (3013) cases, and Semenyih (803) cases.

PERFORMANCE OF LSTM AND CNN

This section analyzed the performances of the deep learning methods LSTM and CNN in predicting COVID-19 cases on three datasets. Table 2 and Figure 3 results indicate that CNN outperformed LSTM in most datasets for all areas. CNN gives lower MAE, RMSE, and MAPE measures and higher R2 than LSTM. CNN gives a slightly higher MAPE measure for the Beranang and Cheras (All Features) and the Beranang (Air Quality) datasets. CNN gives significantly higher R2 values than LSTM, indicating that the variance of the CNN model is closest to the actual value. In contrast, the low R2 value indicates that the two values are not closely related. Figure 4 visualizes the average MAE, RMSE, MAPE and R2 patterns for LSTM and CNN.

Figure 5 depicts the CNN prediction vs actual trends for datasets of six areas in Hulu Langat district in Selangor. The x-axis is the number of cases, and the y-axis represents 29 days (a month). The trend of CNN predicted values is consistent with the actual trends in most areas.

PERFORMANCE OF ALL-FEATURES, METEOROLOGY AND AIR QUALITY DATASET

We analyze the prediction performance of three datasets with different features. Overall, it can be seen in Table 2 and Figure 3 that in both LSTM and CNN models, except for MAPE error, the All-feature datasets give the best prediction models with the lowest MAE, RMSE and highest R2 values. For the MAPE, the meteorology dataset gives the lowest prediction error. The results indicate that combining COVID-19 cases, meteorology, and air quality features

TABLE 2. Performance result of LSTM and CNN prediction on three features dataset of six areas in Hulu Langat district

Area	LSTM matric evaluation				CNN matric evaluation			
	MAE	RMSE	MAPE	R2	MAE	RMSE	MAPE	R2
Ampang (*All Features*)	0.360	0.490	2.731	0.820	0.228	0.253	1.745	0.952
Ampang (*Air Quality*)	0.451	0.639	3.520	0.693	0.398	0.507	3.017	0.807
Ampang (*Meteorology*)	2.447	2.931	1.984	0.485	1.469	1.882	1.188	0.787
Beranang (*All Features*)	0.274	0.490	2.505	0.560	0.311	0.394	2.795	0.715
Beranang (*Air Quality*)	0.345	0.568	3.050	0.407	0.361	0.521	3.257	0.502
Beranang (*Meteorology*)	1.579	2.080	1.368	0.365	1.279	1.723	1.111	0.565
Cheras (*All Features*)	0.310	0.487	2.323	0.878	0.411	0.477	3.063	0.883
Cheras (*Air Quality*)	0.500	0.747	3.551	0.714	0.425	0.554	3.079	0.842
Cheras (*Meteorology*)	3.015	3.941	2.455	0.363	1.958	3.028	1.599	0.624
Hulu Langat (*All Features*)	0.829	1.125	6.343	0.594	0.604	0.744	4.862	0.823
Hulu Langat (*Air Quality*)	0.959	1.463	7.051	0.313	0.578	0.919	4.468	0.729
Hulu Langat (*Meteorology*)	3.859	5.115	3.144	0.330	2.374	3.292	1.952	0.722
Kajang (*All Features*)	1.311	1.896	6.735	0.739	0.621	0.770	3.246	0.957
Kajang (*Air Quality*)	1.638	2.527	8.110	0.536	1.049	1.835	5.231	0.756
Kajang (*Meteorology*)	5.666	8.876	3.997	0.543	4.884	6.305	3.374	0.770
Semenyih (*All Features*)	0.828	1.199	5.773	0.597	0.436	0.573	3.050	0.908
Semenyih (*Air Quality*)	0.587	1.231	5.838	0.575	0.832	1.060	5.660	0.685
Semenyih (*Meteorology*)	3.414	4.455	2.652	0.555	2.188	2.893	1.712	0.813

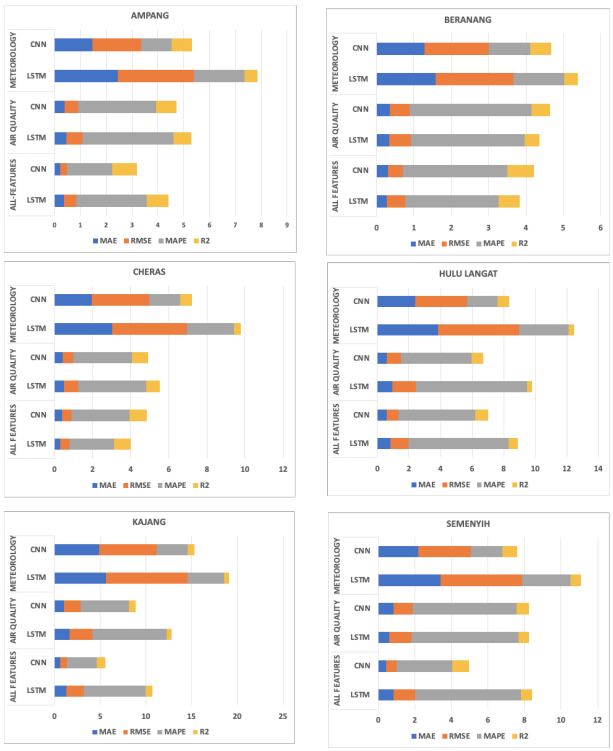


FIGURE 3. Predictive performance of LSTM and CNN for three features dataset in all district

(all-features dataset) can produce more accurate predictions. The Air Quality dataset gives competitive results to the all-features dataset with slight differences in MAE, RMSE, and R2 measures, indicating that the Air Quality data contributes most to the accurate COVID-19 cases prediction. The Meteorology dataset gives the lowest MAPE, suggesting no extremes in the meteorology data. Besides the higher MAPE obtained in the all-features dataset, we consider the MAE and RMSE errors as more significant measures shown by the all-features dataset. Thus, the best prediction model can be obtained by considering the combination of COVID-19 cases, air quality, and meteorology data. Despite the all-features model giving the lowest errors, the feature ranking process shows that meteorology factors are less important than air quality factors, which confirmed the results. Specific to CNN models, the average lowest prediction errors can be observed in Beranang, while the highest R2 values were in Ampang. The comparison within areas is not appropriate as the number of cases varies. For example, the Ampang and Beranang areas give lower prediction errors as the area has 1228 and 222 cases, respectively. Meanwhile, the Kajang area has 3013 cases affecting the prediction errors. However, the comparison will not be fair because of the different number of cases. Figure 6 depicts CNN models' average error and R2 values (by area).

In conclusion, the CNN all-features model can generally predict the daily activity of COVID-19 in the Hulu Langat district more accurately than other models. The dataset includes Air Quality and Meteorology features that potentially affect the COVID-19 cases. The CNN models on individual Air Quality and Meteorology data show that the air quality models lower errors via better prediction than the meteorology dataset.

Despite LSTM being well known to deal with time series data, in the case of Malaysia's meteorology and air quality data, CNN gives better results than LSTM. The reason could be characteristic of Malaysian climate and air quality data depending closely on its neighbors. CNN works well with the kind of data that the neighboring, and it is sufficient when there is no long dependency in the data. Information is supposedly relevant for the analysis of the data. CNN is suitable for forecasting time series because it offers dilated convolutions, in which filters can compute dilations between cells. The size of the space between each cell allows the neural network to understand better the relationships between the different observations in the time series (Madini, Mutasem & Reem 2022).

FEATURE IMPORTANCE

We performed feature selection to identify the rank of features that most contribute to the COVID-19 case prediction accuracy. A machine learning algorithm called Extra-Trees Classifier is employed to calculate the feature importance. Extra-Tree Classifier is based on ensemble learning, a variant of the Random Forest algorithm. The algorithm is implemented on three datasets to identify the critical features in each data set.

Based on Table 3, for the all-features dataset, it can be found that the top-rank features are the air quality features, such as Carbon Monoxide (CO), with an importance value of 0.116. The value shows that CO is one of the features contributing to high prediction accuracy. Following closely behind is Sulfur Dioxide (SO_2) in importance 0.108 and Nitrogen Dioxide (NO_2) with a significance of 0.104. Other features include PM_{10}, Ozone (O_3), Humidity, Wind Direction, Temperature, $PM_{2.5}$, and Wind Speed. These characteristics contribute to the overall understanding and analysis, although with slightly low significance values ranging from 0.100 to 0.090. The higher values indicate the relative importance of each feature in each context.

The top-rank feature in the Air Quality dataset is Carbon Monoxide (CO), with the highest importance value of 0.186 among the listed features. Next followed by Nitrogen Dioxide (NO_2) with an importance of 0.169 and PM_{10} with an importance of 0.167. Other important features, including Ozone (O_3), Sulfur Dioxide (SO_2), and $PM_{2.5}$, contributed significantly to the analysis, although slightly lower significance values ranged from 0.164 to 0.156. The result shows that air quality characteristics have a more substantial effect or influence on the COVID-19 case prediction. Table 3 shows that the top-rank feature for Meteorology dataset is Average Wind Direction with an importance value of 0.259, indicating that wind direction has the highest effect among other features. Next, the second most important feature is temperature, with an importance of 0.251. In addition, humidity and wind speed also contribute significantly to the analysis, with values of 0.249 and 0.241, respectively.

The result of the feature ranking of the individual dataset is evidence of the effect of specific parameters on the accurate prediction of COVID-19 cases. Based on the error performance in the all-features dataset and feature importance, the Air Quality data significantly affects the number of COVID-19 case predictions. The findings are consistent with several studies (Ali & Islam 2020; Ghobakhloo et al. 2022; Yu et al. 2021; Zhou 2022). In contrast, several other studies highlight that meteorology factors gives more effect to the COVID-19 cases than the air quality (Ogunjo, Fuwape & Rabiu 2022; Ramirez-Alcocer et al. 2022).

RESULTS DISCUSSION

The prediction results show that the CNN model has a higher prediction model accuracy with low MAE, RMSE, and MAPE values due to the spatial dataset characteristics. CNN model performs well in capturing spatial patterns and local dependencies. The COVID-19 data has a spatial

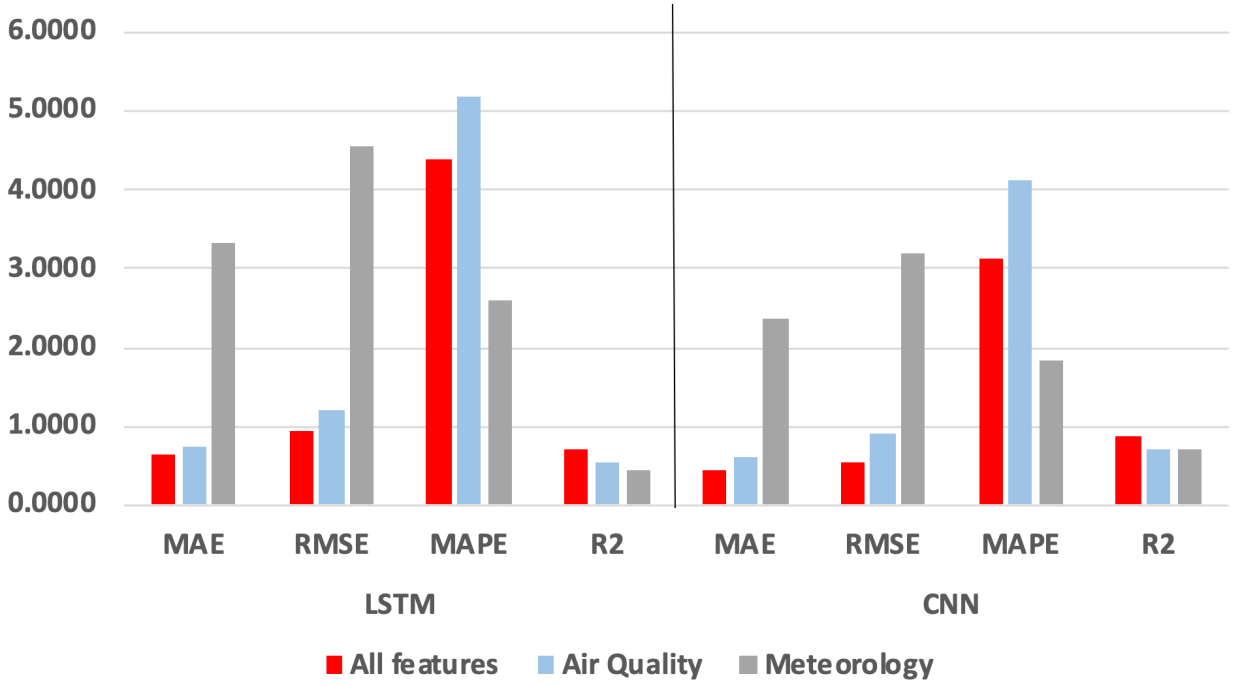


FIGURE 4. Average error values and R2 for six areas in Hulu Langat district (showed All-features dataset obtained the lowest prediction errors (MAE, RMSE) and highest R2)

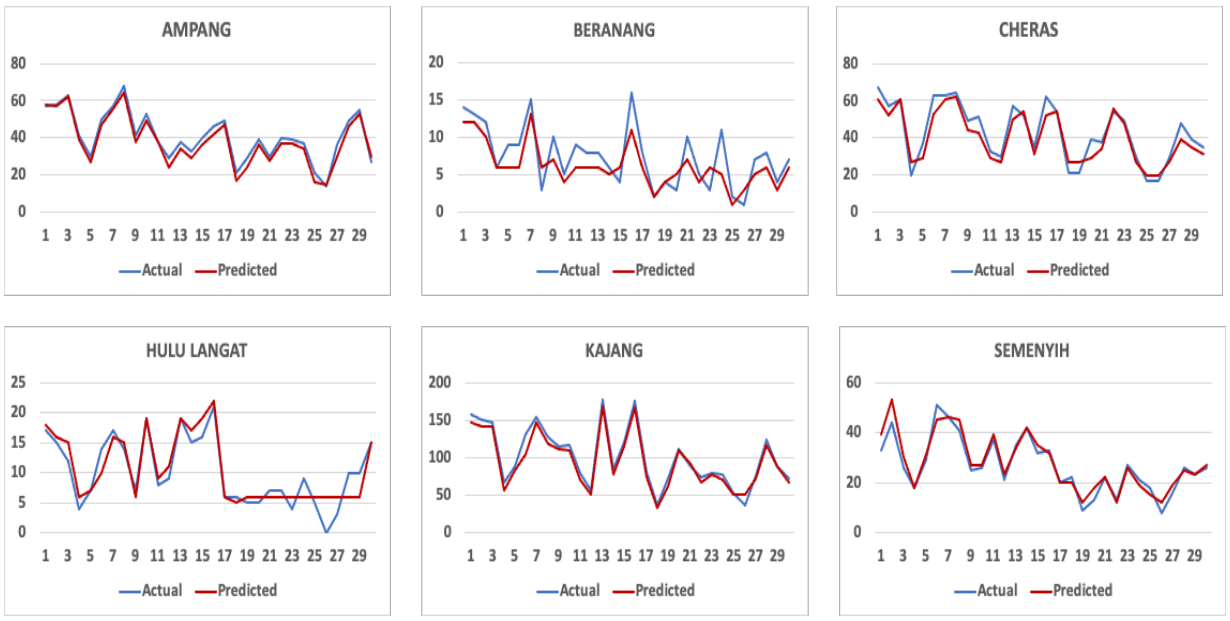


FIGURE 5. CNN predicted vs actual trends in six areas in Hulu Langat district

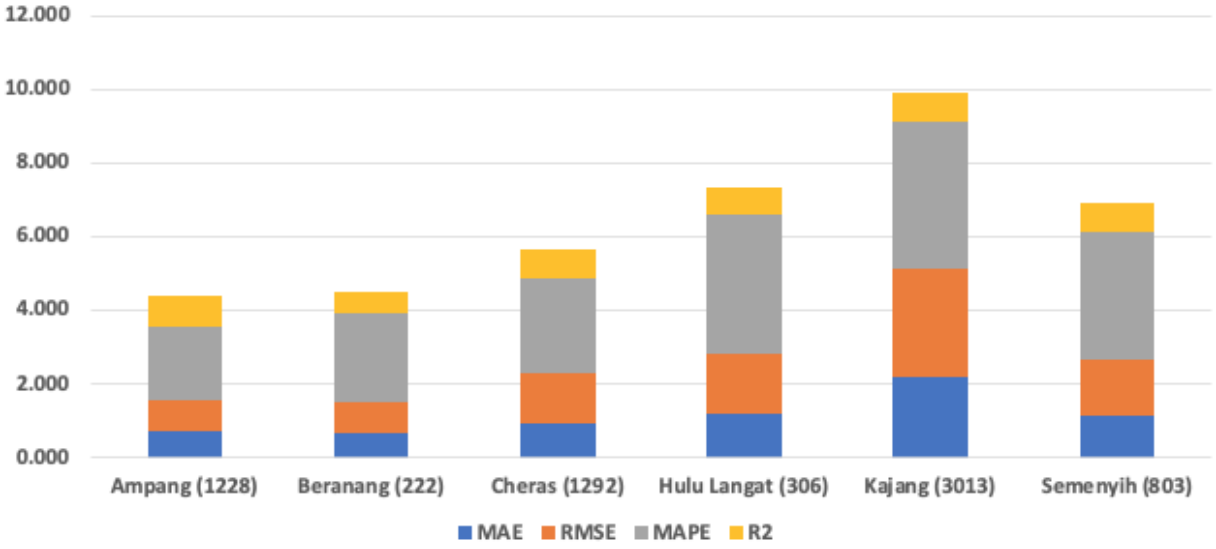


FIGURE 6. Average errors and R2 values (by areas) for CNN models
(The figures in the bracket are the number of cases)

TABLE 3. Feature importance of three datasets

All-features Dataset	Rank	Air Quality Dataset	Rank	Meteorology Dataset	Rank
CO	0.116	CO	0.186	Avg Wind Direction	0.259
SO_2	0.108	NO_2	0.169	Avg Ambient Temperature	0.251
NO_2	0.104	PM_{10}	0.167	Avg Relative Humidity	0.249
PM_{10}	0.101	O_3	0.164	Avg Wind Speed	0.241
O_3	0.100	SO_2	0.159		
Avg Relative Humidity	0.098	$PM_{2.5}$	0.156		
Avg Wind Direction	0.096				
Avg Ambient Temperature	0.095				
$PM_{2.5}$	0.092				
Avg Wind Speed	0.090				

component, such as regional or geographic information; thus, CNN models may better detect and use these patterns for prediction. In addition, CNN models can automatically extract meaningful features from data. The integrated COVID-19 dataset contains complex patterns or features that can be effectively learned and represented by the CNN model; the model will have an advantage in capturing these relevant features for prediction and producing accurate prediction results. Accordingly, CNN models often require large-scale datasets for practical training, and using multiple channels in the input layer, the CNN model can effectively handle multivariate data, such as meteorological variables and air quality. Each channel can represent a different variable and thus allows the model to process and analyze the relationship between these variables simultaneously. The ability to generalize well to observed data is critical in predicting COVID-19. Moreover, it allows the model to capture underlying patterns and make accurate predictions over time with limited data. Feature importance analysis showed outstanding prediction results in air quality data showing that air quality parameters such as CO, NO_2, PM_{10}, O_3, and SO_2 contribute to the COVID-19 case prediction, while meteorology parameters showed lower importance in prediction.

The impact of air quality factors on the spread of the COVID-19 pandemic compared to meteorological factors, especially in certain Asian climates such as Malaysia, can be significant due to several reasons. The poor air quality, prevalent in some areas due to pollution from industrial activities, vehicular emissions, and biomass burning,

can weaken the respiratory system. This compromised respiratory health can make individuals more susceptible to respiratory infections like COVID-19 and exacerbate the severity of the disease. COVID-19 primarily spreads through respiratory droplets, but evidence suggests that the virus can also spread through aerosols. Poor air quality can contribute to the persistence and spread of aerosols, potentially increasing the risk of transmission, especially in enclosed spaces with inadequate ventilation.

This study uses a case study from the Hulu Langat districts, with those six areas among the high-population areas in Malaysia. People living in areas with high air pollution levels may be at a higher risk of severe COVID-19 if infected, as pre-existing respiratory conditions are linked to worse outcomes with the virus. The environmental factors, often associated with poor air quality, such as high humidity or particulate matter in the air, might contribute to the survivability of the virus in the environment. Extended virus survival in such conditions could increase the likelihood of transmission. With the higher pollution levels and dense urban populations, the impact of air quality on the pandemic spread might be more pronounced compared to meteorological factors due to the population density and increased likelihood of exposure.

The findings of this study could benefit policymakers in developing a better systematic policy for Malaysia based on pollution sources and measures to improve air quality. Integrated efforts to control emissions and minimize exposure to air pollutants can improve human health and alleviate the public health burden of COVID-19. Besides that, strict enforcement of the law by the relevant agencies is very crucial, particularly by the Department of Environment (DOE) under environmental policies and laws, such as the Environmental Quality Act 1974, with subsidiary legislation such as the Malaysian Ambient Air Quality Standard 2013, Environmental Quality (Clean Air) Regulations 2014, and the like. For the public, self-control should be practiced even though the Ministry of Health has announced that there is no longer an obligation to wear a face mask and physical distancing. The public should be advised to avoid crowded places and wear a face mask; these measures need to be practiced to avoid contracting COVID-19 infection.

CONCLUSION

The paper presents the deep learning models for COVID-19 case prediction using integrated meteorology, air quality and COVID-19 case data in Hulu Langat Selangor. The deep learning model, CNN, and LSTM have shown that air quality factors have more impact on COVID-19 cases than meteorological factors in various areas. The results are consistent with several previous studies, while others reported that meteorology factors have more impact. Deep learning algorithms such as CNN can extract meaningful features from the spatial component characteristics of COVID-19, air quality and meteorology data. A large COVID-19 data set can ensure that CNN is trained well by producing more accurate prediction results while reducing the processing time needed to build a prediction model of the COVID-19 cases in the study. Lastly, CNN models have shown robust generalization capabilities for meteorological and air quality data analysis. While meteorological factors can influence virus transmission by affecting the persistence of the virus in the environment or human behaviour (like spending more time indoors during extreme weather conditions), in Malaysian climates, the impact of poor air quality on respiratory health and immune responses may play a more significant role in facilitating the spread and severity of COVID-19.

ACKNOWLEDGEMENTS

We acknowledged the Fundamental Research Grant Scheme grant number FRGS/1/2022/ICT02/UKM/02/7, funded by the Ministry of Higher Education (MOHE) Malaysia.

REFERENCES

Adhikari, A. & Yin, J. 2021. Lag effects of ozone, $PM_{2.5}$, and meteorological factors on COVID-19 new cases at the disease epicenter in queens, New York. *Atmosphere* 12(3): 357.

Ali, N. & Islam, F. 2020. The effects of air pollution on COVID-19 infection and mortality - A review on recent evidence. *Frontiers in Public Health* 8: 580057.

Felix Gers, FA., Schraudolph, NN., & Schmidhuber, J. 2002. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research*. 115-143

Ghobakhloo, S., Miranzadeh, M.B., Ghaffari, Y., Ghobakhloo, Z. & Mostafaii, G.R. 2022. Association between air pollution, climate change, and COVID-19 pandemic: A review of recent scientific evidence. *Health Scope* 11(4): e122412. https://doi.org/10.5812/jhealthscope-122412

Goodfellow, I., Bengio, Y. & Courville, A. 2016. *Deep Learning.* Massachusetts: MIT Press. http://www.deeplearningbook.org

Hasan, R.A. & Jamaludin, J.E. 2023. Prediction of COVID-19 cases for Malaysia, Egypt, and USA using deep learning models. *Malaysian Journal of Fundamental and Applied Sciences* 19: 417-428.

Jalaludin, J., Wan Mansor, W.N., Abidin, N.A., Suhaimi, N.F. & Chao, H-R. 2023. The impact of air quality and meteorology on COVID-19 cases at Kuala Lumpur and Selangor, Malaysia and prediction using machine learning. *Atmosphere* 14(6): 973. https://doi.org/10.3390/atmos14060973

Kelleher, J.D., Namee, B.M. & D'Arcy, A. 2020. *Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies*. 2nd ed. Massachusetts: MIT Press.

Khan, Z., Ualiyeva, D., Khan, A., Zaman, N., Sapkota, S., Khan, A., Ali, B. & Ghafoor, D. 2021. A correlation among the COVID-19 spread, particulate matters, and angiotensin-converting enzyme 2: A review. *Journal of Environmental and Public Health* 2021: 5524098.

Lloyd, B.N. & Viswanathan, P.M. 2022. A long term observation of meteorological influence on COVID-19 pandemic spread in Malaysia - A case study. *Journal of Climate Change* 8(1): 67-96.

Madini, O.A., Mutasem, J. & Reem, A. 2022. Time series predicting of COVID-19 based on deep learning. *Neurocomputing* 468: 335-344. https://doi.org/10.1016/j.neucom.2021.10.035

Mohan Viswanathan, P., Sabarathinam, C., Karuppannan, S. & Gopalakrishnan, G. 2022. Determination of vulnerable regions of SARS-CoV-2 in Malaysia using meteorology and air quality data. *Environment, Development and Sustainability* 24(6): 8856-8882. https://doi.org/10.1007/s10668-021-01719-z

Mohd Halim, N.F., Mohd Zahid, A.Z., Salleh, M.Z.M. & Abu Bakar, A.A. 2022. Air quality status during pandemic COVID 19 in urban and sub-urban area in Malaysia. *IOP Conf. Ser.: Earth Environ. Sci.* 1019: 012044.

Nielsen, M.A. 2015. *Neural Networks and Deep Learning.* Determination Press.

Ogunjo, S.T., Fuwape, I.A. & Rabiu, A.B. 2022. Predicting COVID-19 cases from atmospheric parameters using machine learning approach. *GeoHealth* 6(4): e2021GH000509. https://doi.org/10.1029/2021GH000509

Ramirez-Alcocer, U.M., Tello-Leal, E., Macías-Hernández, B.A. & Hernandez-Resendiz, J.D. 2022. Data-driven prediction of COVID-19 daily new cases through a hybrid approach of machine learning unsupervised and deep learning. *Atmosphere* 13(8): 1205. https://doi.org/10.3390/atmos13081205

Shen, N.W., Bakar, A.A. & Mohamad, H. 2023. Univariate and multivariate long short term memory (LSTM) model to predict COVID-19 cases in Malaysia using integrated meteorological data. *Malaysian Journal of Fundamental and Applied Sciences* 19: 653-667.

Valsamatzi-Panagiotou, A. & Penchovsky, R. 2022. Environmental factors influencing the transmission of the coronavirus 2019: A review. *Environmental Chemistry Letters* 20(3): 1603-1610.

WHO. 2022. WHO Coronavirus (COVID-19) Dashboard WHO Coronavirus (COVID-19) dashboard with vaccination data. https://covid19.who.int/ (Accessed on 4 December 2022).

Yu, Z., Abdel-Salam, A.S.G., Sohail, A. & Alam, F. 2021. Forecasting the impact of environmental stresses on the frequent waves of COVID19. *Nonlinear Dynamics* 106(2): 1509-1523.

Zhou, N., Dai, H., Zha, W. & Lv, Y. 2022. The impact of meteorological factors and $PM_{2.5}$ on COVID-19 transmission. *Epidemiology and Infection* 150: e38. https://doi.org/10.1017/S0950268821002570

*Corresponding author; email: azuraliza@ukm.edu.my