International Journal of Informatics, Information System and Computer Engineering

Air Quality Prediction in Smart City's Information System

Ivan Kristianto Singgih

Department of the Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon, 305-701, Republic of Korea
Correspondence: E-mail: ivanksinggih@gmail.com

ABSTRACTS

The introduction of new technology and computational power enables more data usages in a city. Such a city is called a smart city that records more data related to daily life activities and analyzes them to provide better services. Such data acquisition and analysis must be conducted quickly to support real-time information sharing and support other decision-making processes. Among such services, an information system is used to predict the air quality to ensure people's health in the city. The objective of this study is to compare various machine learning techniques (e.g., random forest, decision tree, neural network, naïve Bayes, etc.) when predicting the air quality in a city. For the comparison, we perform the removal of records with empty values, data division into training and testing datasets, and application of the k-fold cross-validation method. Numerical experiments are performed using a given online dataset. The results show that the three best methods are random forest, Gradient Boosting, and k-nearest neighbors with precision, recall, and f1-score values more than 0.63.

ARTICLE INFO

Article History:

Keywords:
Definitions of Shophouses;
Identity;
George Town;
Influence Architecture and Design.

1. INTRODUCTION

MART city integrates the physical world and the virtual world. A concept used for performing such connectivity is called digital twin that is a virtual model representing the physical world (Marr, 2020). Using such a virtual model allows monitoring the physical system, preventing problems from happening, finding new opportunities, and planning the future. The interaction is illustrated in Smart City Korea (2020) through Fig. 1. Massive IoTs, digital twins, and data hubs are utilized to generate the required information in the integration process.

In the smart city, fixed/mobile sensors are installed within the city to observe real behaviors (e.g., the people) and conduct a better operation of the virtual world. There are various subareas within the smart city, including smart mobility, smart buildings, etc. Among them, the smart environment is the one that manages air pollution control to ensure the health of the citizens (Alvear et al., 2018). The smart environment system's continuous improvement is supported by the wide use of the Internet of Things that provides better connectivity between multiple sensors located in the dispersed area and ease the air quality monitoring process (Zhang & Woo, 2020).

The existence of different air pollutants causes harm to the respiratory systems. Such air pollutants are nitrogen dioxide (NO2), carbon monoxide (CO), ozone (O3), sulphur dioxide (SO2), and particulate matter (PM). Real-time monitoring stations are built by many cities to check the air quality, then inform people when it is safe to conduct outside activities and plan better movements (Zhang & Woo, 2020). Systems for collecting and assessing air quality have been installed in several areas, e.g., Peking University (with 100 thousand data from 30 devices) (Hu et al., 2019), Christchurch that is a part of IBM's smart city initiatives (Marek et al., 2017), Los Angeles (Wu et al., 2017), etc. Various information systems are implemented for supporting the data collection and air quality information transfer to the people. An example of the information system used for air quality management in Los Angeles is presented in Fig. 2 (Wu et al., 2017). In this implementation, a remote data collection can be performed using a smartphone. The collected image data are analyzed using machine learning to calculate the particle concentration in the air and evaluate the air quality

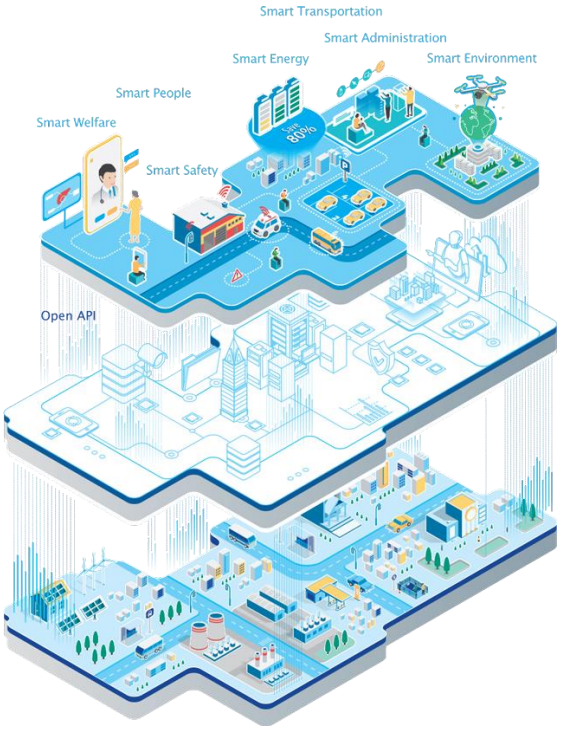

Fig. 1. Interaction between physical and virtual world in smart city concept

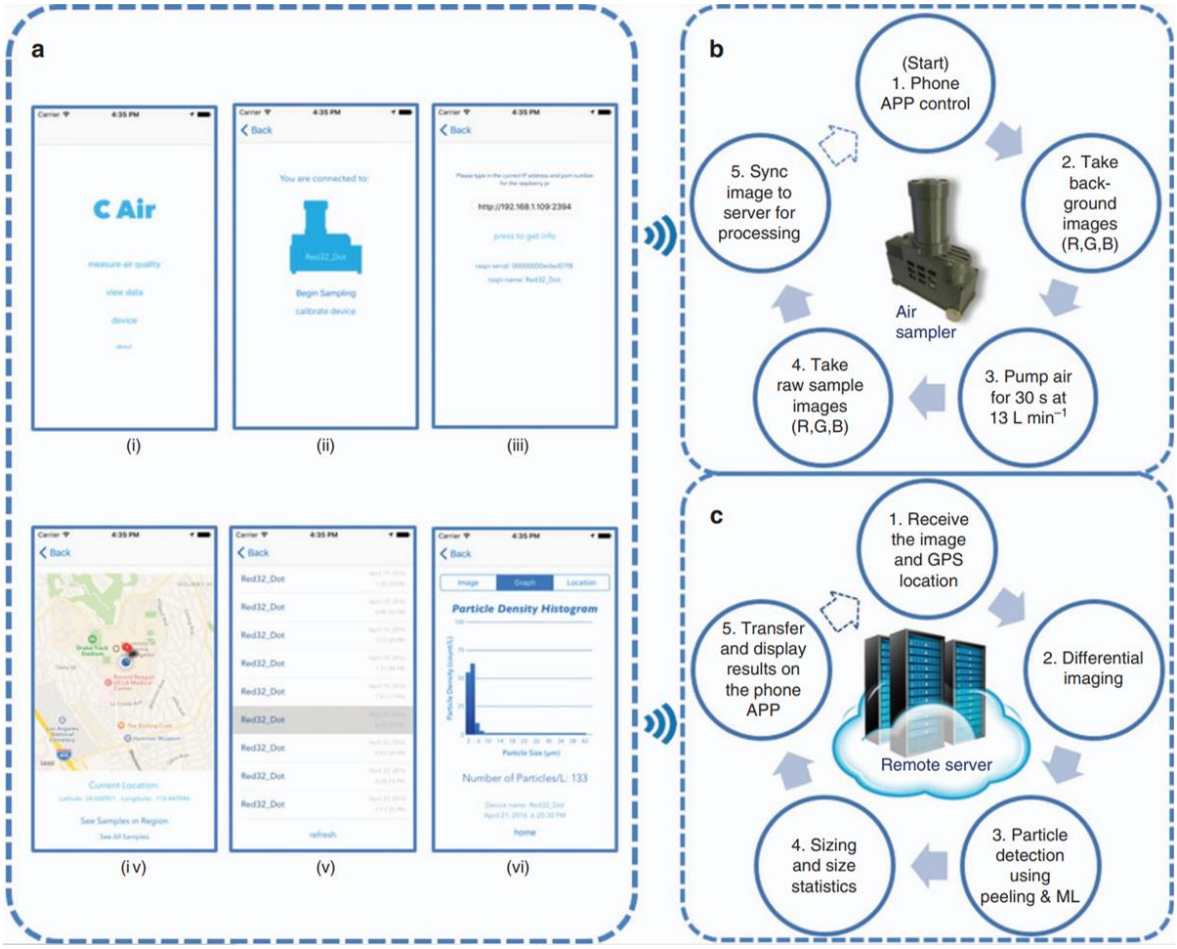

Fig. 2. Information system for air quality management in Los Angeles

2. METHODOLOGY

Information system in smart cities has a component for the data acquisition and a server to store and process the obtained data. The good adoption of technologies for data collection and computation determines the success of smart city developments (Marek et al., 2017). Given that multiple sources of data (e.g., open data, online data sharing) emerge, improving the information system interoperability, including how to utilize existing data, is a great challenge to be solved in smart city projects. Many smart city initiatives have been started. One of them is ERA-PLANET, a wide European network that consists of 118 researchers in 35 institutions and 18 countries (Tsinganos et al., 2017).

The architecture of the information system related to the air quality sensing process performs the following tasks (Alvear et al., 2018):

1) Sampling

The sampling task measures the pollutant in the air that includes the calibration process. By performing such sampling with many mobile sensors, the problem of sampling error can be handled because of the possibility of considering redundant data and statistical analysis.

2) Data filtering

Through the filtering process, redundant data and wrong measurements are removed.

3) Data transfer

The collected data are uploaded from the sensors to the cloud (servers). The upload process is managed based on some IoT protocols.

4) Data processing

The observed data are processed to obtain a conclusion on the air quality. Through this process, a pollution distribution map is generated.

5) Presentation of the analysis result

The results can be presented as a graphical map.

The architecture itself can be divided into three layers (Schürholz et al., 2020):

1) Data layer

The data layer contains a database of historical data and prediction data.

2) Logic layer

The logic layer converts the input data before being used in the analysis and performs the prediction process.

3) Visualization layer

The visualization layer passes the information to be visualized in the end-user devices.

The introduction of inexpensive small sensors allows retrieving a huge amount of data in real-time fashion (Hu et al., 2019). Effective machine learning techniques are implemented in this study to perform such a real-time air quality assessment. The machine learning techniques used in

this study are listed in Table I. The selected methods have been proven to perform well for predicting air quality purposes. Studies that used each method (or its variants) are presented as well.

3. RESULTS AND DISCUSSION

We use the Python sklearn library (Pedregosa et al., 2011) to implement the machine learning techniques. The code is written using the Visual Studio 2019 Community platform. A partial view of the code is presented in Fig. 3. Air quality prediction data presented in (Bhat, 2020) is solved. Among 26,6219 data, we remove records with any empty values and obtain 4,646 records to be used in our study. We exclude the location and time stamp fields from the observed data. The preprocessed data are stored in an Excel input file and is imported into Python. Libraries for performing the calculations and generating a graphical representation of the results are used.

The dependent variable is the air quality with the following values: Severe, Very Poor, Poor, Moderate, Satisfactory, Good. The independent variables are:

1) PM2.5
   PM is the abbreviation of particulate matter that includes potential harmful compounds, which can reach human respiratory systems (Chaparro et al., 2020). PM2.5 refers to cases of air particles with the mass per cubic meter less than 2.5 µm.
2) PM_{10}
3) NO
   NO refers to nitrogen oxide.
4) NO_2
   NO2 refers to nitrogen dioxide.
5) NOx
   NOx is the total amount of NO and NO_2.
6) NH3
   NH3 refers to ammonia.
7) CO
   CO refers to carbon monoxide.
8) SO_2
   SO_2 refers to sulfur dioxide.
9) O_3
   O_3 refers to ozone.
10) Benzene
11) Toluene
12) Xylene

Table 1. Used machine learning techniques in this study

Machine Learning Technique	Reference
adaptive boosting (AB)	Liu and Chen (2020)
linear classifiers with stochastic gradient descent training (SGD)	Ganesh et al. (2017)
neural network (multi-layer perceptron [a]) (NNMLP)	Ganesh et al. (2017), Gu et al. (2020), Sun et al. (2020), Wang et al. (2017), Zhao et al. (2020)
Gradient Boosting (GB)	Zhang et al. (2019), Zhang et al. (2019b), Liu et al. (2019), Yu et al. (2016), Feng et al. (2018)
random forest (RF)	Liu et al. (2019), Yu et al. (2016), Feng et al. (2018)
k-nearest neighbors (KNN)	Zhao et al. (2020)
decision tree (CART)	Zhang et al. (2019b)
Naive Bayes (Gaussian [a]) (NB)	Feng et al. (2018), Melgarejo et al. (2015)
support vector machine (C-Support Vector [a]) (SVM)	Ganesh et al. (2017), Gu et al. (2020), Liu et al. (2019), Melgarejo et al. (2015), Dun et al. (2020)

[a]Specific one considered in this study.

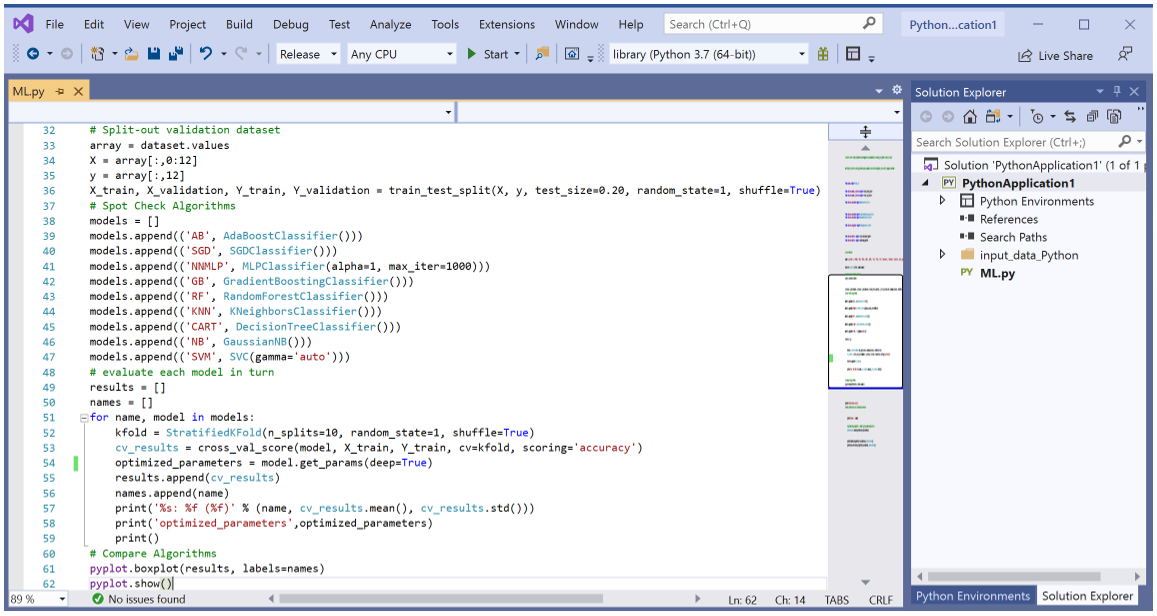

Fig. 3. A partial view of the code

Analysis steps performed in this study are:

1) Dividing the dataset into training and testing data

In our implementation, the percentage of testing data is set into 20%. The data is shuffled before the division.

2) Testing the accuracy of each technique using the k-fold cross-validation

The number of used splits is 10. The training data is shuffled as well before performing the testing.

3) Fitting the testing data

Precision, recall, f1-score, and support metrics are measured for each technique. True positive, false positive, false negative, and true negative cases are observed to calculate such values. The cases are defined in Fig. 4 based on the comparison between results concluded by the test and the real data (Parikh et al., 2008). Definition and formula of precision, recall, f1-score, and support are presented in Table II. Machine learning methods that can predict the air quality better are the ones with higher scores.

Actual State / Test Result	Right	Wrong
Positive	True Positive (TP)*	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)*

*Correct outcomes

Fig. 4. True positive, false positive, false negative, and true negative cases

Result of accuracy testing using k-fold cross validation is presented using boxplots in Fig. 5. Three techniques that have the best accuracy are RF, GB, and KNN. These three methods have good average accuracy and a smaller deviation in the accuracy calculation when considering different training and testing datasets, compared with the others. The worst accuracies are obtained by SGD and SVM methods.

Table 2. Definition and formula of precision, recall, f1-score, and support

Metric	Definition	Formula
Precision	Total number of retrieved data that are relevant/total number of retrieved data (Ting, 2011)	TP / (TP + FP) (Jiang et al., 2017)
Recall	Total number of retrieved data that are relevant/total number of relevant data in the database (Ting, 2011)	TP / (TP + FN) (Jiang et al., 2017)
F1-score	A weighted value obtained from the precision and recall values with 1 as its best value and 0 as its worst	2 * precision * recall / (precision + recall) (Yuan et al., 2020)
Support	Number of occurrences of each class in y_true	-

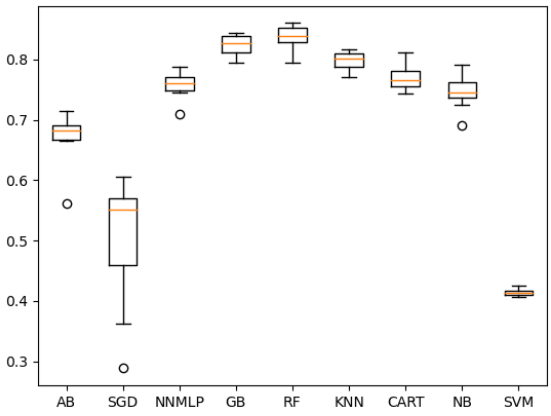

Fig. 5. Accuracy comparison of the machine learning techniques

Table 3. Average value of each metric using the testing data

Machine Learning Technique	Prediction	Recall	F1-score
AB	0.36	0.52	0.41
SGD	0.38	0.41	0.39
NNMLP	0.73	0.64	0.66
GB	0.82	0.77	0.79
RF	0.81	0.76	0.78
KNN	0.81	0.77	0.79
CART	0.75	0.71	0.73
NB	0.65	0.70	0.67
SVM	0.22	0.19	0.14

Table 4. Detailed metric values of rf method

Class	Prediction	Recall	F1-score	Support
Good	0.84	0.64	0.73	67
Satisfactory	0.78	0.84	0.81	269
Moderate	0.83	0.86	0.84	383
Poor	0.75	0.69	0.72	110
Very poor	0.81	0.81	0.81	77
Severe	0.85	0.71	0.77	24

Table 5. Detailed metric values of gb method

Class	Prediction	Recall	F1-score	Support
Good	0.85	0.66	0.74	67
Satisfactory	0.80	0.84	0.82	269
Moderate	0.84	0.86	0.85	383
Poor	0.72	0.68	0.70	110
Very poor	0.81	0.79	0.80	77
Severe	0.90	0.79	0.84	24

Table 6. Detailed metric values of knn method

Class	Prediction	Recall	F1-score	Support
Good	0.70	0.63	0.66	67
Satisfactory	0.76	0.81	0.78	269
Moderate	0.84	0.84	0.84	383
Poor	0.77	0.73	0.75	110
Very poor	0.86	0.81	0.83	77
Severe	0.90	0.79	0.84	24

The fitting results of the testing data are presented in Tables III-VI. In Table III, the average value of each metric calculated from all classification class is presented. The detailed metric values for the three best techniques are presented in Tables IV-VI. In these tables, evaluation is performed for each class (Severe, Very Poor, Poor, Moderate, Satisfactory, Good). It can be seen that the value of each metric is similar for each class when a certain method is implemented.

4. CONCLUSION

In this study, we implement several machine learning techniques to predict air quality as part of the smart city's information system. Based on the numerical experiments, random forest, Gradient Boosting, and k-nearest neighbors have the best accuracies. Future studies must assess whether it is necessary to include all input values in the models and consider how to deal with incomplete records.

REFERENCES

Alvear, O., Calafate, C. T., Cano, J.-C., & Manzoni, P. (2018). Crowdsensing in smart cities: Overview, platforms, and environment sensing issues. *Sensors*, 18(2), 460.

Bhat, N. (2020, October 1). *Air quality level of different cities in India (2015-2020)*. Kaggle Dataset. https://www.kaggle.com/nareshbhat/air-quality-pre-and-post-covid19-pandemic

Chaparro, M. A. E., Chaparro, M. A. E., Castañeda-Miranda, A. G., Marié, D. C., Gargiulo, J. D., Lavornia, J. M., Natal, M., & Böhnel, H. N. (2020). Fine air pollution particles trapped by street tree barks: In situ magnetic biomonitoring. *Environmental Pollution*, 266(1), Article 115229.

Dun, M., Xu, Z., Chen, Y., & Wu, L. (2020). Short-term air quality prediction based on fractional grey linear regression and support vector machine. *Mathematical Problems in Engineering*, 2020(1), Article 8914501.

Feng, C., Tian, Y., Gong, X., Que, X., & Wang, W. (2018). MCS-RF: mobile crowdsensing–based air quality estimation with random forest. *International Journal of Distributed Sensor Networks*, 14(10), 1–15.

Ganesh, S. S., Arulmozhivarman, P., & Tatavarti, R. (2017). Forecasting air quality index using an ensemble of artificial neural networks and regression models. *Journal of Intelligent Systems*, 28(5), 893–903.

Gu, K., Zhou, Y., Sun, H., Zhao, L., & Liu, S. (2020). Prediction of air quality in Shenzhen based on neural network algorithm. *Neural Computing and Applications*, 32(7), 1879–1892.

Hu, Z., Bai, Z., Bian, K., Wang, T., & Song, L. (2019). Real-time fine-grained air quality sensing networks in smart city: Design, implementation, and optimization. *IEEE Internet of Things Journal*, 6(5), 7526–7542.

Jiang, C., Liu, Y., Ding, Y., Liang, K., & Duan, R. (2017). Capturing helpful reviews from social media for product quality improvement: A multi-class classification approach. *International Journal of Production Research*, 55(12), 3528–3541.

Liu, H., & Chen, C. (2020). Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: A case study in China. *Journal of Cleaner Production*, 265(1), Article 121777.

Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Applied Sciences*, 9(19), 4069.

Marek, L., Campbell, M., & Bui, L. (2017). Shaking for innovation: The (re)building of a (smart) city in a post disaster environment. *Cities*, 63(1), 41–50.

Marr, B. (2020, October 3). *What is digital twin technology - And why is it so important?* Forbes. https://www.forbes.com/sites/bernardmarr/2017/03/06/what-is-digital-twin-technology-and-why-is-it-so-important/#3388188e2e2a

Melgarejo, M., Parra, C., & Obregón, N. (2015). Applying computational intelligence to the classification of pollution events. *IEEE Latin America Transactions*, 13(7), 2071–2077.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 56(1), 45–50.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(1), 2825–2830.

Schürholz, D., Kubler, S., & Zaslavsky, A. (2020). Artificial intelligence-enabled context-aware air quality prediction for smart cities. *Journal of Cleaner Production*, 271(1), 121941.

Smart City Korea. (2020, October 3). *Introducing Core 1, 2, and 3 smart city innovation growth engines*. Ministry of Land, Infrastructure and Transport. https://smartcity.go.kr/en/

Sun, X., Xu, W., Jiang, H., & Wang, Q. (2020). A deep multitask learning approach for air quality prediction. *Annals of Operations Research*.

Ting, K. M. (2011). Precision and recall. In *Encyclopedia of machine learning*, 781-781. Springer, Boston, MA.

Tsinganos, K., Gerasopoulos, E., Keramitsoglou, I., Pirrone, N., & The ERA-PLANET Team. (2017). ERA-PLANET, a European network for observing our changing planet. *Sustainability*, 9(6), 1040.

Wang, J., Zhang, X., Guo, Z., & Lu, H. (2017). Developing an early-warning system for air quality prediction and assessment of cities in China. *Expert Systems with Applications*, 84(1), 102–116.

Wu, Y.-C., Shiledar, A., Li, Y.-C., Wong, J., Feng, S., Chen, X., Chen, C., Jin, K., Janamian, S., Yang, Z., Ballard, Z. C., Göröcs, Z., Feizi, A., & Ozcan, A. (2017). Air quality monitoring using mobile microscopy and machine learning. *Light: Science & Applications*, 6(1), 17046.

Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. A. (2016). RAQ–A random forest approach for predicting air quality in urban sensing systems. *Sensors*, *16*(1), Article 86. https://doi.org/10.3390/s16010086

Yuan, J., Zhang, L., Guo, S., Xiao, Y., & Li, Z. (2020). Image captioning with a joint attention mechanism by visual concept samples. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *16*(3), Article 83. https://doi.org/10.1145/3394955

Zhang, D., & Woo, S. S. (2020). Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network. *IEEE Access*, *8*(1), 89584–89594. https://doi.org/10.1109/ACCESS.2020.2993547

Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., Wang, Q., & Huang, L. (2019a). A predictive data feature exploration-based air quality prediction approach. *IEEE Access*, *7*(1), 30732–30743. https://doi.org/10.1109/ACCESS.2019.2897754

Zhang, Y., Zhang, R., Ma, Q., Wang, Y., Wang, Q., Huang, Z., & Huang, L. (2019b). A feature selection and multi-model fusion-based approach of predicting air quality. *ISA Transactions*, *100*(1), 210–220. https://doi.org/10.1016/j.isatra.2019.11.023

Zhao, Z., Qin, J., He, Z., Li, H., Yang, Y., & Zhang, R. (2020). Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China. *Environmental Science and Pollution Research*, *27*(23), 28931–28948. https://doi.org/10.1007/s11356-020-08948-1

Zhao, X., Song, M., Liu, A., Wang, Y., Wang, T., & Cao, J. (2020). Data-driven temporal-spatial model for the prediction of AQI in Nanjing. *Journal of Artificial Intelligence and Soft Computing Research*, *10*(4), 255–270. https://doi.org/10.2478/jaiscr-2020-0017