International Journal of Computational Intelligence and Applications
Vol. 21, No. 2 (2022) 2250014 (11 pages)

DOI: 10.1142/S1469026822500146

World Scientific
www.worldscientific.com

Comparison of Nitrogen Dioxide Predictions During a Pandemic and Non-pandemic Scenario in the City of Madrid using a Convolutional LSTM Network

Ditsuhi Iskandaryan*, Francisco Ramos and Sergio Trilles
Institute of New Imaging Technologies
Universitat Jaume I, Avinguda de Vicent Sos Baynat
s/n Castelló de la Plana 12071, Spain
**iskandar@uji.es*

Received 12 September 2021
Revised 4 April 2022
Accepted
Published

Traditionally, machine learning technologies with the methods and capabilities available, combined with a geospatial dimension, can perform predictive analyzes of air quality with greater accuracy. However, air pollution is influenced by many external factors, one of which has recently been caused by the restrictions applied to curb the relentless advance of COVID-19. These sudden changes in air quality levels can negatively influence current forecasting models. This work compares air pollution forecasts during a pandemic and non-pandemic period under the same conditions. The ConvLSTM algorithm was applied to predict the concentration of nitrogen dioxide using data from the air quality and meteorological stations in Madrid. The proposed model was applied for two scenarios: pandemic (January–June 2020) and non-pandemic (January–June 2019), each with sub-scenarios based on time granularity (1-h, 12-h, 24-h and 48-h) and combination of features. The Root Mean Square Error was taken as the estimation metric, and the results showed that the proposed method outperformed a reference model, and the feature selection technique significantly improved the overall accuracy.

Keywords: Air quality prediction; machine learning; ConvLSTM; COVID-19.

1. Introduction

Many studies have confirmed the effectiveness of machine learning technologies, for instance, for analyzing the time series of a recurrent neural network.[1,2] Because forecasting air quality can be viewed as a time series analysis, all applied time series analysis methods and algorithms can also be used to forecast air quality. In addition to forecasting along the time axis, it is also important to know the air quality value in places with no stations. Several authors have focused on the spatial factor in their studies.[3,4] In order to capture spatiotemporal patterns more efficiently and make very accurate predictions, various studies have suggested using Convolutional LSTM (ConvLSTM) to predict rainfall,[5] traffic accidents[6] and air quality,[7] amongst other

things. The importance of forecasting air quality and handling its consequences is growing day by day and continues to be the center of government and scientific attention. The study shows that short-term and long-term exposures to air pollutants cause about seven million death annually.[8] Various approaches and control measures have been implemented to reduce the concentration of these pollutants. Having information about the future concentration in advance can prompt decision-makers to implement certain strategies that can decrease concentration. To improve the accuracy of forecasts and the choice of the model, it is also very important to consider the factors that may directly or indirectly affect air quality. One of these factors is the lockdowns imposed due to the coronavirus disease 2019 (COVID-19) pandemic. To control the COVID-19 outbreak, all countries adopted severe traffic restrictions and self-quarantine measures,[9] which resulted in a decrease in air pollution.[10] An example of this was especially apparent in Madrid, where due to COVID-19 restrictions, the concentration of nitrogen dioxide (NO_2) dropped to 62%.[11] Taking into account the above information, the main objective of this work is to predict NO_2 concentration using ConvLSTM and compare it with LSTM, which, looking at the Table 1, can be considered as a benchmark model (Table 1 shows publications focused on NO_2 prediction and implemented methods, these results are extracted from the following paper).[12] The analysis is done for two scenarios: pandemic (January–June 2020) and non-pandemic (January–June 2019), in each of which the following sub-scenarios were defined, based on time intervals (1-h, 12-h,

Table 1. Publications focused on the prediction of nitrogen dioxide and implemented algorithms.

Work	ML Algorithm	Work	ML Algorithm
15	BRT, SVM, XGBoost, RF, GAM, Cubist	22	LSTM
1	LSTM	23	MLR, MLPNN, ELM, OSMLR, OSELM
2	LSTM	24	ELM
16	ANN	25	SVM, M5P model trees, ANN
17	Cluster-based bagging	26	MLP with hierarchical clustering, SOM and k-means clustering.
18	GAM, Bagging, RF, GBM, ANN, KRLS, SVR, Linear stepwise regression algorithms, regularization or shrinkage algorithms	27	Ensemble model with DRR
19	AIS-RNN (RNN, LSTM, GRU)	28	SVM
20	RF partition model	29	SVM
21	LSTM		

*BRT is Boosted Regression Trees, SVM is Support Vector Machine, XGBoost is EXtreme Gradient Boosting, RF is Random Forest, GAN is Generalized Additive Model, LSTM is Long Short Term Memory, ANN is Artificial Neural Network, GBM is Gradient Boosting Machines, KRLS is Kernel-based Regularized Least Squares, AIS is Adaptive Input Selection, RNN is Recurrent Neural Network, GRU is Gated Recurrent Unit, MLR is Multiple Linear Regression, MLPNN is Multi-layer Perceptron Neural Networks, ELM is Extreme Learning Machine, OSMLR is Online Sequential Multiple Linear Regression, OSELM is Online Sequential Extreme Learning Machine, SOM is Self-organizing Map, DRR is Discounted Ridge Regression.

24-h and 48-h) and features combinations. The Root Mean Square Error (RMSE) metric was applied to evaluate the results provided by each of the models.

2. Materials and Methods

The study area of this work was the city of Madrid (Fig. 1). According to the study by Khomenko *et al.*[13] related to premature mortality due to air pollution in European cities, which considered $PM_{2.5}$ and NO_2 as the main pollutants, Madrid was found to be in the top position of the ranking of European cities with the highest NO_2 mortality burden. Therefore, bearing in mind the importance of NO_2 for Madrid, it was selected as an air pollutant for predictive analysis.

The datasets used in this work are NO_2 data and meteorological data from January to June 2019 (non-pandemic scenario) and from January to June 2020 (pandemic scenario) and the location of the control stations, which was used to generate grid cells. The data were obtained from *Open data portal of the Madrid City Council*.[14] There are 24 air quality control stations and 26 meteorological control stations. The meteorological data includes ultraviolet radiation, wind speed, wind direction, temperature,

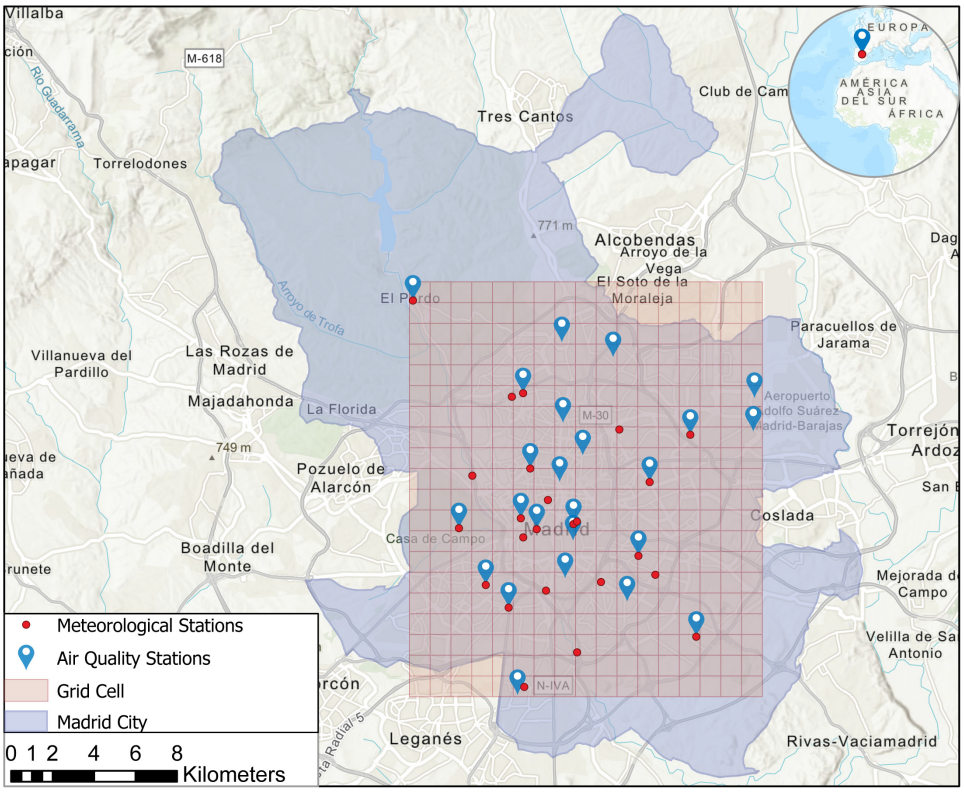

Fig. 1. Air quality stations, meteorological stations and grid cells segments on the defined area of the city of Madrid.

Table 2. Summary statistics of the periods January–June 2019 (Non-pandemic) and 2020 (pandemic) for each data type.

	Descriptors	Non-pandemic Period (2019)	Pandemic Period (2020)
Nitrogen dioxide	Mean (SD)	36.69 (30.85)	26.03 (25.35)
	Median [Min, Max]	27.0 [0.0, 328]	17.0 [0.0, 326]
UV	Mean (SD)	15.83 (30.27)	—
	Median [Min, Max]	1.0 [0.0, 199]	—
Wind speed	Mean (SD)	1.41 (1.11)	1.31 (1.05)
	Median [Min, Max]	1.14 [0.0, 8.75]	1.05 [0.0, 8.97]
Wind direction	Mean (SD)	167.80 (105.72)	140.82 (98.35)
	Median [Min, Max]	182.0 [0.0, 359]	135.0 [0.0, 359]
Temperature	Mean (SD)	13.38 (8.09)	13.63 (7.6)
	Median [Min, Max]	12.5 [−55.0, 47.3]	12.6 [−55.0, 44.6]
Humidity	Mean (SD)	48.73 (21.60)	60.76 (22.77)
	Median [Min, Max]	47.0 [−25, 100]	62.0 [−25, 100]
Pressure	Mean (SD)	943.3 (34.91)	940.62 (63.28)
	Median [Min, Max]	945.0 [0.0, 962.0]	945.0 [0.0, 1073.0]
Solar irradiance	Mean (SD)	220.73 (301.06)	191.95 (279.83)
	Median [Min, Max]	11.0 [0.0, 1103.0]	9.0 [0.0, 1113.0]
Precipitation	Mean (SD)	0.03 (0.41)	0.03 (0.27)
	Median [Min, Max]	0.0 [0.0, 30.4]	0.0 [0.0, 13.5]

relative humidity, barometric pressure, solar irradiance and precipitation. Both datasets have an hourly rate. Table 2 shows summary statistics of each data type during both periods. The datasets and the code implemented are available at the following link (https://bit.ly/3d5tatN). The selection of the two periods is based on

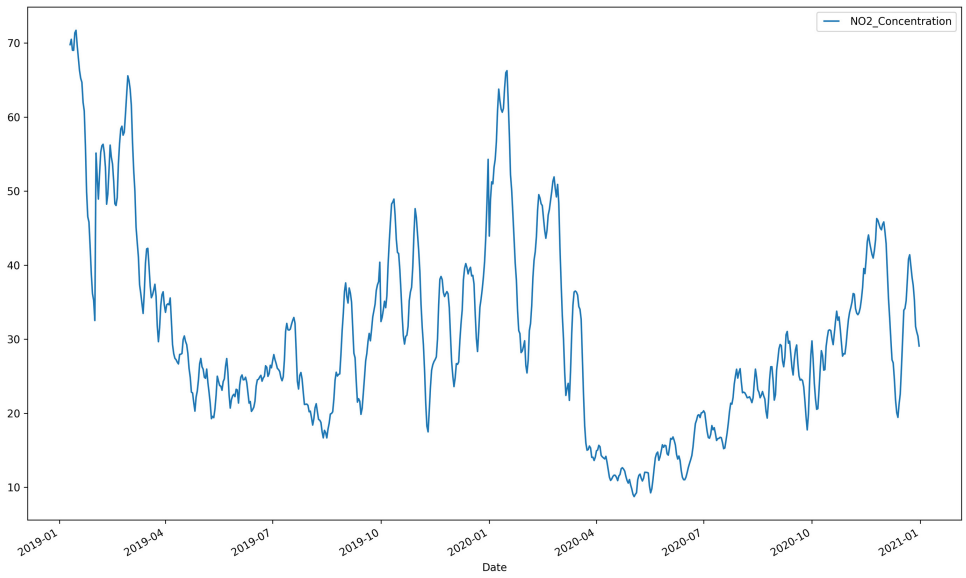

Fig. 2. Concentration of NO_2 during the period of 2019–2020.

the output illustrated in Fig. 2, which presents the concentration of NO_2 (average daily concentration from all the stations visualized by the moving average smoothing technique) during 2019–2020.

ConvLSTM is a type of recurrent neural network, similar to LSTM with only one difference — convolution operations are used instead of internal matrix multiplications.[30] Having convolutional structures in the input-to-state and state-to-state transitions makes it possible to consider a spatial factor in addition to a temporal factor. In this analysis, the model architecture was constructed by stacking several convLSTM layers which were combined with dropout and batch normalization layers and the entire network was finished with a Conv2D layer.

3. Experiments and Results

This section presents a detailed description of the experiments implemented and the results obtained. The overall workflow of the analysis is presented in Fig. 3. It can be seen that the workflow consists of two segments formed on the basis of the software applied: ArcGIS Pro and Google Colab. The first step was to create grid cells in a given area. In this study, the part of Madrid within the following extent was selected as the given area: Top — 4,486,449.725263 m; Bottom — 4,466,449.725263 m; Left — 434,215.234430 m; Right — 451,215.234430 m. The grid cells were created using the Fishnet tool (https://bit.ly/3vUpBxj). Both cell size width and height were 1000 m. There are a total of 340 cells (20 by 17), which cover 340 km^2 or 56.27%

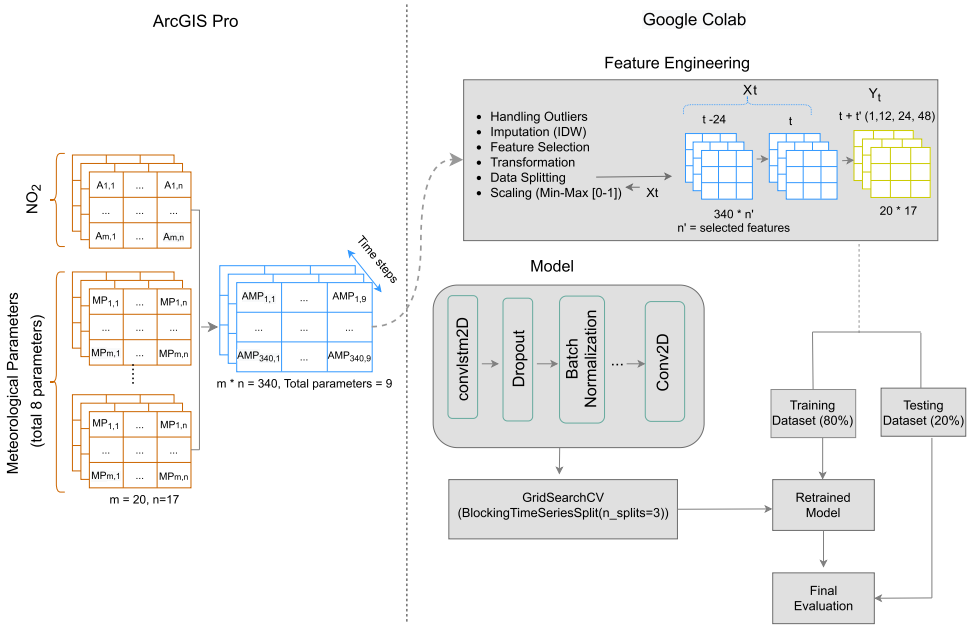

Fig. 3. The detailed workflow of the analysis.

of the total area of the city of Madrid. The logic behind selecting this area was to select a minimum extension that included all air quality control stations to achieve higher accuracy. The value of each cell includes the values of NO_2 and meteorological attributes obtained from assigned stations at a certain time. A zero value was assigned to the cells that did not include any stations. After generating grid cells, the next step was to export the output as Comma Separated Values (CSV) files, which were used as input in further stages of the analysis. Overall, 4344 and 4368 CSV files were generated corresponding to every hour during January-June 2019 and January-June 2020, respectively.

The above process is the first segment. The second segment presents machine learning techniques applied to the data obtained. In the figure, it can be seen that the machine learning process begins with feature engineering. The following feature engineering techniques were applied in this work: (a) handling outliers: Before outliers can be processed they must be detected, and summary statistics from Table 2 can help detect them. The minimum values of humidity and temperature show that they are outliers. Temperatures below -3° (https://bit.ly/3gOxLD0) and humidity with negative values were considered outliers and replaced with the average of the previous and next values. (b) imputation: As already mentioned, there are 24 air pollution control stations and 26 meteorological stations, which means that of 340 cells around 8% have data. In order to solve the problem related to missing data, inverse distance weighting was applied. (c) feature selection: A significant step in machine learning analysis is the feature selection process. As already mentioned, only nine features were included. From Table 2, it can be seen that no ultraviolet radiation was recorded for the pandemic period. Therefore, this feature has been removed. Also, regarding precipitation, it was found out that around 99% of data were 0, so this feature was also eliminated. For remaining variables, the mutual information (MI) technique was implemented.[31] It calculates the mutuality between additional datasets and the target dataset (NO_2). The formula to calculate MI is presented as follows:

$$\begin{aligned} \mathrm{MI}(x;y) &= \iint P(x_i,y)\log\frac{P(x_i,y)}{P(x_i)P(y)}dx_i dy \\ &= H(x) - H(x|y), \end{aligned} \tag{1}$$

where $P(x_i, y)$ is the joint probability distribution of two variables, $P(x_i)$ and $P(y)$ are marginal distributions, $H(x)$ is the entropy for x, and $H(x|y)$ is the conditional entropy.

Figure 4 shows the feature importance scores of 5 additional datasets based on MI. It should be noted that the wind direction is not included in Fig. 4. The wind direction is circular data and needs to be converted for later use (the details are mentioned below). Looking at Fig. 4, it can be seen that the wind speed has a higher score compared to other variables, so it was chosen for further analysis with the wind direction, given their interconnection. (d) transformation: Regarding wind direction,

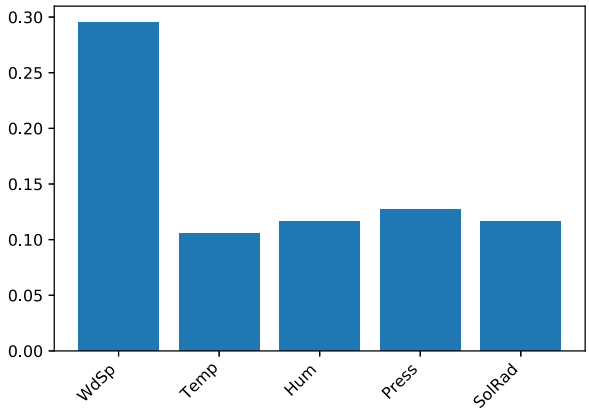

Fig. 4. The feature importance scores based on MI.

which is circular data, it was converted in categorical data with the following categories: north, east, south, west, southwest, northeast, southeast, northwest, and later by implementing One Hot Encoder[32] it was used for predictive analysis. (e) data splitting: In this step, independent and dependent datasets were generated based on the time granularity (to predict NO_2 in t' hours based on the data for the previous 24 h, where $t' \in \{1, 12, 24, 48\}$). (f) scaling: The ranges of each feature can be different from each other. In order to achieve better results, scaling can be a very useful technique. The input data was normalized with min–max (0–1) normalization.

After preprocessing the data with the aforementioned techniques, the dataset was split into training and testing sets with a fraction of 0.2: 80% training set and 20% testing set. Apart from this split, GridSearchCV with blocking time series split was applied on the training set for parameter optimization. Blocking time series split was chosen instead of cross-validation because it considers the time series aspect and prevents leakage from one set to another. In order to reduce computation time for parameter optimization, GridSearchCV was applied on the sampled dataset, which was generated by sampling the data every 6 h. Table 3 shows optimized parameters with the options that were tried, and the option that was finally selected is indicated in bold.

After parameter optimization, it was decided that the finalized model should be implemented in the following scenarios: (a) Including all features; and (b) Including

Table 3. Parameter optimization with GridSearchCV.

Parameters	Options
Number of filters	8, **16**, 32
Kernel size	(3, 3), (5, 5), (7, 7), **(9, 9)**
Dropout rate	**0.2**, 0.3, 0.5
Optimizer	RMSprop, **Adam**
Kernel Initializer	Uniform, normal, glorot_normal, **glorot_uniform**

Table 4. RMSE of ConvLSTM and LSTM for the periods January–June 2019 (Non-pandemic) and 2020 (Pandemic) in terms of features combination and time granularities.

		All Features		Selected Features	
Method	Hours	Non-pandemic Period (2019)	Pandemic Period (2020)	Non-pandemic Period (2019)	Pandemic Period (2020)
ConvLSTM	1	13.46	11.55	1.46	1.22
	12	21.05	25.11	2.09	1.63
	24	26.02	20.17	2.21	1.58
	48	25.23	26.15	2.12	1.62
LSTM	1	27.94	32.16	1.51	1.46
	12	34.01	32.12	2.89	2.52
	24	33.69	32.0	2.57	2.00
	48	33.8	32.16	2.52	2.29

only selected features (NO_2, wind speed, and wind direction). RMSE was taken as an evaluation metric, and EarlyStopping callback was used to prevent overfitting.

Table 4 presents the executed results. It can be seen that the feature selection significantly improved the results, given the fact that often the presence of many features prevents the model from effectively generalizing to the data due to the curse of dimensionality, which also exists in this work. Regarding machine learning algorithms, it should be noticed that convLSTM outperformed LSTM, particularly in the first scenario compared to the second scenario the differences between the two models are significant. Regarding the two different periods, it should be noted that the pandemic period exceeds the non-pandemic period in most sub-scenarios; however, the difference is not large. Although the variance of a pandemic year is smaller than for a non-pandemic year, nevertheless, the algorithms are trained and tested for each period separately, which means that the models during training most likely learn and generalize all existing patterns for both periods. In terms of time granularity, 1-h granularity outperformed other granularities in all sub-scenarios, but this trend does not maintain for other time granularities, which can be related to the selection of the historical time lags.[33] Based on the above findings, it can be concluded that analysis involved feature selection yields greater accuracy. ConvLSTM being able to convey spatial information in addition to temporal information has a clear advantage over LSTM, which can also be noted from the final results.

4. Conclusions and Future Work

Taking into account the fact that the concentration of NO_2, in addition to seasonal changes, has undergone the impact of COVID-19, this work uses ConvLSTM to compare the results for the pandemic and non-pandemic periods. The analysis was carried out for different time resolutions (1-h, 12-h, 24-h and 48-h) with different feature combinations. The strength of the chosen algorithm is that, despite the temporal prediction, it can also perform higher predictions in the spatial plane.

RMSE was chosen as the assessment metric. The final results showed that the proposed model outperformed the LSTM, which can be explained by the ability of the convLSTM to generalize and transfer the spatiotemporal information. In terms of datasets, the analyzes performed with selected features surpassed the results performed with all features due to the problem of high dimensionality. Regarding future work, there are several aspects to be considered. It may be useful to include traffic data since it plays a decisive role in raising the level of NO_2. Another extension of this work could be to apply the analysis to another city and compare the results for different case studies.

Acknowledgments

Ditsuhi Iskandaryan has been funded by the predoctoral programme PINV2018—Universitat Jaume I (PREDOC/2018/61). S.T. has been funded by the Juan de la Cierva — Incorporacin postdoctoral programme of the Ministry of Science and Innovation — Spanish government (IJC2018-035017-I). This work has been funded by the Generalitat Valenciana through the Subvenciones para la realización de proyectos de I+D+i desarrollados por grupos de investigación emergentes program (GV/2020/035).

References

1. I. H. Fong, T. Li, S. Fong, R. K. Wong and A. J. Tallón-Ballesteros, Predicting concentration levels of air pollutants by transfer learning and recurrent neural network, *Knowl. Based Syst.* **192** (2020) 105622.
2. W. Zhai and C. Cheng, A long short-term memory approach to predicting air quality based on social media data, *Atmos. Environ.* **237** (2020) 17411.
3. W. Yang, M. Deng, F. Xu and H. Wang, Prediction of hourly PM2.5 using a space-time support vector regression model, *Atmos. Environ.* **181** (2018) 12–19.
4. J. Ma, Y. Ding, V. J. Gan, C. Lin and Z. Wan, Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM, *IEEE Access* **7** (2019) 107897–107907.
5. S. Kim, S. Hong, M. Joh and S.-k. Song, arXiv:1711.02316.
6. Z. Yuan, X. Zhou and T. Yang, Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data, in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, 2018, pp. 984–992.
7. V.-D. Le, T.-C. Bui and S. K. Cha, arXiv:1911.12919.
8. Air pollution. https://bit.ly/3gScSXF. Accessed 24 June 2021.
9. W. Liu, X.-G. Yue and P. B. Tchounwou, Response to the COVID-19 epidemic: The chinese experience and implications for other countries (2020).
10. Z. S. Venter, K. Aunan, S. Chowdhury and J. Lelieveld, COVID-19 lockdowns cause global air pollution declines, *Proc. Natl. Acad. Sci.* **117**(32) (2020) 18984–18990.
11. J. M. Baldasano, COVID-19 lockdown effects on air quality by NO2 in the cities of Barcelona and Madrid (Spain), *Sci. Total Environ.* **741** (2020) 140353.
12. D. Iskandaryan, F. Ramos and S. Trilles, Features exploration from datasets vision in air quality prediction domain, *Atmosphere* **12**(3) (2021) 312.

13. S. Khomenko, M. Cirach, E. Pereira-Barboza, N. Mueller, J. Barrera-Gómez, D. Rojas-Rueda, K. de Hoogh, G. Hoek and M. Nieuwenhuijsen, Premature mortality due to air pollution in European cities: A health impact assessment, *Lancet Planet. Health* **5**(3) (2021) e121–e134.
14. Portal de Datos Abiertos del Ayuntamiento de Madrid. https://bit.ly/2TZzwEo. Accessed 24 June 2021.
15. Z. Li, S. H. L. Yim and K.-F. Ho, High temporal resolution prediction of street-level PM2.5 and NOx concentrations using machine learning approach, *J. Clean. Prod.* (2020) 121975.
16. L. Goulier, B. Paas, L. Ehrnsperger and O. Klemm, Modelling of urban air pollutant concentrations with artificial neural networks using novel input variables, *Int. J. Environ. Res. Public Health* **17**(6) (2020) 2025.
17. L. Li *et al.*, Cluster-based bagging of constrained mixed-effects models for high spatio-temporal resolution nitrogen oxides prediction over large regions, *Environ. Int.* **128** (2019) 310–323.
18. J. Chen *et al.*, A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide, *Environ. Int.* **130** (2019) 104934.
19. L. Munkhdalai, T. Munkhdalai, K. H. Park, T. Amarbayasgalan, E. Erdenebaatar, H. W. Park and K. H. Ryu, An end-to-end adaptive input selection with dynamic weights for forecasting multivariate time series, *IEEE Access* **7** (2019) 99099–99114.
20. J. A. Kamińska, A random forest partition model for predicting No 2 concentrations from traffic flow and meteorological conditions, *Sci. Total Environ.* **65** (2019) 475–483.
21. E. Pardo and N. Malpica, Air quality forecasting in Madrid using long short-term memory networks, in *Int. Work-Conf. Interplay between Natural and Artificial Computation* (Springer, 2017), pp. 232–239.
22. M. Krishan, S. Jha, J. Das, A. Singh, M. K. Goyal and C. Sekar, Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India, *Air Qual. Atmos. Health* **12**(8) (2019) 899–908.
23. H. Peng, A. R. Lima, A. Teakles, J. Jin, A. J. Cannon and W. W. Hsieh, Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods, *Air Qual. Atmos. Health* **10**(2) (2017) 195–211.
24. J. Zhang and W. Ding, Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong, *Int. J. Environ. Res. Public Health* **14**(2) (2017) 114.
25. K. B. Shaban, A. Kadri and E. Rezk, Urban air pollution monitoring system with forecasting models, *IEEE Sens. J.* **16**(8) (2016) 2598–2606.
26. W. Tamas, G. Notton, C. Paoli, M.-L. Nivet and C. Voyant, Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks, *Aerosol Air Qual. Res.* **16**(2) (2016) 405–441.
27. E. Debry and V. Mallet, Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM10 on the Prev'Air platform, *Atmos. Environ.* **91** (2014) 71–84.
28. C.-M. Vong, W.-F. Ip, P.-k. Wong and J.-y. Yang, Short-term prediction of air pollution in Macau using support vector machines, *J. Control Sci. Eng.* **2012** (2012) Article ID 518032, 11 pages, https://doi.org/10.1155/2012/518032.
29. W. Wang, C. Men and W. Lu, Online prediction model based on support vector machine, *Neurocomputing* **71**(4–6) (2008) 550–558.
30. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong and W.-c. Woo, arXiv:1506.04214.

31. H. Peng, F. Long and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8) (2005) 1226–1238.
32. One hot encoder. https://bit.ly/3nwzXCe. Accessed 24 June 2021.
33. H. Zheng, H. Li, X. Lu and T. Ruan, A multiple kernel learning approach for air quality prediction, *Adv. Meteorol.* **2018** (2018) Article ID 3506394, 15 pages, https://doi.org/10.1155/2018/3506394.