Prediction of Indoor Air Quality using Long Short-Term Memory with Adaptive Gated Recurrent Unit

Muhamad Sharifuddin Abd Rahim[1]*, Fitri Yakub [2], Mas Omar [1], Rasli Abd Ghani [2], Sheikh Ahmad Zaki Shaikh Salim [3], Shiro Masuda [4], Inge Dhamanti [5].

[1] WEE Laboratory, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia, (msharifuddin6@graduate.utm.my, masomar@graduate.utm.my)
[2] Department of Electronic System Engineering, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia, (mfitri.kl@utm.my, rasli.kl@utm.my)
[3] Department of Mechanical Precision Engineering, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia, (sheikh.kl@utm.my)
[4] Faculty of Systems Design and Graduate School of Systems Design, Tokyo Metropolitan University, Japan, (smasuda@tmu.ac.jp).
[5] Department of Health Policy and Administration, Faculty of Public Health, Universitas Airlangga, Surabaya, Indonesia, (inge-d@fkm.unair.ac.id)

Abstract. There is significant evidence that the COVID-19 virus may be spread by inhaling aerosols leading to risk of infections across indoor environments. Having said that, it is clear that the formulation of indoor air quality (IAQ) needs to be carefully examined. In general, IAQ can be controlled by proper ventilation system across buildings. Nevertheless, different buildings require different mechanistic approaches and it may not be an effective solution for the buildings. Thus, statistical approaches have great potential to evaluate the IAQ in real occupied buildings. Numerous machine learning (ML) techniques were introduced to forecast the indoor environmental risk across buildings. However, there is inadequate data available on how well these ML techniques perform in indoor environments. Recurrent neural network (RNN) is a ML technique that deals with sequential data or time series data. However, the RNN model gradient tends to explode and vanish, leading to inaccurate prediction outcomes. Therefore, this study presents the development of a time based prediction model, Long Short-Term Memory (LSTM) with adaptive gated recurrent units for the prediction of IAQ. Using an advanced LSTM model, the study focuses on the performance of the prediction accuracy and the loss during training and validation. Also, the developed model will be assessed with other RNN models for data validation and comparisons. A set of particulate matter (PM2.5) dataset from commercial buildings is assessed, preprocessed and clean to ensure quality prediction outcomes. This study demonstrates the performance of the hybrid LSTM model to remember past information, minimize gradient error and predict the future data precisely, ensuring a healthier indoor building environment.

Keywords. Indoor air quality, prediction, machine learning, Long Short-Term memory, hybrid

1 Introduction

The sick building syndrome describes a situation in which building occupants experience acute health and/or comfort effects that appear to be linked to time spent in a particular building, but where no specific illness or cause can be identified. The complaints may be localized in a particular room or zone, or may be spread throughout the building. Also, indoor air quality (IAQ) describes how inside air can affect a person's health, comfort, and ability to work. It can include but not limited to temperature, humidity, mould, bacteria, poor ventilation, or exposure to other chemicals. Indoor air pollution has received little attention in the past compared with air pollution in the outdoor environment. Nowadays, indoor air quality has becoming world-wide concern especially after the pandemic of COVID-19 where people mostly spend 90% of their time indoor. Therefore, IAQ is ranked as one of the top five environmental risks to global health and well-being [1].

Research in the field IAQ has a long tradition in the environmental engineering field. However, the determination to design such robust forecasting model is technically challenging, especially when dealing with non-linear data. Recently due to the pandemic of COVID-19, research have shown significant interest in modelling prediction model for IAQ monitoring or statistical model which has been referred as data driven

* Corresponding author: msharifuddin6@graduate.utm.my

model. The type of model used usually require the concept of input and output of the dataset without needing the mechanistic model of the buildings. The model is basically constructed based on the sequential data.

In this modern day, more robust recurrent neural network (RNN) has been introduced in order to solve the vanishing and exploding gradient of the conventional RNN. Before that, RNN processes to stop those events from occur are by having minimum clip at the gradient of RNN during the process of backpropagation. But, still this idea of solution did not able to memorize long term of the time series sequential data. This is where the Long Short-Term Memory (LSTM) comes in. The model is one of the variants of the RNN model. LSTM has solid abilities to memories short and long-term series of historical data. To date, there are still few of published research that deploy LSTM approach in the environmental engineering problem, especially indoor air quality. Therefore, it is necessary to use LSTM in order to predict short and long term of IAQ historical data. This study examines the ability of LSTM to enhance the data driven method in order to predict better accurate forecast and reduce loss during the deep learning training.

This paper organized as follows, in Section 2, discussed the LSTM network architecture, Gated Recurrent Unit (GRU) model and datasets of the IAQ. Further, the section includes the deep learning evaluation metrics to calculate the model performances. Results and discussions of the application of the LSTM and GRU model studies are presented in Section 3. Followed by Section 4, deduces the summary of the research. Last but not least, the challenges of the research and future direction of the research are addresses in this manuscript.

2 Methodology

2.1 LSTM Network Architecture

A vanilla and efficient model for IAQ prediction is introduced in this research. The IAQ parameters from a commercial building in Kuala Lumpur, United State (US) Embassy is used as an input dataset. The total available dataset was used for training and testing using LSTM and GRU prediction model. LSTM is a variant of RNN which deal with vanishing and gradient problem. The highlight feature of LSTM model is it has layer which is called memory cells. The model also consists of input layer and output layers. Each memory cell has three input gates which control and maintain the memory state (S_t). The gates are the forget gate (f_t), input gate (i_t) and output gate (o_t). The overview of the LSTM model architecture shown in Fig.1.

LSTM layer also consists of hidden state, which is known as the output state. At every timestep t, it contains the output of the LSTM. Other than hidden state, the cell state also part of the LSTM layer. The purpose of the cell state is to store the present information of time step t that the layer learned in the previous time steps. Every time step, cell state will decide either to keep or remove the information that state learned. The cell state controls and updates the information by using the three gates mentioned earlier. Also, at every time step, the information x_t is comes from the output of the previous time step, output h_t. Based on the references [2-5] the gates in the cell state have their own responsible. The purpose of the three gates is mentioned in the Table 1.

Table 1. Purpose of gates in LSTM cell state.

Gate	Role
Forget	Decides which information need to be removed from the current cell state
Input	Decides which information need to feed to the current cell state
Output	Decides which information will be used in the current cell state

Overall architecture of the LSTM model is based on the hyperparameters defined in the Table 2, to forecast the time steps by one step and update the new information at next memory cell.

Table. 2. Hyperparameter tuning

Hyperparameters	Setting
Input layer	50 nodes
Activation function	tanh
Optimizer	Stochastic gradient descent
Learning rate	0.01
Epochs	50
Batch size	32
Dropout rate	0.2

The network architecture is tuned according to the Tensorflow packages available. In order to train the proposed network with the IAQ parameters, the overall mechanism of LSTM is shown in the **Fig.1** according the [3]**.** The fully connected layers are then connected to the proposed model.

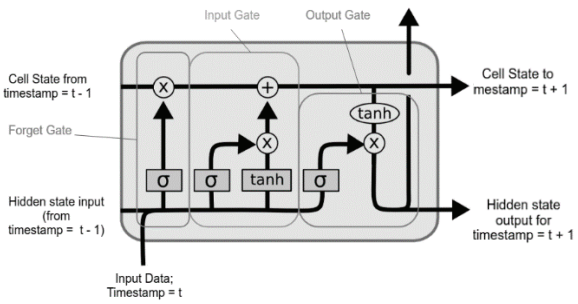

Fig. 1. The overall of LSTM cell state architecture.

2.2 Gated Recurrent Unit Model

After 17 years of the presence of vanilla and improved version of LSTM [3,5,6,18], another variant of RNN was developed. Gated Recurrent Unit (GRU) was introduced to simplify the computational power of LSTM while improving the output from the RNN model. At the same time, GRU preserves the LSTM performance while optimising the network layout. The GRU network structure, which can address the prediction issue of long interval long delay time series, only has two gate structures compared to the LSTM network structure's four. The update gate is used to regulate how much information from one moment is incorporated into the current moment. The reset gate is used to regulate how much of the information from the previous moment is ignored.

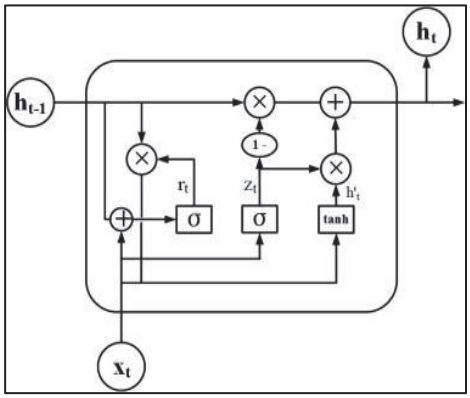

Fig. 2. GRU overall architecture.

The output of the reset gate r_t at time t. Meanwhile the output of the update gate at time t and h_t is z_t. h_{t-1} is the output at time t respectively. The input at time t is x_t and σ is the activation function at each gate. The computation for GRU is referred in equation (1).

$$r_t = \sigma(W_r \ [h_t\text{-}1, x_t \,]) \quad (1)$$

2.3 Hybrid LSTM with GRU model

The LSTM model has been proven in terms of their capacity to remember long term and short-term predictions. GRU on the other hand, has significant computational power due to the simplified gates architecture while preserving reliable performances. Combining these two models into one powerful model, will realize a powerful sequential forecasting performance, deep learning approach using the LSTM-GRU. This is where a hybrid LSTM with GRU comes in, to improvise the performance of the IAQ prediction model. Training data will be fed into the LSTM model first, then continue to be fed into the GRU model. Finally, the data will go through the output layer which contains the dense layer. The overall flow of the LSTM with GRU model is presented in the figure below.

In the LSTM with GRU model, the LSTM has two layers, each layer has 50 hidden units with tanh activation function. Then the training data will be fed into dropout layer before it is pass to the second layer of LSTM. Dropout layer is set at 20 percent so that it can regularize the output of data from the previous layer. The overall parameter tuning can be referred to Table 2. This process intents to prevent complex computation during the training process [8,9].

The GRU section also has two layers for training. Similar to the previous LSTM layer, it also contains 50 hidden units and a dropout layer for each GRU layer. Finally, the data will be passed to the dense layer for the deep learning neural network learning process. This is where the process of the learning and tuning of the overall model dimension occurs. The output of the training model will be validated and test data in order to evaluate the performance of the LSTM with GRU model.

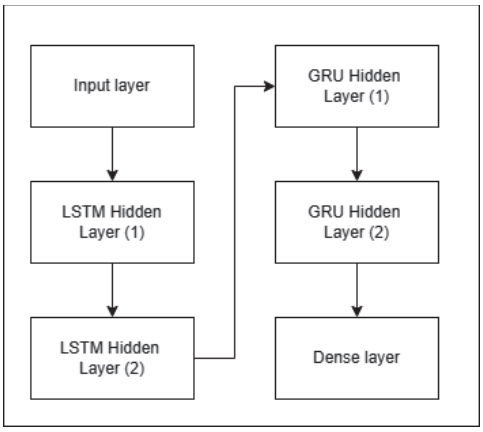

Fig. 3. LSTM with GRU model.

2.4 Prediction model evaluation metrics

Evaluation metric is one of the crucial stages of building a good prediction model. The purpose of an evaluation metric is to benchmark the desired prediction model efficiency. Based on the evaluation metric, one can deduce the performance of any prediction model [10,13]. Therefore, to demonstrate the performance of the IAQ prediction model in this research, root mean square error formulation in equation (2) is assessed.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y - y'|^2} \quad (2)$$

In the equation (2) the data point of the dataset is defined as n, the predicted PM2.5 is represented by y'. Meanwhile y is the observed PM2.5 at the time i, y is the mean of the actual of PM2.5 at the US embassy. In this research, Python Keras library and Tensorflow were used as the platform for the prediction model computation and development.

2.5 Indoor air quality data description and analysis

The dataset used in this research comprises of the database from the United State Embassy and Consulate building in Kuala Lumpur. The set of data collection consists of the particulate matter 2.5 micron (PM2.5). In general, PM2.5 which is referred as fine particles are one of the high risk to the deterioration of respiratory system [16]. Plus, PM 2.5 are harmful because of their size which allow them to pass to the lung and even blood circulation directly. Back then in 2008 in Beijing, US Embassy began air quality monitoring in their premises in order to make sure their citizen receive accurate information related to the air surrounding air quality [12].

In 2015 the United State Environment Protection Agencies (EPA) dealt and agreement to step up the scale of the PM2.5 monitoring at US embassies where the local PM2.5 data were not available [8]. Therefore, US Department of State is working together to monitor and control the indoor air qualities in U.S embassies around the world. In Kuala Lumpur, the consulates effectively links the information of the IAQ to the EPA AirNow International website [8]**.**

All the daily and hourly averaged PM2.5 from the website mentioned previously is scrapped and downloaded through June until November 2022. The website, AirNowTech checks the lowest, highest and the performance of the PM2.5 reading (Analysis of fine particle pollution data measured at 29 US diplomatic posts worldwide). The dataset also provides a column for the Air Quality Indices (AQI), indicating the air quality catogeries ("Good" and "Moderate"). The daily PM2.5 concentration in the dataset were calculated using formula provided [8]. It is stated that the between 18 to 24 hours should have valid data in the measurement. On the other hand, for hourly data measurement (1-h average), it uses the same measurement validation. And the formula calculated for 1-h average, it used the EPA NowCast theory [8]. The data consists of 3,236 of hourly sampling for the past five months. 2,589 data were examined as training and 647 data used as validation respectively.

3 Results and Discussions

The simulation of the IAQ prediction model was done using Keras, Tensorflow and Numpy packages in the Python library. In order to have a good prediction training process, data standardization is implemented to leverage the overall data into a simpler feature range. By having this process, one can transform the scale of the real-world data into a scale of zero to one. By having this method, the learning algorithm performs better when the actual numerical data is scale to range of value which have the most precision.

In this study, the input of the architecture was initialized using 2588 data sampling and it was simulated using 647 validation data using a set of time step (25 percent of the final time series), and it was continued by repeating the remaining network learning of the LSTM model and LSTM with GRU model. The simulation of the validation results are illustrated in **Fig. 4** for LSTM model and in **Fig. 5** for LSTM with GRU model with orange line colours.

Quantitatively, the results of the validations predictions are presented in the **Fig. 6** and **Fig. 7** with orange line colours. From **Fig. 7**, the LSTM with GRU model shows considerable precision in term of the average of the pattern flows.

The RMSE was computed in order to evaluate the performance of the actual PM2.5 concentration at the embassy building. These two assessed performances were computed from the actual PM2.5 value. The results presented in Table 3 show the LSTM model RMSE is 0.31862, while the LSTM with GRU model shows RMSE 0.20345. Also, the computational process time during both models training did not give any differences during the experimental simulation. In deduction, the LSTM with GRU model outperformed the LSTM model with lower RMSE value.

Table. 3. Evaluation metric

Model	LSTM	LSTM with GRU
RMSE	0.31862	0.20345

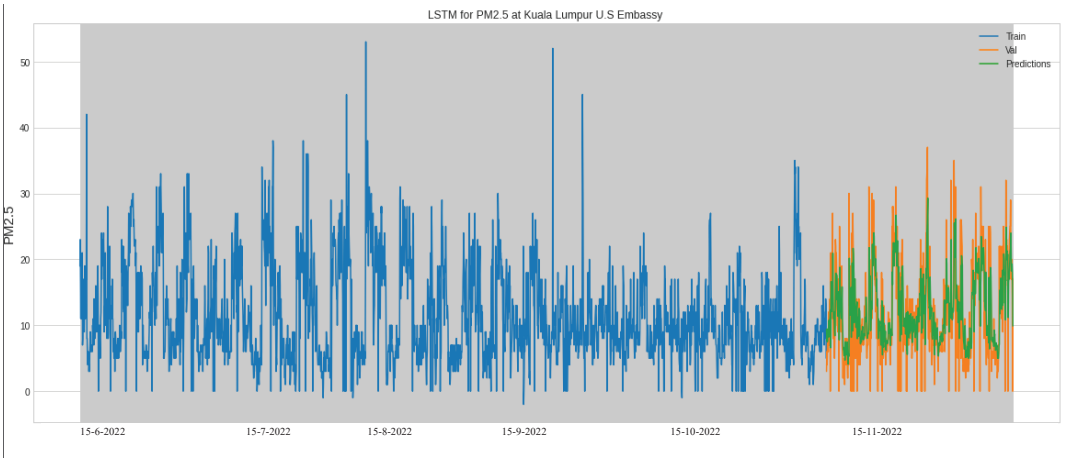

Fig. 4. LSTM model for PM2.5 concentration at Kuala Lumpur US Embassy building.

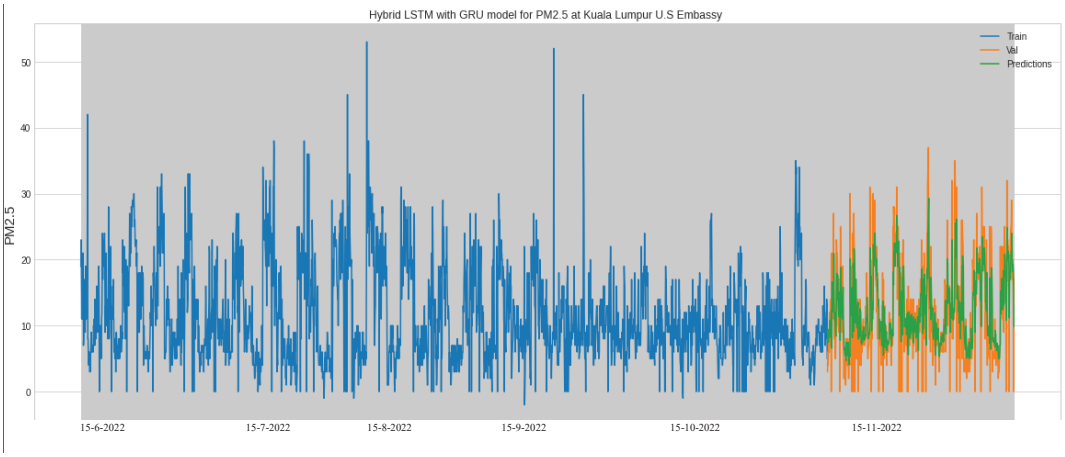

Fig. 5. LSTM with GRU model for PM2.5 concentration at Kuala Lumpur US Embassy building.

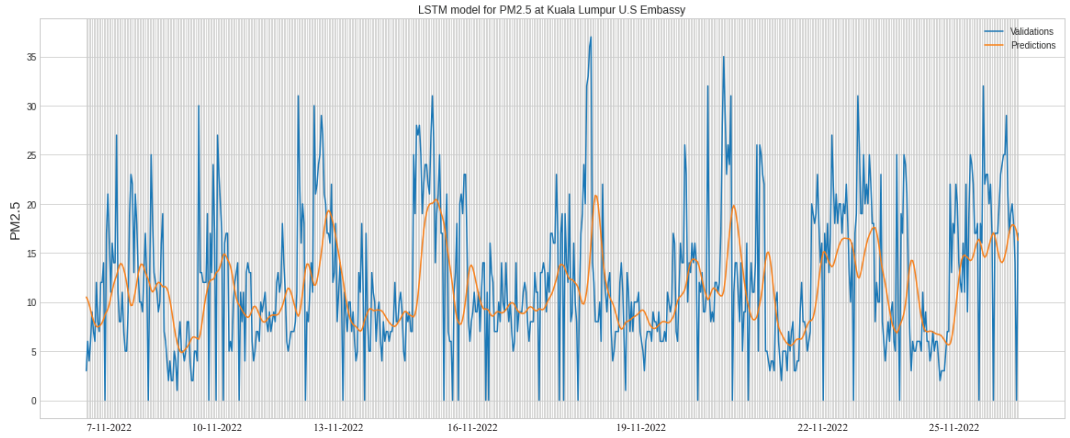

Fig. 6. LSTM validation and prediction results for PM2.5 concentration at Kuala Lumpur US Embassy building.

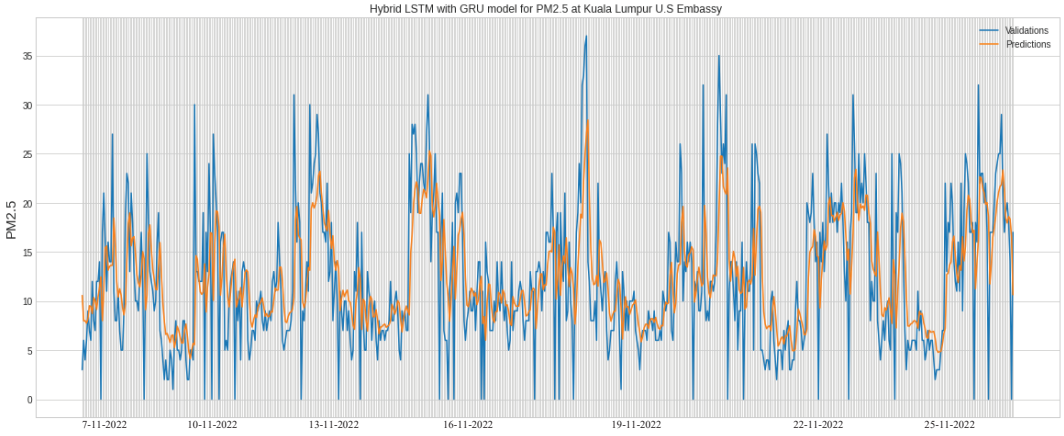

Fig. 7. LSTM with GRU validation and prediction results for PM2.5 concentration at Kuala Lumpur US Embassy building.

4 Conclusion

This paper was arranged to develop a hybrid and improvised version of LSTM for IAQ prediction using vanilla LSTM and hybrid LSTM with GRU model. The research has investigated the idea of RNN and its variances. The hybrid of LSTM with GRU not only outperforms the conventional LSTM, but also combines both RNN variants to improvise loss function of the time series prediction algorithm. With the evolution of ML architectures, the hybrid version of LSTM with GRU is potentially to be deployed in indoor space environmental problems.

For upcoming study, engaging multivariate input features can be included in order to find the relationship impact of poor IAQ based on the prediction performance. Furthermore, the escalation of the RNN variants in order to have promising performance with significant lower computation power will be more challenging in the process of development of building and environmental forecasting cases.

The project is under by the Ministry of Higher Education under FRGS, Registration Proposal No: FRGS/1/2022/TK07/UTM/02/52.

References

1. J. M. Seguel, R. Merrill, D. Seguel, and A. C. Campagna, “Indoor Air Quality,” American Journal of Lifestyle Medicine, vol. 11, no. 4, pp. 284–295, 2016.
2. C. Hu, Q. Wu, H. Li, S. Jian, N. Li, and Z. Lou, “Deep learning with a long short-term memory networks approach for rainfall-runoff simulation,” Water, vol. 10, no. 11, p. 1543, 2018.
3. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
4. A. Gensler, J. Henze, B. Sick, and N. Raabe, “Deep learning for solar power forecasting — an approach using AutoEncoder and LSTM Neural Networks,” 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016.
5. T. Fischer and C. Krauss, “Deep learning with long short-term memory networks for financial market predictions,” European Journal of Operational Research, vol. 270, no. 2, pp. 654–669, 2018.
6. A. Iosifidis, A. Tefas, and I. Pitas, “Dropelm: Fast neural network regularization with dropout and DropConnect,” Neurocomputing, vol. 162, pp. 57–66, 2015.
7. W. Opinion, “Opinion: How the US embassy tweeted to clear Beijing's air,” Wired, 06-Mar-2015. [Online]. Available: https://www.wired.com/2015/03/opinion-us-embassy-beijing-tweeted-clear-air/. [Accessed: 25-Nov-2022].
8. “System alerts,” AirNow.gov. [Online]. Available: https://www.airnow.gov/. [Accessed: 30-Nov-2022].
9. Du, Shengdong, Tianrui Li, Yan Yang, and Shi-Jinn Horng. "Deep air quality forecasting using hybrid deep learning framework." IEEE Transactions on Knowledge and Data Engineering 33, no. 6 (2019)
10. Athira, V., P. Geetha, Rab Vinayakumar, and K. P. Soman. "Deepairnet: Applying recurrent networks for air quality prediction." Procedia computer science 132 (2018)
11. Hossain, Emam, Mohd Arafath Uddin Shariff, Mohammad Shahadat Hossain, and Karl Andersson. "A novel deep learning approach to predict air quality index." In Proceedings of International Conference on Trends in Computational and Cognitive Engineering, pp. 367-381. Springer, Singapore, (2021)
12. Liao, Qi, Mingming Zhu, Lin Wu, Xiaole Pan, Xiao Tang, and Zifa Wang. "Deep learning for air quality forecasts: a review." *Current Pollution Reports* 6, no. 4 (2020)
13. Samal, K. Krishna Rani, Korra Sathya Babu, Abhirup Acharya, and Santos Kumar Das. "Long term forecasting of ambient air quality using deep learning approach." In 2020 IEEE 17th India Council International Conference (INDICON), pp. 1-6. IEEE, 2020.
14. R. Chuentawat and Y. Kan-Ngan, " The comparison of PM 2.5 forecasting methods in the form of multivariate and univariate time series based on support vector machine and genetic algorithm ", Proc. 15th Int. Conf. Electr. Eng./Electron. Comput. Telecommun. Inf. Technol. (ECTI-CON), pp. 572-575, Jul. 2018.
15. M. A. Elangasinghe, N. Singhal and K. N. Dirks, " Complex time series analysis of PM 10 and PM 2.5 for a coastal site using artificial neural network modelling and k-means clustering ", Atmos. Environ., vol. 94, pp. 106-116, Sep. 2014.
16. J. Xie, " Deep neural network for PM 2.5 pollution forecasting based on manifold learning ", Proc. Int. Conf. Sens. Diag. Prognostics Control (SDPC), pp. 236-240, Aug. 2017.
17. Y. Zhang, Y. He and J. Zhu, "Research on forecasting problem based on multiple linear regression model PM2.5", J. Anhui Sci. Technol. Univ., vol. 30, pp. 92-97, 2016.
18. K. R. Baker and K. M. Foley, " A nonlinear regression model estimating single source concentrations of primary and secondarily formed PM 2.5 ", Atmos. Environ., vol. 45, no. 22, pp. 3758-3767, 2011.