Science on Natural Resources and Environment
Journal homepage: tapchikhtnmt.hunre.edu.vn

AIR QUALITY PREDICTION IN HANOI USING A DEEP LEARNING APPROACH

Dang Thi Khanh Linh[*], Vu Van Huan
Hanoi University of Natural Resources and Environment, Vietnam

Received 12 November 2024; Accepted 23 December 2024

Abstract

Air pollution is becoming a serious global crisis, threatening human health, disrupting the balance of the environment, negatively affecting ecosystems, and contributing to climate change. Accurate long-term air quality prediction plays a key role in building early warning systems to mitigate these negative impacts. Efforts to forecast air quality through the combination of knowledge from environmental science, statistics, and computer science have attracted much attention. Among them, deep learning and advanced machine learning have demonstrated an outstanding ability to detect complex non-linear patterns from environmental data. However, the application of deep learning to air quality prediction is still quite new. This paper proposes a deep-learning model using the LSTM (Long Short-Term Memory) network to predict air quality in Hanoi. The research results demonstrate that the proposed model is capable of predicting the air quality index with high accuracy, close to actual values from monitoring data.

Keywords: Air quality index; Deep learning; LSTM.

[*]**Corresponding author, Email:** dtklinh@hunre.edu.vn

DOI: http://doi.org/10.63064/khtnmt.2024.650

1. Introduction

Air pollution is the biggest environmental threat, mainly due to toxic emissions from rapid industrialization and population growth. Air quality has deteriorated significantly, seriously affecting human health. The AQI index is used to measure pollution levels, based on 12 parameters such as PM2.5, PM10, O_3, NO_2, SO_2, CO, NH_3, Pb, Benzene, Toluene, and Arsenic. Typically, six major substances (PM10, PM2.5, SO_2, CO, and O_3) are used to calculate the AQI, depending on the target, data, and monitoring technique. A high AQI indicates severe pollution with a high health risk. AQI data is monitored in real-time and collected daily at many weather stations [1].

Accurately predicting air quality is a significant challenge due to the spatial and temporal heterogeneity of data from sources such as ground monitoring stations, satellites, and aircraft. Ground monitoring stations are particularly important as they are strategically located near highly polluted areas, providing valuable data for analyzing air quality fluctuations. However, many regions still lack comprehensive monitoring systems, leading to sparse and limited data. Additionally, the complexity of chemical processes involved in air pollution, combined with their spatial and temporal variability, makes it challenging to rely on fixed formulas for predictions. As a result, achieving accurate air quality predictions remains a demanding task [2].

Today, two main methods have been used to predict air quality: traditional methods based on hand-crafted features and machine learning algorithms based on deep learning. Traditional methods include basic numerical, statistical, and machine-learning models. Numerical models are based on atmospheric dynamics and chemical processes, using meteorological data to simulate pollution. Meanwhile, statistical models focus on integrating meteorological data with historical air quality data but have limitations in handling nonlinear problems [3]. To overcome this, researchers have applied nonlinear machine learning such as SVR and RFR, which have better performance. However, these methods still have difficulty capturing complex time series patterns and modeling the long-term impact of pollution, leading to reduced prediction accuracy [4, 5, 6].

In recent years, deep learning techniques have emerged and attracted considerable attention in various fields. These methods are effective in recognizing complex patterns and extracting high-level features, achieving great success in fields such as computer vision, speech processing, and natural language processing. However, there is still a lack of systematic reviews on the application of deep learning to air quality prediction [2].

Recently, Artificial Intelligence (AI) algorithms have been widely applied in air quality prediction. Thanks to the advancement of computational technology and algorithms, deep learning models have been deployed to analyze and predict non-linear relationships between data variables. These models have demonstrated outstanding effectiveness in improving the accuracy and reliability of data analysis results [7]. The Random Forest Regression (RFR) and Support Vector Regression (SVR) algorithms were applied to build a regression model to predict the air quality index (AQI) in the study of Liu H et al., [8]. Data related to nitrogen oxide concentrations in a city in Italy, obtained from a public data source, were analyzed. The performance of the models was evaluated based on the correlation coefficient, Root Mean Square Error (RMSE), and coefficient of determination R2. The results show that both methods provide superior prediction performance.

Similarly, Wu Q et al., proposed a hybrid and optimal method that combines secondary analysis (SD) with an AI optimization algorithm [9]. AQI data collected from China during 2016 - 2018 were used to validate the prediction model. Wavelet analysis was performed on high-frequency, low-frequency, and variable modes, and sample entropy was used to smooth the analysis. Then, a Neural Network (NN) based on Long Short-Term Memory (LSTM) was deployed to predict AQI [10].

In addition, the parameters of the least squares support vector machine (LS - SVM) were optimized using the BAT algorithm, taking into account factors related to air pollutants. The results show that this method effectively models the characteristics of AQI data, contributing to improved prediction ability.

Air pollution data, a typical form of time series data, has the potential to be improved through the integration of dynamic data [11], which is an important direction in real-time forecasting. Current deep learning models often have many parameters, resulting in high computational complexity. However, due to the time-sensitive air quality prediction requirements, the models must ensure fast and accurate prediction. Therefore, reducing the computational complexity of deep learning models is an important challenge for future research.

In addition, many factors significantly affect air pollution data. For example, high wind speed can reduce PM2.5 concentrations, while high humidity often increases pollution levels. High atmospheric pressure is also often correlated with improved air quality. These meteorological features play an important role in building air quality prediction models, emphasizing the importance of incorporating weather factors into the analysis and forecasting process.

To address the challenges in enhancing the accuracy of air quality prediction, this study makes the following key contributions:

i) Propose extended features to improve the accuracy of air quality predictions in Hanoi, Vietnam.

ii) Development of a predictive model for the AQI index using the LSTM deep learning algorithm.

2. Data and method

2.1. The dataset

The datasets reflect various environmental conditions associated with air pollutant concentrations. Data on six pollutants - PM2.5, PM10, NO_2, O_3, SO_2, and CO - were collected from multiple monitoring stations in Hanoi, the capital of Vietnam. These pollutants served as predictors for the analysis and were recorded over the period from July 5, 2016, to July 1, 2020.

The raw dataset from the Hanoi monitoring stations consists of 2,115 rows, stored in a CSV file. This dataset is divided into two subsets: The first 2,005 rows are used for training, while the remaining 110 rows are reserved for testing. Before the training phase, data preprocessing is performed to address missing or invalid values in the raw dataset. Missing values are handled through an imputation process, which replaces them with the nearest data points. This approach is applied when the percentage of missing values is below 16 % for a row or 1 % for a column within the station datasets.

2.2. Feature selection

In Vietnam, pollution levels are assessed using parameters such as PM2.5, PM10, NO_2, O_3, SO_2, and CO. In this study, we forecast the values of PM2.5 and PM10 based on data from the preceding 15 days. To enhance forecasting performance, we incorporate extended features into the dataset in addition to the raw parameters from the collected data.

Since Particulate Matter (PM) can be directly emitted from sources or formed in the atmosphere through chemical reactions involving gases such as SO_2, NO_x, and certain organic compounds, and as pollutants often interact with one another, we propose an extended feature set. This set includes the following: the total measured values of pollutants for a given day, the average daily values of PM10 and PM2.5, and the average daily values of NO_2, CO, SO_2, and O_3. The extended features are detailed in Table 1.

Table 1. Description of extended features

No	Feature	Type	Description
1	S_Data	Numeric	Total of value 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 15 days earlier (Sliding window size = 15)
5	A_PM10	Numeric	Average of PM10 values in 15 days earlier

2.3. Sliding window

The sliding window method is a technique that utilizes prior time steps to predict the next time step. By applying this method, the input data is restructured into segments based on a defined window size. In this study, the window size was set to 15, meaning that data from the previous 15 days is used to predict the value for the subsequent day. After restructuring, the data resembles a supervised learning dataset, allowing it to be used with any machine learning algorithm for time series modeling [12]. If the dataset consists of m input time series, denoted as X, and an output series, denoted as Y, then:

$$\begin{aligned} X &= \{x_t^i\} \ \mathrm{t} \in \mathrm{T} = \{1, 2, \ldots, n\} \text{ and } i = 1, 2, \ldots, m \\ Y &= \{y_t\} \ \mathrm{t} \in \mathrm{T} \end{aligned} \tag{1}$$

The sliding window transformation process generates the following dataset:

$$\begin{aligned} X = \{ &\{x_{t-1}^1, x_{t-2}^1, \ldots, x_{t-w}^1\}, \{x_{t-1}^2, x_{t-2}^2, \ldots, x_{t-w}^2\} \ldots, \\ &x_{t-1}^m, x_{t-2}^m, \ldots, x_{t-w}^m\}, \ \{y_{t-1}, y_{t-2}, \ldots, y_{t-w}\}, y_t\} \ \textit{with} \ \mathrm{t} \in \mathrm{T} \end{aligned} \tag{2}$$

w: represents the window size.

2.4. Performance evaluation

Similar to the previous study, the performance of the models is evaluated using the metrics RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) [12]. These metrics are calculated based on the differences between the predicted results and the true values. Additionally, R2 (R-squared) is used to quantify the strength of the relationship between the predictive models and the target variables [13]. The mathematical expressions for these metrics are defined as follows:

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

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

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

Where $\hat{y}_i$ is the i^{th} predicted value, y_i is the i^{th} obsered value.

2.5. Model for predicting AQI

After preprocessing the data, the next steps involve preparing it for training and testing. In deep learning methods, model construction is typically achieved through experimental processes. In this study, we focus on utilizing a multi-step LSTM model. The proposed model for predicting AQI is shown in Fig 1.

3. Results and discussion

The models were trained on a Lenovo X270 computer with 8GB of RAM, using Python 3.9 in an integrated development environment (IDE).

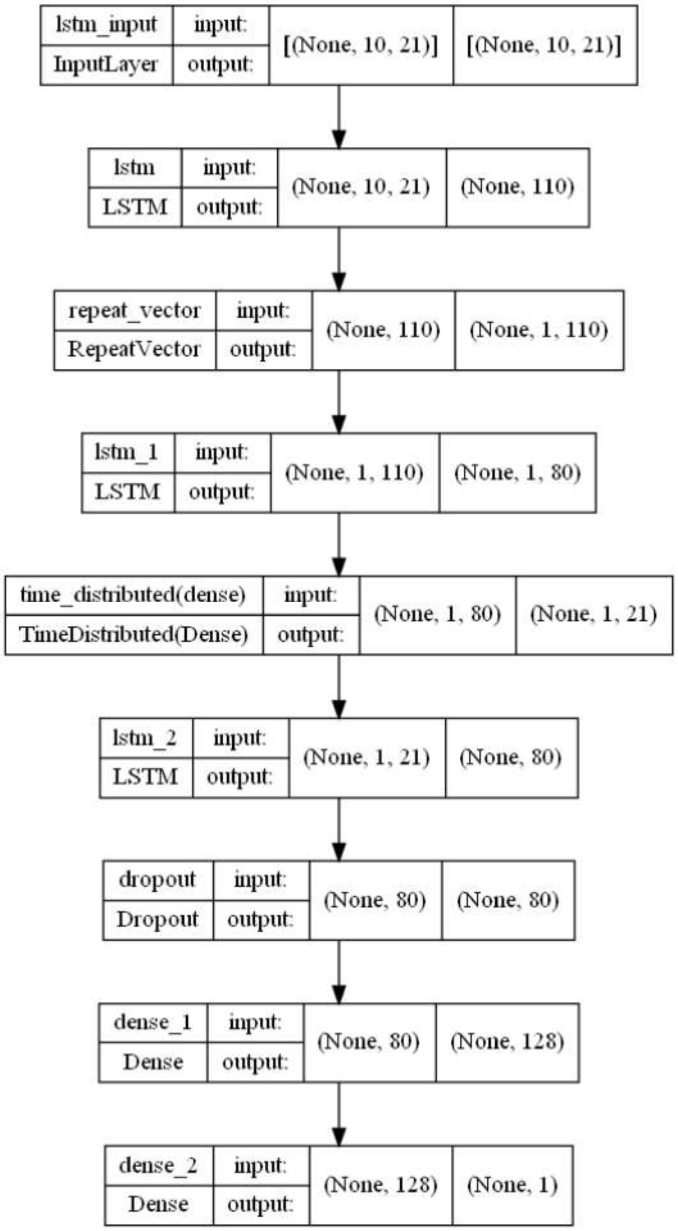

Figure 1: The proposed model for predicting AQI

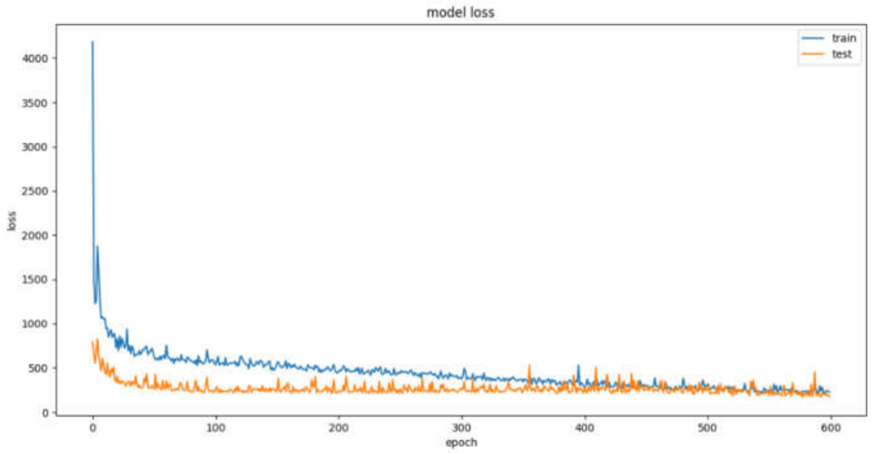

Figure 2: The loss model graph

Fig. 2 illustrates the loss graph for both the training and test sets. The loss on the training set decreases progressively with each epoch, indicating that the model is effectively learning from the data and improving its accuracy on the training set. Beyond epoch 450, the losses on the training and test sets converge, demonstrating that the proposed model adequately fits the data and aligns well with the prediction target.

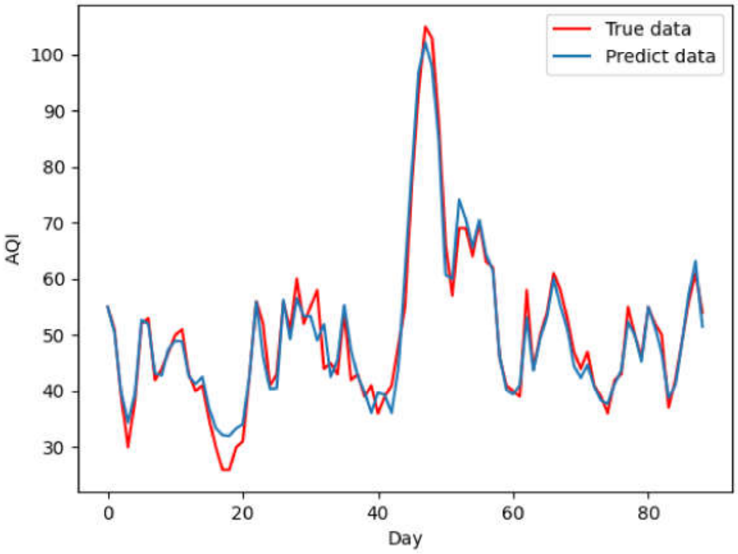

Figure 3: The trend of AQI real value and predicted value

Fig. 3 describes the trends of actual and predicted values for the AQI index in Hanoi during April, May, and June 2020, as modeled by our proposed approach. We employed the Encoder LSTM algorithm with a sliding window size of 15. The results demonstrate that our model effectively captures and predicts the trend of the AQI index.

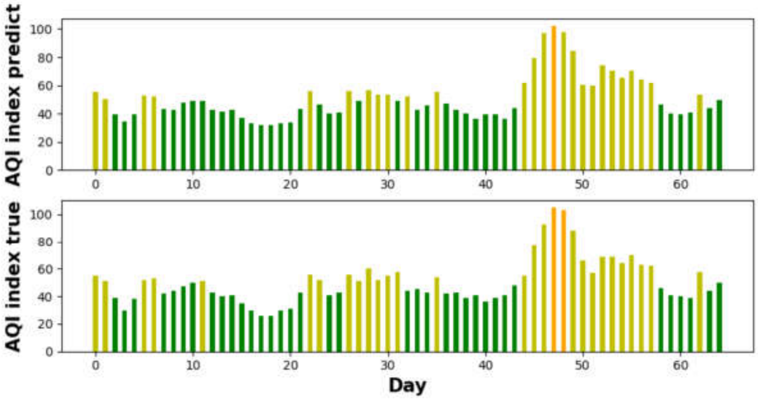

Figure 4: The true AQI values and AQI values from April to June 2020 in Hanoi

Fig. 4 shows the actual AQI values alongside the predicted AQI values in Hanoi from April to June 2020. The colors represent different levels of health concern: green indicates good air quality, yellow signifies moderate conditions, and orange or red denotes unhealthy levels [15].

Table 2 presents the detailed evaluation results of the prediction of the Air Quality Index (AQI) in Hanoi for the next three days, using the LSTM algorithm. The prediction was made based on the data of the previous 15 days, using the LSTM model with a sliding window. The results show that the machine learning algorithms achieved high performance in predicting the AQI index for the first day, with evaluation indices including MAE = 2.456, RMSE = 3.243, and $R2$ =0.96. This proves that the proposed model is suitable for the initial training data and target, demonstrating good learning and generalization ability.

However, when predicting the AQI index for the second and third days, the accuracy of the model tends to decrease. The reason is that the input data for the following days is taken from the prediction results of the previous days. This leads to an accumulation of errors over each prediction step, which gradually reduces the accuracy over time. However, the overall performance of the model is still reliable, showing potential for application to time-series AQI prediction problems, especially in applications requiring short-term forecasting.

Table 2. Performance of the LTSM AQI predicted model

The step ahead	MAE	RMSE	R2
1 day ahead	2.456	3.243	0.96
2 day ahead	6.782	8.587	0.723
3 day ahead	7.235	9.111	0.693

4. Conclusion

In this paper, we proposed a deep learning model to predict air quality based on AQI. The effectiveness of this method has been verified through the criteria: high accuracy with small MAE, short training time, and simple model structure. To evaluate the performance of the model, we use the MAE, RMSE, and R2 indices. The results show that the deep learning method with the LSTM algorithm, using a 15-day sliding window, is capable of predicting 1 day in advance with high reliability. Our experimental results achieved R2 = 0.96, based on a dataset collected over 4 years in Hanoi, Vietnam. This shows the potential application of the model not only in research but also in real-time air quality forecasting systems. This method can be used to assist managers in taking timely measures to protect public health and minimize the impacts of air pollution.

REFERENCES

[1]. N. Srinivasa Gupta, Yashvi Mohta, Khyati Heda, Raahil Armaan, B. Valarmathi, G. Arulkumaran (2023). *Prediction of air quality index using machine learning techniques: A comparative analysis.* Journal of Environmental and Public Health. https://doi.org/10.1155/2023/4916267.

[2]. Zhen Zhang, Shiqing Zhang, Caimei Chen, Jiwei Yuan (2024). *A systematic survey of air quality prediction based on deep learning*. Alexandria Engineering Journal, Volume 93, p. 128 - 141. ISSN: 1110-0168. https://doi.org/10.1016/j.aej.2024.03.031.

[3]. S. Du, T. Li, Y. Yang, S.-J. Horng (2019). *Deep air quality forecasting using a hybrid deep learning framework*. IEEE Trans. Knowl. Data Eng., 33 (2019), p. 2412 - 2424.

[4]. H. Huang, X. Wei, Y. Zhou (2022). *An overview on twin support vector regression*. Neurocomputing, 490 (2022), p. 80 - 92.

[5]. D. Borup, B.J. Christensen, N.S. Mühlbach, M.S. Nielsen (2023). *Targeting predictors in random forest regression*. Int. J. Forecast., 39 (2023), p. 841 - 868.

[6]. F. Ricardo, P. Ruiz-Puentes, L.H. Reyes, J.C. Cruz, O. Alvarez, D. Pradilla (2023). *Estimation and prediction of the air-water interfacial tension in conventional and peptide surface-active agents by Random Forest regression*. Chem. Eng. Sci., 265 (2023), Article 118208.

[7]. Kim D, Han H, Wang W, Kang Y, Lee H, Kim HSJAS (2022). *Application of deep learning models and network method for comprehensive air-quality index prediction*. Appl Sci. 2022; 12(13):6699.

[8]. Liu H, Li Q, Yu D, Gu YJAS (2019). *Air quality index and air pollutant concentration prediction based on machine learning algorithms*. Appl Sci. 2019; 9(19):4069.

[9]. Wu Q, Lin HJSOTTE (2019). *A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors*. Sci Total Environ. 2019;683:808 - 21.

[10]. Phruksahiran NJUC (2021). *Improvement of air quality index prediction using geographically weighted predictor methodology.* Urban Climate. 2021;38: 100890.

[11]. A. Ali, Y. Zhu, M. Zakarya (2022). *Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flow prediction*. Neural Netw., 145 (2022), p. 233 - 247.

[12]. Raquel Espinosa, José Palma, Fernando Jiménez, Joanna Kamińska, Guido Sciavicco, Estrella Lucena-Sánchez (2021). *A time series forecasting-based multi-criteria methodology for air quality prediction*. Applied Soft Computing 113 (2021) 107850..

[13]. Yun-Chia Liang, Yona Maimury, Angela Hsiang-Ling Chen, and Josue Rodolfo Cuevas Juarez (2020). *Machine learning-based prediction of air quality*. Appl. Sci. 2020, 10, 9151. Doi:10.3390/app10249151.

[14]. Dufour, J. M., (2011). *Coefficients of Determination*. McGill University: Québec, QC, Canada.

[15]. Nathaniel Mopa Wambebe, Xiaoli Duan (2020). *Air quality levels and health risk assessment of particulate matters in Abuja municipal area, Nigeria*. Atmosphere 2020, 11, 817. Doi:10.3390/atmos110808.