Enhancing Air Quality Prediction Accuracy Using Hybrid Deep Learning

Trang Pham Thi Quynh*, Tuyen Nguyen Viet, Hang Duong Thi, and Kha Hoang Manh

***Abstract*—PM2.5 (Particulate Matter) and PM10 are the most common pollutants, and the increasing of concentration in the air will threaten people's health. The machine learning method has recently been of particular interest to many researchers due to its effectiveness in air quality prediction models. Many solutions employing deep learning-based techniques such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and hybrid CNN-LSTM models to enhance air quality prediction accuracy have been developed. This paper proposes a hybrid Encoder STM model for predicting the next day to the next five days' PM2.5 and PM10 concentrations in Hanoi. Additionally, we proposed five extended features to increase the accuracy of prediction. Then other models, namely the LSTM model and the Bidirectional LSTM model, are also considered for PM2.5 and PM10 concentration prediction. Our results show that the proposed approaches outperform other state-of-the-art deep learning-based methods on both Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) due to low error and the small number of features.**

***Index Terms*—Urban air quality, PM2.5, PM10 prediction analysis, machine learning, hybrid deep learning**

I. Introduction

PM, which stands for particulate matter is also called particle pollution. PM is a complicated mixture of solids and aerosols, including tiny droplets of liquid, dry solid fragments, and solid cores with liquid coatings. Therefore, it is an intricate mixture of many chemical species instead of a single pollutant. They are called PM10 if their diameter is equal to *10* microns or less and PM2.5 if they are *2.5* microns or less in diameter. Many studies show air quality pollution's harmful effects on human health. According to [1], the increase in PM concentration may directly lead to elevated morbidity and even mortality. Microorganisms in PM2.5 and PM10 are suspected of causing allergies and spreading respiratory diseases [2]. Air pollution is a serious problem in many countries, especially in developing countries including Vietnam. Air pollution is one of the causes of heart disease, respiratory disease, cardiovascular disease, and lung disease in Vietnam [3]. Hanoi is the capital and the cultural and economic center of Vietnam. Air pollution increases with economic development and growth in the number of factories in Hanoi. The increase in PM concentrations and its human effects present an urgent problem that needs to be solved. Pollution assessment objectives may support policies to prevent air pollution. Therefore, predicting PM concentration is also one action plan to reduce and limit polluting activities. The benefit of forecasting helps local government give planning or measures to prevent emissions increases and allows people to make reasonable decisions about outdoor activities.

In recent years, artificial neural networks have been widely used in predicting pollutant concentrations, and the accuracy of prediction is increasing based on the improvement of models. In this paper, we developed AI-based daily mean PM2.5 and PM10 forecasting models to predict the next five days'' PM2.5 and PM10 concentrations in the upcoming day. We chose the last 15 days for input data and used traditional LSTM, Bidirectional LSTM, and encoder-decoder LSTM models to predict PM2.5 and PM10 concentration in Hanoi, Vietnam. The proposed model has a fast training time. The accuracy of prediction is evaluated through MAE, RMSE, and R^2 parameters. Those parameters are calculated in the test dataset.

The main contributions of this paper can be declared as follows:

- Make the expanded features from unprocessed data.
- Propose a model using the Bidirectional LSTM algorithm.

II. Related Works

There are many types of air quality prediction models in urban areas. They have included traditionally statistical models, numerical models, and machine learning methods. The traditional models used chemical transfer and atmospheric dispersion models [4].

A. Statistical Models

Statistical models used historical data for learning and then the prediction of future behavior of desired variables. The advantages of those models were high accuracy. However, the prediction accuracy reduces extremely if there is the dynamic behavior of meteorological parameters [5, 6].

B. Numerical Models

The researchers normally use mathematical equations to simulate the atmospheric process and predict the air quality using numerical methods. However, it is difficult to map the physics of pollutants based on simpler assumptions. These models were not appropriate for short-term predictions with exceptional cases of high variation in data [4].

C. Machine Learning Models

Today, artificial intelligence-based algorithms are being widely used for prediction. Unlike a purely statistical model, machine learning considers multiple parameters for

Manuscript received June 23, 2022; revised August 15, 2022; accepted September 7, 2022.

The authors are with Faculty of Electronics, Hanoi University of Industry, Hanoi, 100000, Vietnam. E-mail: nvtuyen@haui.edu.vn (T.N.V.), hangdt@haui.edu.vn (H.D.T.), khahoang@haui.edu.vn (K.H.M.)

*Correspondence: pham.trang@haui.edu.vn (T.P.T.Q.)

prediction, increasing the accuracy of the result. A recent study in Vietnam developed daily average PM2.5 forecasting models for HCM City, this daily PM2.5 forecasting used six machine learning algorithms and gave a conclusion that the Extra Trees Regression model gives the best forecast with statistical evaluation indicators including RMSE = 7.68 μg/m^3, MAE = 5.38 μg/m^3, R-squared = 0.68, and the confusion matrix accuracy of 74% [7]. The authors of [8] used a spatiotemporal Convolutional Neural Network (CNN) and Long-Short-Term Memory (LSTM) model to predict the next day's daily average PM2.5 concentrations in Beijing City using data collected over three years from January 1st, 2015 to December 31st, 2017. They showed that the predictive model using air quality data is more effective than that using meteorological data. The performance indexes of the proposed PM predictor in [8] include RMSE = 2.997, MAE = 2.21, and MAPE = 0.039. Rajnish and Quan *et al.* [3] analyzed and discussed the change in PM2.5 concentration over time at different locations in Ho Chi Minh City (HCMC), Vietnam. This study developed several deep learning-based one-shot multi-step PM2.5 forecasting models, an hourly forecast (1h to 24h), and a 24-hour rolling mean. The accuracy of models was evaluated by RMSE and MAE index and the best performance pertained to SGDRegressor with the lowest average RMSE of 3.38 μg/m^3 and MAE of 2.64 μg/m^3 [3]. According to Wang *et al.* [9], the authors improved neural networks using genetic algorithms. The neural network optimized by the genetic algorithm has better performance in PM2.5 mass concentration prediction, which increases the accuracy of prediction results and lessens the error rate [9]. To predict PM2.5 concentrations, the researchers in [10] think that features are significant for prediction in Tehran's urban area. The authors implement random forest, extreme gradient boosting, and Machine Learning (ML) algorithms for their study. They used 23 features, including satellite and meteorological data, ground-measured PM2.5, and geographical data, in their modeling. The best result pertaining to the XGBoost approach, incorporating the elimination of unimportant features with R^2 = 0.81, MAE = 9.93 μg/m^3, and RMSE = 13.58 μg/m^3 [10]. Xia and Yang *et al.* [11] developed a weighted long short-term memory neural network extended model (WLSTME) to predict the daily average PM2.5 concentrations. This study considers the effect of the density of sites and wind conditions on the spatiotemporal correlation of air pollution concentration when combined with multilayer models in deep learning. They mix MLP and LSTM methods to improve PM2.5 prediction accuracy with the lowest RMSE equaling 40.67 and the MAE equaling 26.10 [8]. The other study in Beijing chooses the last week's (7-day) air quality data as the input for forecasting the PM2.5 concentration of the next day [12]. This study combined the convolutional neural network (CNN) with the long short-term memory (LSTM) neural network for forecasting (called the hybrid CNN-LSTM model). To evaluate the accuracy of those models via mean absolute error (MAE = 6.779) and root mean square error (RMSE = 8.119). The final results of [12] show that the proposed model improves the accuracy of prediction and reduces the training time.

To improve the accuracy of prediction we aim to the importance of features like [10] but we did not use the meteorological data, we made the expanded features from unprocessed data. The result shows that the MAE and RMSE indexes are considerably reduced.

III. Data and Method

The procedures in this study consist of six steps followed as shown in Fig. 1. The models were simulated by Python programming scripts. The details of each step are explained below:

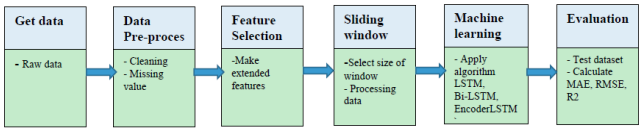

Fig. 1. Process for estimating PM2.5 and PM10 values daily.

A. Data Collection and Preprocessing

The datasets represent different environmental conditions related to air pollutant concentration. Six pollutants have been collected from several monitoring stations in Hanoi, the capital of Vietnam. The data used as predictors to perform the analysis involve PM2.5, PM10, NO_2, O_3, SO_2, and CO, collected from January 1, 2018 to May 10, 2022.

The number of raw data points of the Hanoi monitoring stations is 1598 rows that were saved by the CSV file. Each of these databases is divided into two databases, one with 1500 first rows for training and the last 98 rows for the test. Before the learning phase, data preprocessing is operated. In the raw data file, there are some missing values or invalid variables, these values are treated by imputation to recover the corresponding values. The imputation process uses the nearest data field to substitute the rest of the invalid or missing data if the percentage of missing values is lower than 16% for a row or 1% for a column in station datasets.

B. Feature Selection

In Vietnam, assessing pollution levels is based on parameters like PM2.5, PM10, NO_2, O_3, SO_2, and CO [3]. In this study, we forecast the values of PM2.5 and PM10 based on the previous 15 days' data. To increase forecasting performance, in addition to the raw parameters from the collected data set we add the extended features to those data sets. Because PM may be directly emitted from sources formed in the atmosphere through the chemical reaction of gases such as SO_2, NO_x, and certain organic compounds, furthermore pollutants interact with each other, we propose an extended feature that includes the following: total measured values of pollutants on a day; average values of PM10 and PM2.5 on a day; and the average of NO_2, CO, SO_2, and O_3 on a day. The extended features are shown in Table I.

TABLE I: Description of Extended Features

No	Feature	Type	Description
1	S_Data	Numeric	Total of valuable data on day (sum of PM2.5, PM10, NO_2, O_3, SO_2, and CO values)
2	A_PM	Numeric	Median of PM2.5 and PM10 values

3	A_O	Numeric	Median of NO_2, O_3, SO_2, and CO values
4	S_PM10	Numeric	Total of PM10 values in 15 days earlier (Sliding window size = 15)
5	A_PM10	Numeric	Average of PM10 values 15 days earlier

Table II shows the output of models, which include PM2.5 and PM10 concentration in the next *n* days' (*n* = 1,2,3,4,5).

TABLE II: DESCRIPTS THE TARGET VARIABLES

No	Feature	Type	Description
1	PM2.5_1/ PM10_1	Numeric	PM2.5, PM10 values for the next day
2	PM2.5_2/ PM10_2	Numeric	PM2.5, PM10 values for the next 2 days
3	PM2.5_3/ PM10_3	Numeric	PM2.5, PM10 values for the next 3 days
4	PM2.5_4/ PM10_4	Numeric	PM2.5, PM10 values for the next 4 days
5	PM2.5_5/ PM10_5	Numeric	PM2.5, PM10 values for the next 5 days

C. Sliding Window

The sliding window method uses the previous time steps to predict the next time step. We restructure the input data with arbitrary window size using the sliding window method. The value for time step for LSTM models is a very important hyper-parameter. Within every sliding window, assume the time-step is t. It means the LSTM has learned from t time-step and has attempted to predict the next t time step in the future. Next, the sliding window slides a one-time step to the right, and the whole procedure restarts. In the multi-input models, if the time step is large then the input data for learning is large too. In the multi-step models, to ensure accuracy the size of the window is always a smaller time step value. we predict the PM value for the next five days so the time step must be greater than 5. In this study, we chose the window size as 15, which means, using the value at the previous time step (in this study, we used data from 15 days ago), to predict the deal at the next time step. After restructuring, the data looks like a supervised learning dataset, so that any machine learning algorithm can use to model time series [13]. The window size in the study was selected experimentally.

If there are m input time series in the data set called X and output called Y , then

$$X = \{x_t^i\}\ \mathrm{t} \in \mathrm{T}=\{1,2,...,n\} \text{ and } i = 1,2,...,m$$
$$Y = \{y_t\}\ \mathrm{t} \in \mathrm{T} \quad (1)$$

n: is the size of the window.

T: is the time recorded in the dataset.

m: is the number of fields in the dataset [13].

The sliding window transformation process creates the following dataset:

$$X = \{\{x_{t-1}^1, x_{t-2}^1,..., x_{t-\mathrm{w}}^1\}, \{x_{t-1}^2, x_{t-2}^2,..., x_{t-\mathrm{w}}^2\}..., x_{t-1}^m, x_{t-2}^m,..., x_{t-\mathrm{w}}^m\}, \{\mathrm{y}_{t-1}, y_{t-2},..., \mathrm{y}_{t-\mathrm{w}}\}, y_t\} \text{ with } \mathrm{t} \in \mathrm{T} \quad (2)$$

D. Performance Evaluation

Like the research [14], to evaluate the performance of models, we calculate the parameters RMSE (root mean squared error) and MAE (mean average error). Such parameters were calculated based on the difference between the prediction result and the actual value. R^2 (R-squared) is needed to explain the strength of the relationship between predictive models and target variables [15]. The mathematical expressions of the metrics are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - avg(\hat{y}_i)\right)^2} \quad (5)$$

where $\hat{y}_i$ is the i^{th} predicted value, y_i is the i^{th} observed value [15].

N is the number of records in the test dataset [15].

E. Model Building

The following steps, after pre-processing the data, can be prepared for training and testing. In deep learning methods, processing to build a model is made by experiments. In this paper, we aim to use the multi-step LSTM model and variants of the LSTM model like the Bidirectional LSTM and Encoder-LSTM. Finding the best hyper-parameter combination for each model and transformed training database is the first step in the model training process [13, 16]. Hyper-parameters in our models include epochs, number of steps in, and number of steps out. When hyper-parameters have been found, we add different hidden layers with varying algorithms of optimization. The number of hidden layers and optimization algorithms were found by testing. Table III shows the hyper-parameter and optimization algorithms used for this study, and the proposed model used in this paper is shown in Fig. 2.

TABLE III: HYPER-PARAMETER AND OPTIMIZATION ALGORITHMS

No	Model	Hyper-parameter	Optimization algorithm
1	LSTM	epochs=500, verbose=0, number of steps in =15, number of steps out = 1/2/3/4/5	LSTM layer: activation='relu' Output layer: activation='linear'; optimizer='adam', loss='mse'
2	Bidirectional LSTM	epochs=1000, verbose=0, number of steps in =15, number of steps out = 1/2/3/4/5	LSTM layer: activation = 'relu' Output layer: activation='linear'; optimizer='adam', loss='mse'
3	Encoder LSTM	epochs=1000, verbose=0, number of steps in =15, number of steps out = 1/2/3/4/5	LSTM layer: activation ='relu' Output layer: activation='linear'; optimizer='adam', loss='mse'

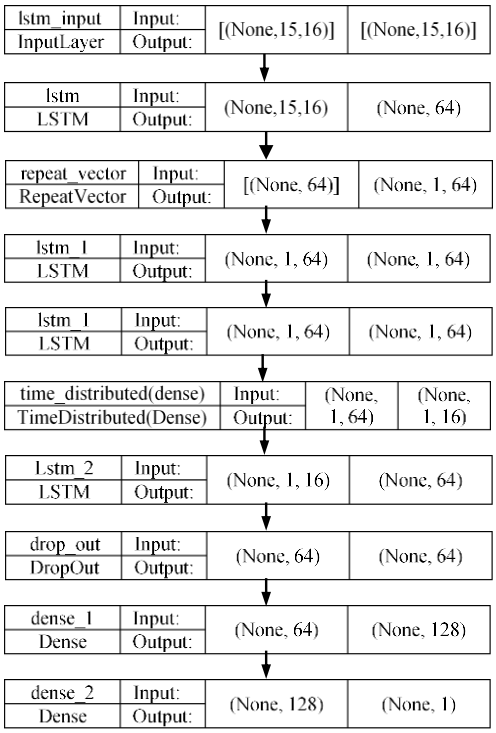

Fig. 2. Proposed model used for predict PM2.5 and PM10.

IV. Results and Discussion

The models have been trained on Lenovo computers with X270 8GB RAM using Python 3.9 IDE and the training takes 404 seconds. Fig. 3 describes the true PM2.5 and PM10 values and the predicted PM2.5 and PM10 values from February to May 2022 in Hanoi. The color expresses the level of health concern: green is good, yellow means moderate, and orange or red means unhealthy [17]. Fig. 4 shows the trend of real value and predicted value according to our proposed model of the PM2.5 index in Hanoi in February, March, and May 2022. In this, we used the Encoder LSTM algorithm with a sliding window size of 15. The results show that our model gives good trend prediction results.

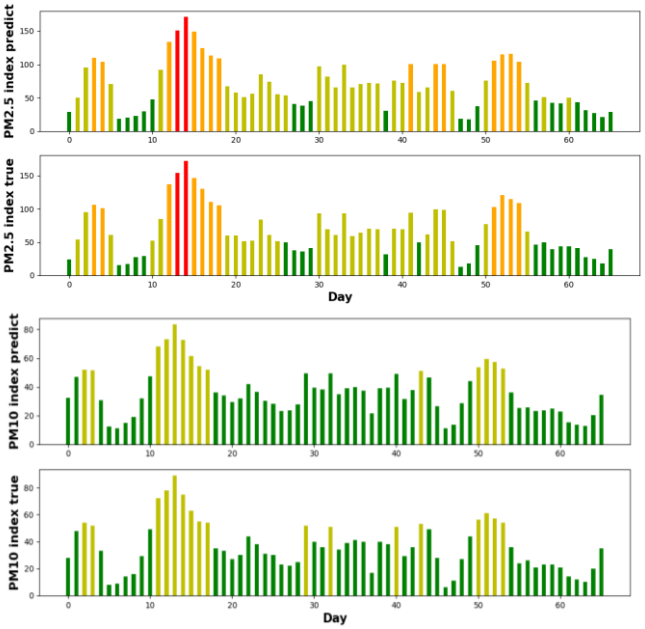

Fig. 3. The true PM2.5, and PM10 values and predicted PM2.5, and PM10 values from February to May 2022 in Hanoi.

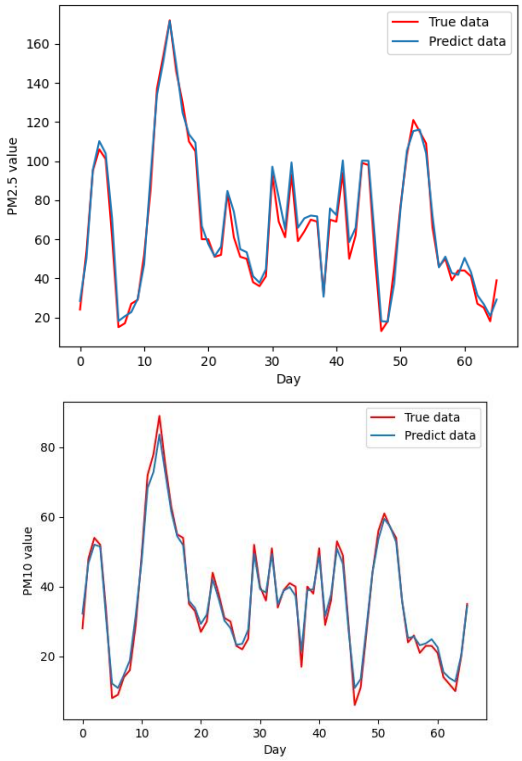

Fig. 4. The trend of PM2.5, PM10 real value and predict value.

Table IV describes the evaluation results of PM2.5 for 1 to 5 days ahead predictions in Hanoi using three methods. The result of predictions is based on data from the previous 15 days. It can be realized that machine learning algorithms performed very well in predicting the PM2.5 index for the following day in Hanoi. The results were shown to be good with R^2 of 0.964 when predicting one day ahead (step ahead = 1) using the LSTM method. The exactness of a model decreases if it increases the number of steps predicted. The precision of forecasting for the next 5th day with R^2 only equals 0.24. The encoder LSTM algorithm is the best with R^2 equal to 0.991 among three algorithms for performance.

The results in Table V show the performance of forecasting PM10. The same as above the best result belongs to the Encoder LSTM method with mean absolute error (*MAE*) equal to 1.850, Root mean square error RMSE = 2.209, and R^2 = 0.985.

TABLE IV: Performance and Hyper-Parameters of PM2.5 Using the LTSM Method

Method	The step-ahead	MAE	RMSE	R^2
LSTM	1 day ahead	5.391	6.659	0.964
	2 days ahead	12.261	17.414	0.764
	3 days ahead	19.724	23.757	0.563
	4 days ahead	23.364	29.806	0.314
	5 days ahead	25.143	31.264	0.243
Bidirectional LSTM	1 day ahead	4.762	6.419	0.968
	2 days ahead	15.611	19.358	0.708
	3 days ahead	17.989	22.627	0.604
	4 days ahead	19.994	27.433	0.418
	5 days ahead	22.997	28.625	0.366
Encoder LSTM	1 day ahead	2.77	3.348	0.991
	2 days ahead	12.438	17.146	0.771
	3 days ahead	18.089	23.410	0.576
	4 days ahead	21.981	29.007	0.351
	5 days ahead	26.313	32.682	0.175

TABLE V: PERFORMANCE OF PM10 PREDICTION

Method	The step-ahead	MAE	RMSE	R^2
LSTM	1 day ahead	4.111	5.091	0.919
	2 days ahead	6.263	8.001	0.803
	3 days ahead	7.055	8.799	0.763
	4 days ahead	8.199	10.232	0.679
	5 days ahead	8.249	10.416	0.667
Bidirectional LSTM	1 day ahead	2.963	3.620	0.959
	2 days ahead	4.785	5.905	0.893
	3 days ahead	5.635	6.938	0.853
	4 days ahead	6.863	8.450	0.773
	5 days ahead	7.052	8.850	0.759
Encoder LSTM	1 day ahead	1.850	2.209	0.985
	2 days ahead	2.866	3.669	0.959
	3 days ahead	6.172	7.874	0.811
	4 days ahead	8.093	10.466	0.663
	5 days ahead	9.390	12.151	0.545

Several considerations can be seen when analyzing our results. Firstly, how good algorithms that we have used can fit the past pollution data. Table IV and Table V show that the encoder LSTM is better than the basic LSTM and the Bidirectional LSTM algorithm. Secondly, the addition of extended features also results in better efficiency. When predicting PM2.5, we used the extended features in Table I, whereas when predicting PM10, we set the opened features as in Table I. The result shown in Table IV and Table V indicates that the *MAE* and *RMSE* indexes reduce and the R^2 index increases if there are more extended features

V. CONCLUSIONS

In this paper, the deep learning model for predicting air quality based on PM2.5 and PM10 indexes has been proposed. The effectiveness of the proposed approach has been verified through the following aspects: the air prediction accuracy with tiny *MAE*, the training time is short and the model is not complicated. We have used *MAE*, *RMSE*, and R^2 to evaluate the goodness of a prediction technique and concluded that the deep-learning approach, and in Encoder LSTM, with windows from 10 to 15 days, allow for a very reliable 1-day ahead prediction. Our results include prediction with correlation indexes in many cases greater than 0.95 with data that were collected years ago in Hanoi, Vietnam. Moreover, simulation results prove that the proposed approach outperforms the other state-of-the-art methods in terms of prediction accuracy with the small number of features and the training time of the proposed method is much faster than the others. Due to the above reasons, the proposed method can be used in real-time air prediction applications.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Tuyen Nguyen Viet, Kha Manh Hoang proposed an idea, Trang Pham Thi Quynh, and Hang Duong Thi contributed to the simulation and analyzed the data. All authors had written the paper and approved the final version.

FUNDING

This work was supported by the Hanoi University of Industry (HaUI) under Grant No. 23-2021-RD/HĐ-ĐHCN.

REFERENCES

[1] Y. F. Xing, Y. H. Xu, M. H. Shi, *et al*., "The impact of PM2.5 on the human respiratory system," *Journal of Thoracic Disease*, vol. 8, no. 1, E69–E74, 2016.
[2] C. Cao, W. Jiang, and B. Wang *et al*., "Inhalable microorganisms in Beijing's PM2.5 and PM10 pollutants during a severe smog event," *Environ Sci Technol*, vol. 48, no. 3, pp. 1499–1507, 2014.
[3] R. Rajnish, L. Quan, H. B. Quoc, V. Khue, and C. R. Simon. (2022). Ai based air quality PM2.5 forecasting models for developing countries: A case study of Ho Chi Minh city, Vietnam. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4111434
[4] C. Srivastava and S. Singh, "Estimation of air pollution in delhi using machine learning techniques," in *Proc. International Conference on Computing, Power and Communication Technologies*, 2018.
[5] Li, N. Hsu and S. T say, "A study on the potential applications of satellite data in air quality monitoring and forecasting," *Atmos. Environ.*, vol. 45, no. 22, pp. 3663-3675, 2011.
[6] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, Hoboken: Wiley S. Pro., 1970.
[7] V. T. T. Minh, T. T. Tin, and T. T. Hien, "PM2.5 forecast system by using machine learning and WRF model, A case study: Ho Chi Minh City, Vietnam," *Aerosol and Air Quality Research*, vol. 21, no. 12, 210108.
[8] U. Pak, J. Ma, U. Ryu, *et al.*, "Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China," *Science of The Total Environment*, vol. 699, 133561, 2020.
[9] X. Wang and B. Wang, "Research on prediction of environmental aerosol and PM2.5 based on artificial neural network," *Neural Computing and Applications*, vol. 31, pp. 8217–8227, 2019.
[10] J. M. Zamani, C. Cao, X. Ni, *et al.*, "PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data," *Atmosphere*, vol. 10, no. 7, 373, 2019.
[11] F. Xiao, M. Yang, H. Fan, G. Fan, and M. A. Al-Qaness, "An improved deep learning model for predicting daily PM2.5 concentration," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.
[12] T. Li, M. Hua, and X. U. Wu, "A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5)," *IEEE Access*, vol. 8, pp. 26933-26940, 2020
[13] R. Espinosa, J. Palma, F. Jiménez, J. Kamińska, G. Sciavicco, and E. Lucena-Sánchez, "A time series forecasting based multi-criteria methodology for air quality prediction," *Applied Soft Computing*, vol. 113, 107850, 2021.
[14] Y.-C. Liang, Y. Maimury, H.-L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," *Appl. Sci*., vol. 10, 9151, 2020.
[15] J. M. Dufour, *Coefficients of Determination*, McGill University: Québec, QC, Canada, 2011.
[16] J. Brownlee, "Deep learning for time series forecasting: Predict the future with MLPs, CNNs and LSTMs in Python," *Machine Learning Mastery*, 2018.
[17] N. M. Wambebe and X. Duan, "Air quality levels and health risk assessment of particulate matters in Abuja Municipal Area, Nigeria," *Atmosphere*, vol. 11, 817, 2020.