Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 05
Issue 01

June, 2024

Research Article

Hybrid CNN-LSTM Model for Air Quality Prediction: A Case Study for Gurugram

Anıl Utku[1]*

Computer Engineering, Faculty of Engineering. Munzur University, 62000, Tunceli, Turkey

ARTICLE INFO

Article history:
Received April 16, 2024
Revised April 25, 2024
Accepted May 02, 2024

Keywords:
Air quality
Deep learning
Machine learning
CNN
LSTM

ABSTRACT

One of the most important environmental problems brought about by rapid population growth and industrialization is air pollution. Today, air pollution is generally caused by heating, industry and motor vehicles. In addition, factors such as unplanned urbanization, topographic structure of cities, atmospheric conditions and meteorological parameters, building and population density also cause pollution to increase. Pollutants with concentrations above limit values have negative effects on humans and the environment. In order to prevent people from being negatively affected by these pollutants, it is necessary to know the pollution level and take action as soon as possible. In this study, a hybrid ConvLSTM model was developed in order to quickly and effectively predict air pollution, which has such negative effects on humans and the environment. ConvLSTM was compared with LR, RF, SVM, MLP, CNN and LSTM using approximately 4 years of air quality index data from the city of Gurugram in India. Experimental results showed that ConvLSTM was significantly more successful than the base models, with 30.645 MAE and 0.891 R^2.

1. Introduction

Air Quality Index (AQI) allows inferences to be made about how polluted the air of a particular area is and what health effects this pollution may cause [1]. AQI indicates the health effects that may occur in the short or long term after breathing polluted air [2]. AQI can be thought of as a scale ranging from 0-500. It is thought that as the AQI value increases, air pollution increases and the health risk increases. An AQI value 0-50 indicates that the air quality is good and there is little risk of affecting health [3, 4]. An AQI value above 300 indicates that the air quality is dangerous and therefore the health risk is high [5]. An index value below 100 is generally indicative of good air quality. When the AQI value exceeds 100, it is inferred that the air quality is unhealthy [5, 6].
Each limit for AQI corresponds to a different level of health. "Good" means the AQI value is in the range of 0-50. It means there is no health risk to air pollution [7]. “Moderate” means the AQI value is in the range of 51-100 and some pollutants may have moderate adverse health effects on certain segments of society [8, 9]. “Unhealthy for sensitive groups” means that the AQI value is in the range of 101-150 and people who are sensitive to certain pollutants are likely to be affected by this level of pollution [9]. An "unhealthy" AQI value in the range of 151-200 means health problems are likely to occur for all segments of society. "Very unhealthy" means an AQI value in the range of 201-300. All segments of society can be seriously affected [10]. "Dangerous" means an AQI value above 300. It is an emergency point that can affect all segments of society.
AQI is a scale used to measure the effects of atmospheric particulate matter on human health and the environment. AQI is calculated using the amount of particulate matter and meteorological observation

[1] * Corresponding author
e-mail: anilutku@munzur.edu.tr
DOI: 10.55195/jscai.1469468

parameters [10, 11]. Artificial intelligence-based models that can be developed using historical air quality data can provide higher accuracy forecasts. Artificial intelligence techniques analyze large amounts of data and model complex patterns among the data [12, 13]. Deep learning models, in particular, are very successful in modeling and learning complex relationships due to their structure. The motivation of this study is to increase the efficiency and accuracy of predicting in order to reduce the effects of air pollution on human health. The main purpose of this study are to protect human health, monitor environmental conditions, detect emergency situations in advance, and help ensure that necessary precautions are taken by predetermining risks for risk groups such as the elderly and children. In this study, a hybrid ConvLSTM model for AQI prediction was developed. With ConvLSTM, it was aimed to benefit from the prominent and effective features of CNN and LSTM. The effectiveness of CNN in the feature extraction phase and the success of LSTM in modeling and learning complex relationships were utilized. The advantages of ConvLSTM are that it can capture long-term dependencies in the data, effectively model time series data, and model complex relationships between data.

Contributions of this study to the literature:

- A hybrid ConvLSTM model was developed to improve the prediction quality and accuracy.
- This is the first study in the literature using this dataset.
- ConvLSTM was compared with base models such as Convolutional Neural Network (CNN), Linear Regression (LR), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Long-Short Term Memory (LSTM).

Experimental results showed that ConvLSTM is more effective than compared models. ConvLSTM has a very successful prediction performance with 30.645 MAE and 0.891 R^2.

2. Related Works

Artificial intelligence methods are successfully used in application areas such as forecasting meteorological data and air quality forecasting. In the rest of this section, studies in the literature using artificial intelligence methods are examined.

Mishra and Gupta presented a comparative analysis of statistical models and machine learning and deep learning models for air quality prediction [13]. Beijing's air quality data between 2014 and 2018 were used. Boosting algorithms, LSTM, Decision Tree (DT), Autoregressive Integrated Moving Average (ARIMA), Huber Regressor, k-Nearest Neighbor (kNN), and Dummy Regressor were compared. Experiments showed that LSTM outperformed compared models in short-term forecasts.

Ravindiran et al. presented a comparative analysis of XGBoost, LightGBM, Catboost, RF, and Adaboost algorithms for air quality prediction [14]. Approximately 5 years of air quality data from Visakhapatnam, India, were used in the study. Experiments showed that Catboost outperformed comparison models with 0.9998 R^2.

Van et al. presented a comparative analysis of DT, XGBoost, and RF for air quality prediction [15]. In the study, 6-year data provided by the Central Pollution Control Board of India and 1-month India Open Government Data were used. Experiments showed that XGBoost outperformed benchmark models with 0.9214 R^2 and 0.9993 R^2.

Maltare and Vahora presented an applied analysis of SVM, SARIMA and LSTM for predicting air quality of Ahmedabad city [16]. Approximately 7 years of data provided by the Central Pollution Control Board of India were used in the study. Experiments showed that SVM using RBF kernel outperformed the compared models with 4.94 RMSE.

Drewil and Al-Bahadili proposed a model in which the hyper-parameters of LSTM for air quality prediction are determined by genetic algorithm [17]. The developed model aimed to predict PM_{10}, CO, $PM_{2.5}$, and NO_X concentrations. India's air quality data between 2017 and 2020 was used as the dataset. Experiments showed that the developed model outperformed the compared models with 9.58 RMSE.

Kurnaz and Demir developed a Recurrent Neural Network (RNN)-based model for PM_{10} and SO_2 prediction [18]. Air quality data of Sakarya province for the years 2018–2020 were used. Experiments showed that the prediction result for SO_2 was 2.84 RMSE and for PM_{10} was 4.09 RMSE.

Kristiani et al. presented a comparative analysis of deep learning models for predicting meteorological parameters and concentrations of air pollutants such as SO_2, O_3, and CO_2 [19]. Data from 2017-2019 provided by the Taiwan Environmental Protection Administration was used as the dataset. Experiments showed that LSTM has 1.9 RMSE, CNN has 3.5 RMSE, Bi-LSTM has 2.5, Bi-GRU has 2.7 and RNN has 2.4 RMSE.

3. Material and Method

In this study, the AQI of Gurugram city in India was used [20]. The dataset used includes 1488 lines of AQI data between March 5, 2020 and March 31, 2024. Table 1 shows the first 5 lines of the dataset as an example.

Table 1 The first 5 lines of the dataset

Date	AQI
2020-03-05	73.0
2020-03-06	55.0
2020-03-07	78.0
2020-03-08	120.0
2020-03-09	179.0

Figure 1 shows the change in AQI according to date.

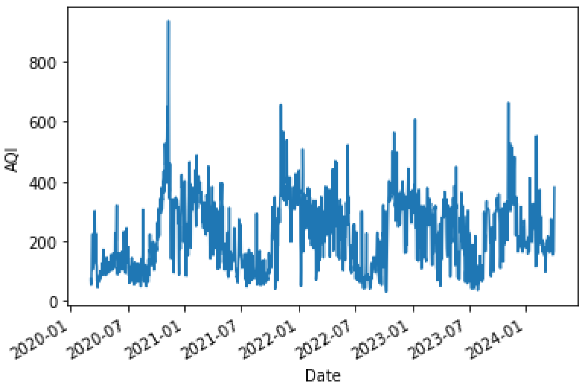

Figure 1 The change in AQI according to date

During the data pre-processing phase, missing or incorrect fields in the dataset were checked. The sliding window module was used to structure time series data as a supervised learning problem. Sliding window allows data to be configured as input and output according to the specified window size [21]. As a result of the experimental studies, the lowest error rates were obtained when the window size was 3. The dataset was structured so that the data points at time t_1, t_2, and t_3 were input, and the data point at time t_4 was output, as seen in Figure 2.

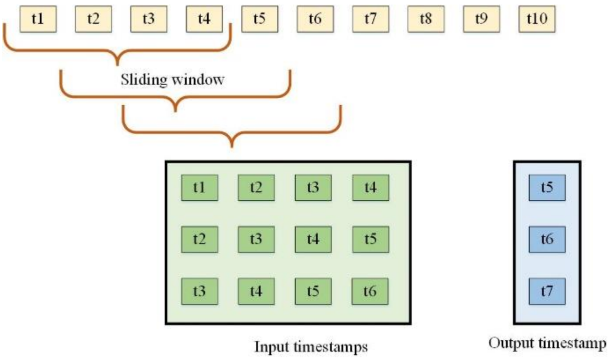

Figure 2 Sliding window method

MinMax normalization was used to scale the values in the dataset to a certain range. Normalization increases the performance of the model by scaling the data to specified ranges [22]. 80% of the dataset was used for training the models and 20% for testing the models. 10% of the training data was used for optimization of model hyper-parameters. For each applied model, the hyper-parameters with the highest prediction accuracy were determined using GridSearch.

3.1. Prediction Models

LR is a statistical analysis and machine learning method used to model the correlation between two or more variables and make predictions [23]. LR is based on the assumption that this relationship is linear. The relationship between dependent and independent variables can be expressed by a linear equation [24]. In order to perform linear regression analysis, both variables must be continuous data type.

RF is an ensemble learning algorithm created by combining many decision trees [25]. Each decision tree is trained on a randomly sampled subset of data, and these trees are created with random features. RF is highly resistant to noise and overfitting in the data [26]. The model does general learning rather than being too specific to the training data, and thus can make better predictions on unseen data. It is created by the combination of Bagging and Random Subspace methods [27]. Observations for trees are selected with the boostrap random sample selection method and variables are selected with the subspace method.

SVM is a generalization of a simple and intuitive classifier called the maximum marginal classifier [28]. The basic idea of maximum marginal classification is to maximize the margin (gap) between classes. Classes are separated from each other by as wide a space as possible. Margin refers to the distance of the closest data points between classes to the hyperplane. This distance is the sum of the distance of one class to the hyperplane and the distance of the other class to the hyperplane. Maximum marginal classification attempts to maximize this gap [29]. The best hyperplane is the hyperplane that maximizes the margin (space) between classes. SVM helps express nonlinear relationships by moving data into high-dimensional spaces using kernel functions.

MLP consists of multiple layers of interconnected neurons, where each layer processes information from the previous layer [30]. MLP is a mathematical model that attempts to mimic the way the humanbrain processes information. MLP has multiple layers of interconnected neurons. These layers are usually organized as an input layer, output layer and hidden layers. In MLP, which is trained using methods such as backpropagation or gradient descent, the input layer is responsible for transferring the data to the model, and the output layer is responsible for presenting the prediction results [31]. Complex relationships in the data are learned through trace layers located between the input and output layers. Additionally, neurons are enabled to learn complex and non-linear patterns with the help of activation functions.

CNN is an effective model used for image processing, classification and segmentation. The main purpose of CNN is feature extraction and pattern recognition. The input layer provides input data to the network [32]. Convolution layers perform convolution operations to identify different features in the data. Each convolution layer pans over the image using filters. Pooling layers are used to reduce the size of the feature map. Fully connected layers enable feature maps to be flattened [33].

LSTM is a model developed to overcome the limitations of recurrent neural network models. LSTM has gate mechanisms as a solution to the vanishing gradient problem experienced in training RNN [34]. Gate mechanisms enable analysis of long-term dependencies by deciding what information to remember/forget. LSTM updates the hidden state in memory cells at each time step. The gates allow new data to enter the LSTM cells. The forgetting gate determines what information will be stored in the cells. The output gate is responsible for transferring information in the cells. LSTM's architecture that enables selective remembering or forgetting makes it suitable for problems that require modeling long-term dependencies [35].

3.2. Developed Hybrid Model

The ConvLSTM model was developed to model complex and long-term dependencies in time series data. CNN is an efficient model that uses convolution layers to identify local patterns in data and extract features. LSTM is a model that can store information from past time steps and learn relationships that change depending on time. The structure of the developed model is seen in Figure 3. In the developed system, data is first pre-processed and the dataset is converted into a supervised learning structure using a sliding window. Data is scaled using MinMax normalization and hyper-parameters with the lowest error value are determined with the help of GridSearch. The features extracted by CNN are presented as input to LSTM. LSTM performs the learning and prediction process and provides the predicted AQI value as output.

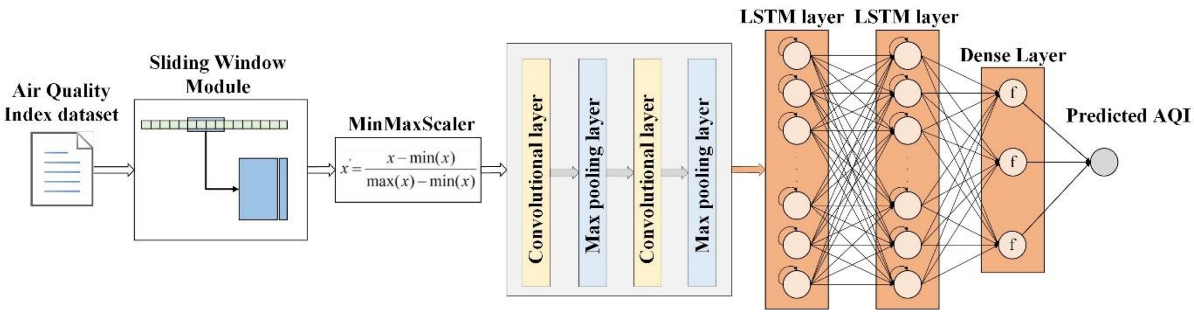

Figure 3 The structure of ConvLSTM

CovLSTM consists of TimeDistributed Conv1D layers, TimeDistributed MaxPooling1D layer and Flatten layer in the CNN component. In Conv1D layers, filter size is 32 and kernel size is 1. The activation function is ReLU. The LSTM component has a double-layered structure, each consisting of 64 neurons. LSTM's activation function is ReLU, number of epoch is 100, number of neurons is 64 and its optimizer is Adam.

4. The Experimental Results

In this study, a comparative analysis of LR, RF, SVM, MLP, CNN and LSTM with the developed ConvLSTM-based deep learning model for estimating the daily AQI value of Gurugram, one of the leading industrial centers of India, is presented.

For each algorithm and model applied, the results obtained according to MSE, RMSE, MAE and R^2 metrics were comparatively analyzed. Table 2 and Figure 4 show comparative experimental results.

Table 2 The experimental results

Models	MSE	RMSE	MAE	R^2
LR	4415.332	66.447	47.944	0.699
RF	4364.763	66.066	45.775	0.704
SVM	4338.909	65.870	45.709	0.706
MLP	4212.788	64.903	45.037	0.726
CNN	4332.878	65.825	45.918	0.707
LSTM	3846.480	62.020	43.263	0.760
ConvLSTM	**2004.157**	**44.767**	**30.645**	**0.891**

Experimental results have shown that ConvLSTM is more successful than compared models. ConvLSTM's MSE value is 2004.157, RMSE value is 44.767, MAE value is 30.645 and R^2 value is 0.891. After ConvLSTM, LSTM, MLP, SVM, RF, CNN and LR were successful respectively.

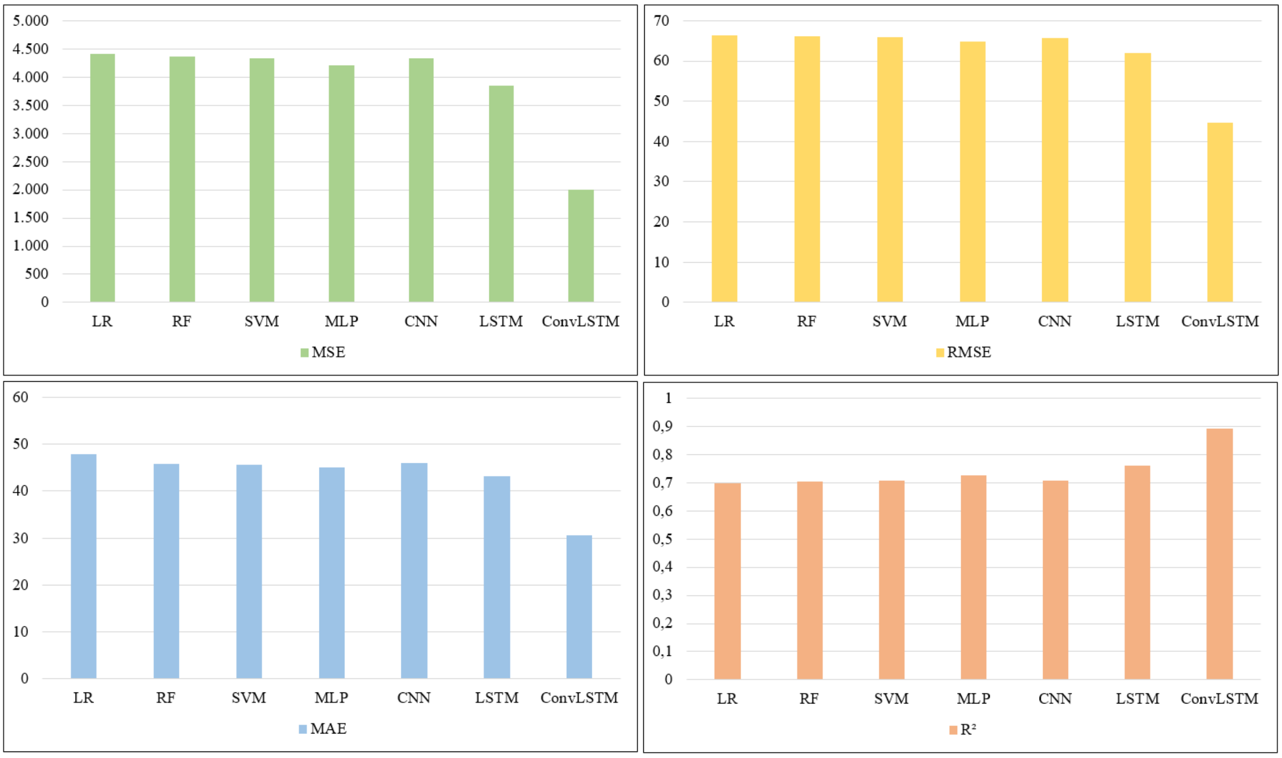

Figure 4 The experimental results

Figure 4 presents the experimental results of the compared models according to performance evaluation metrics. As seen in Figure 4, ConvLSTM has lower MSE, RMSE, and MAE than the compared models. The value of these metrics being close to 0 indicates that the model is more successful. The R2 is a measure of how well the applied model fits the data set, and being close to 1 indicates that the model is successful. ConvLSTM outperformed comparison models with 0.891 R2.

Figure 5 shows the prediction graphs of the compared models.

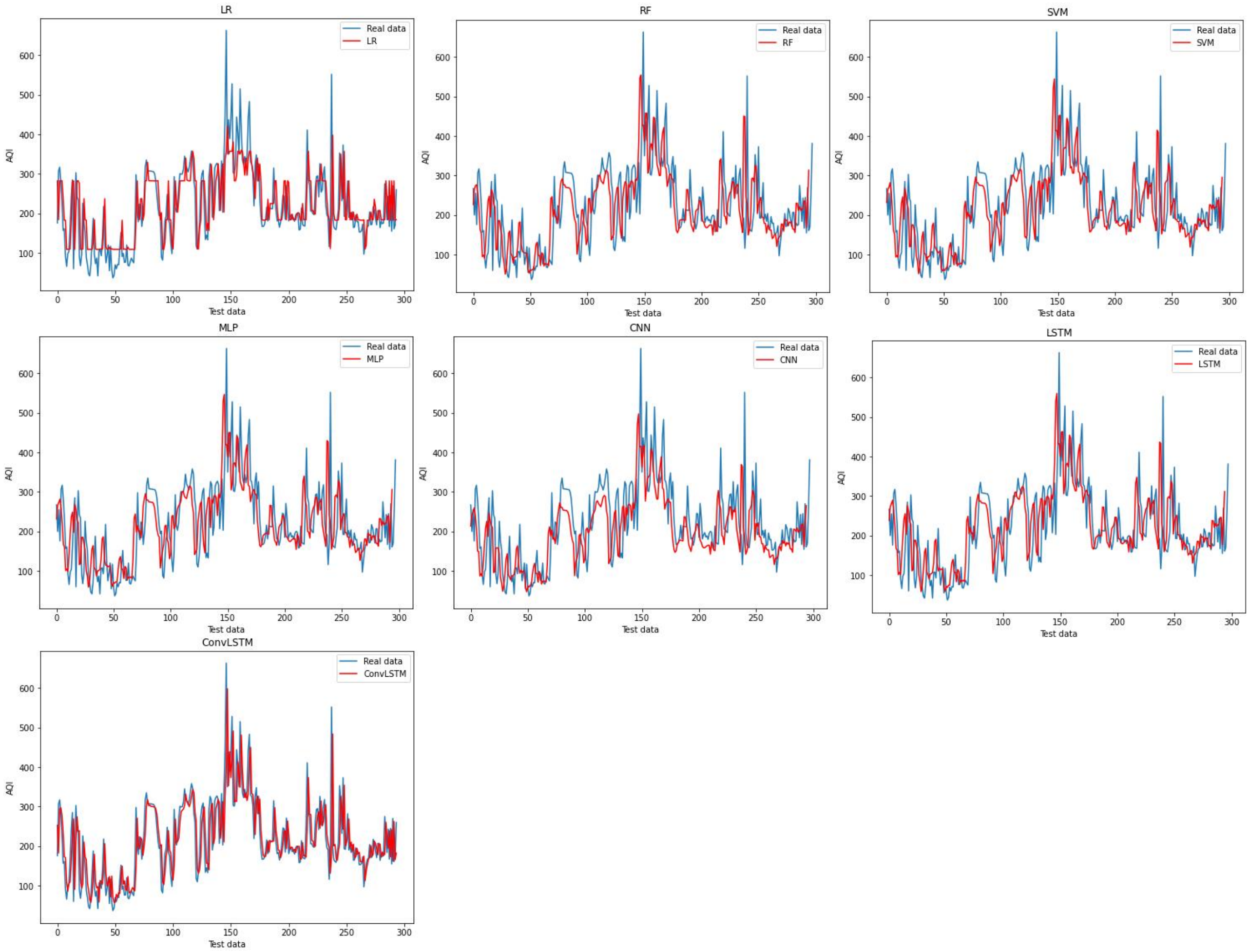

Figure 5 Prediction graphs of the compared models

As seen in Figure 5, ConvLSTM was able to predict the changes in the dataset more successfully than other models. Looking at the prediction graphs of the models, it is seen that ConvLSTM models the fluctuations and dynamics in the dataset more successfully than the compared models.

The fact that ConvLSTM is more successful than the compared models can be explained by the fact that ConvLSTM can model long-term dependencies better by using the iterative structure of LSTM. Additionally, ConvLSTM has the ability to perform better modeling by extracting complex relationships from local patterns extracted by CNN. In addition, since the relationships in time series data are nonlinear, ConvLSTM is more successful than nonlinear models such as LR and MLP. LR, RF and SVM are insufficient to model long-term dependencies in time series data. Additionally, ConvLSTM can capture hidden patterns in the data, such as seasonal variations and trends.

5. Discussion

In this study, a hybrid ConvLSTM model was created using CNN and LSTM models for air quality prediction. Experiments showed that ConvLSTM outperformed the compared models. However, ConvLSTM also has its limitations. The predictive power and success of ConvLSTM depends on the amount and quality of data. ConvLSTM processes long-term dependencies thanks to its LSTM component, but it may be insufficient to model very long-term dependencies. Additionally, hyper-parameter optimization is very important for ConvLSTM. If the most appropriate hyper-parameter combinations cannot be determined, the model may not be successful enough.

6. Conclusions

Air pollution, the most important environmental problems today, seriously threatens the world of the future. With increasing population and urbanization, increased energy use, fossil fuel use, and air pollution resulting from industrialization have adverse effects on human health and the environment. Air pollution is the disruption of the natural composition of the air by pollutants such as particulate matter. It is the presence of particulate matter in the air at concentrations that can harm human health and ecological balance. Meteorological factors, location and geographical structure of the region also affect air pollution. Therefore, it is important to constantly monitor and analyze air quality.

In this study, the ConvLSTM hybrid model was developed to more effectively predict the AQI of Gurugram, one of the industrial cities in India. In order to test the effectiveness of the developed model, ConvLSTM was compared with base models such as LR, RF, SVM, MLP, CNN and LSTM. Experimental results showed that ConvLSTM outperformed the compared models with MSE, RMSE, MAE and R^2 value of 2004.157, 44.767, 30.645 and 0.891, respectively. These results demonstrate the effectiveness of artificial intelligence techniques in monitoring and evaluating air quality. Additionally, the results are promising in terms of developing more successful prediction models by using more comprehensive data in real-world applications.

References

[1] X. Tan, L. Han, X. Zhang, W. Zhou, W. Li, & Y. Qian, A review of current air quality indexes and improvements under the multi-contaminant air pollution exposure. J. Env. Manag., c. 279, 2021.

[2] M. Leili, A. Nadali, M. Karami, A. Bahrami, & A. Afkhami, Short-term effect of multi-pollutant air quality indexes and PM2. 5 on cardiovascular hospitalization in Hamadan, Iran: a time-series analysis. Env. Sci. and Poll. Res., c. 28, sy 38, ss. 53653-53667, 2021.

[3] P. Kumar, A critical evaluation of air quality index models (1960–2021). Environmental Monitoring and Assessment, c. 194, sy 5, ss. 1-45, 2022.

[4] R. Cao, Y. Wang, J. Huang, Q. Zeng, X. Pan, G. Li, & T. He, The construction of the air quality health index (AQHI) and a validity comparison based on three different methods. Env. Res., sy 197, 2021.

[5] X. Sui, K. Qi, Y. Nie, N. Ding, X. Shi, X. Wu, & W. Wang, Air quality and public health risk assessment: A case study in a typical polluted city, North China. Urban Climate, sy 36, 2021.

[6] Y. Wang, L. Huang, C. Huang, J. Hu, & M. Wang, High-resolution modeling for criteria air pollutants and the associated air quality index in a metropolitan city. Env. Int., sy 172, 2023.

[7] F. O. Abulude, I. A. Abulude, S. D. Oluwagbayide, S. D. Afolayan, & D. Ishaku, Air Quality Index: A case of 1-day monitoring in 253 Nigerian urban and suburban towns. Journal of Geovisualization and Spatial Analysis, c. 6, sy 1, 2022.

[8] Z. Jiang, Y. Gao, H. Cao, W. Diao, X. Yao, C. Yuan, & Y. Chen, Characteristics of ambient air quality and its air quality index (AQI) model in Shanghai, China. Sci. Total Env., sy 896, 2023.

[9] F. Abulude, I. Abulude, S. Oluwagbayide, S. Afolayan, & D. Ishaku, Air Quality Index: Case of One-Day Monitoring of 253 Urban and Suburban Towns in Nigeria. Env. Sci. Proc., c. 8, sy 1, 2021.

[10] D. P. K. Meena, & D. V. Singh, Air Quality Monitoring and Pollution Control Technologies, Int.l J. Multidiscip. Res. Sci., Eng. Tech., c. 7, sy 2, ss. 4409-4426, 2024.

[11] A. Sengupta, G. Govardhan, S. Debnath, P. Yadav, S. H. Kulkarni, A. N. Parde, & S. D. Ghude, Probing into the wintertime meteorology and particulate matter (PM2. 5 and PM10) forecast over Delhi. Atmos. Poll. Res., c. 13, sy 6, 2022.

[12] H. Nozari, J. Ghahremani-Nahr, & A. Szmelter-Jarosz, AI and machine learning for real-world problems. Adv. in Comp., sy 134, ss. 1-12, 2024.

[13] A. Mishra, & Y. Gupta, Comparative analysis of Air Quality Index prediction using deep learning algorithms. Spat. Inf. Res., c. 32, sy 1, ss. 63-72, 2024.

[14] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, & C. Sonne, Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam. Chemosphere, sy 338, 2023.

[15] N. H. Van, P. Van Thanh, D. N. Tran, & D. T. Tran, A new model of air quality prediction using lightweight machine learning. Int. J. Env. Sci. Tech., c. 20, sy 3, ss. 2983-2994, 2023.

[16] N. N. Maltare, & S. Vahora, Air Quality Index prediction using machine learning for Ahmedabad city. Dig. Chem. Eng., sy 7, 2023.

[17] G. I. Drewil, & R. J. Al-Bahadili, Air pollution prediction using LSTM deep learning and metaheuristics algorithms. Measurement: Sensors, sy 24, 2022.

[18] G. Kurnaz, & A. S. Demir, Prediction of SO2 and PM10 air pollutants using a deep learning-based recurrent neural network: Case of industrial city Sakarya. Urban Climate, sy 41, 2022.

[19] E. Kristiani, H. Lin, J. R. Lin, Y. H. Chuang, C. Y. Huang, & C. T. Yang, Short-term prediction of PM2. 5 using LSTM deep learning methods. Sustainability, c. 14, sy 4, 2022.

[20] “Gurugram's Air Quality Index Time-Series Dataset”, Kaggle. https://www.kaggle.com/datasets/pranaii/test-aqi/data (Erişim 1 Mart 2024)

[21] S. Park, S. Jung, S. Jung, S. Rho, & E. Hwang, Sliding window-based LightGBM model for electric load forecasting using anomaly repair. J. Supercomp, sy 77, ss. 12857-12878, 2021.

[22] H. Henderi, T. Wahyuningsih, & E. Rahwanto, Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. Int. J. Inf Sys., c. 4, sy 1, ss. 13-20, 2021.

[23] X. Song, X. Liu, F. Liu, & C. Wang, Comparison of machine learning and logistic regression models in

predicting acute kidney injury: A systematic review and meta-analysis. Int. J. Med. Inf., sy 151, 2021.
[24] S. W. Lee, Regression analysis for continuous independent variables in medical research: statistical standard and guideline of Life Cycle Committee. Life cycle, sy 2, 2022.
[25] M. M. Ghiasi, & S. Zendehboudi, Application of decision tree-based ensemble learning in the classification of breast cancer. Comp. Bio. Med., sy 128, 2021.
[26] X. Zhou, H. Wen, Y. Zhang, J. Xu, & W. Zhang, Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. Geosci. Front., c. 12, sy 5, 2021.
[27] S. Talukdar, K. U. Eibek, S. Akhter, S. K. Ziaul, A. R. M. T. Islam, & J. Mallick, Modeling fragmentation probability of land-use and land-cover using the bagging, random forest and random subspace in the Teesta River Basin, Bangladesh. Eco. Indic., 126, 2021.
[28] A. Utku. Derin Öğrenme Tabanlı Trafik Yoğunluğu Tahmini: İstanbul İçin Bir Vaka Çalışması. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, c. 11, sy 3, ss. 1584-1598, 2023.
[29] A. Rizwan, N. Iqbal, R. Ahmad, & D. H. Kim, WR-SVM model based on the margin radius approach for solving the minimum enclosing ball problem in support vector machine classification. Appl. Sci., c. 11, sy 10, 2021.
[30] A. Utku. Deep Learning Based an Efficient Hybrid Model for Urban Traffic Prediction. Bilişim Teknolojileri Dergisi, c. 16, sy 2, ss. 107-117, 2023.
[31] H. Alla, L. Moumoun, & Y. Balouki, A multilayer perceptron neural network with selective-data training for flight arrival delay prediction. Sci. Prog., ss. 1-12, 2021.
[32] Y. Liu, H. Pu, & D. W. Sun, Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. Trends in Food Sci. & Tech., sy 113, ss. 193-204, 2021.
[33] X. Meng, N. Shi, D. Shi, W. Li, & M. Li, Photonics-enabled spiking timing-dependent convolutional neural network for real-time image classification. Optics Express, c. 30, sy 10, ss. 16217-16228, 2022.
[34] Q. Wang, R. Q. Peng, J. Q. Wang, Z. Li, & H. B. Qu, NEWLSTM: An optimized long short-term memory language model for sequence prediction. IEEE Access, sy 8, ss. 65395-65401, 2020.
[35] Y. Kaya, Z. Yiner, M. Kaya, & F. Kuncan, F. A new approach to COVID-19 detection from X-ray images using angle transformation with GoogleNet and LSTM. Measurement Science and Technology, c. 33, sy 12, 2022.