ICIC Express Letters
Part B: Applications
Volume **14**, Number **10**, October **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **1021–1028**

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR PREDICTION OF AIR POPULATION

Sewon Park, Dongyeol Baek, Insoo Choi and Gun Ho Lee

Department of Industrial Information Systems Engineering
Soongsil University
369 Sangdoro, Dongjakku, Seoul 06978, Korea
{ se1.park; ins.choi }@samsung.com; dongyeol98@kdiwin.com; ghlee@ssu.ac.kr

Received December 2022; accepted March 2023

Abstract. *This study analyzes the correlation between air pollutants in Korea and in Jiangsu, Hebei, and Shandong provinces in China, which are closest to Korea. The regression models in this study predict the amount of sulfur dioxide, carbon monoxide, ozone, nitrogen dioxide, particulate matter, and ultra-particulate matter in the atmosphere of the Korean peninsula. We use linear regression, k-nearest neighbor, AdaBoost, gradient boost, random forest, bagging, and XGBoost algorithms for predictive regression models. Through feature importance, we confirm that Jiangsu's air pollutants have the most significant effect on the atmosphere of the Korean peninsula and identify the importance of other independent features. We evaluate and compare the results of six models using performance measures of R2-Score, mean squared error, root mean squared error, and mean absolute error. The model using XGBoost shows the best results.*
Keywords: Comparative analysis, Machine learning models, Air pollution, Prediction model, Ultra-particulate matter, Correlation between air pollutants

1. **Introduction.** Air pollutants cause many chronic diseases, harm the human body, and are a constant source of fear for people with chronic illnesses [1]. China grows economically as a manufacturing base for global supply but relies heavily on thermal power generation for manufacturing production. Due to the geographical location of the Korean peninsula, there is an issue that mainland China influences the concentration of air pollutants in the Korean peninsula. The Korean government issued and implemented emergency reduction measures for high-concentration particulate matter and implemented a seasonal management system to reduce the concentration of particulate matter. It is necessary to promote preemptive air pollution reduction measures, manage air pollution emission sources, and strengthen scientific investigation and analysis of air pollution. Harishkumar et al. [2] implemented a machine-learning model to predict particulate matter in Taiwan's major cities. Shin et al. [3] implemented a prediction model to predict Japan's average monthly nitrogen dioxide concentration. Chiwewe and Ditsela [4] used linear regression and neural network models as prediction models for ozone and derived 0.579 and 0.77 accuracy, respectively. According to Park and Shin [5], the maximum and average values of the ultra-particulate matter in the air in Shandong province in China were 94.62 and 160 $\mu g/m^3$, higher than those in Korea. The correlation coefficient between Korea's ultrafine particulate matter and the wind direction ratio of the west wind in Shandong province was a positive value of 0.402. The correlation coefficient between particulate matter and the westward wind ratio in China was about 0.358. They derived significant results from the correlation between ultrafine particulate matter in the atmosphere of China and Korea. They confirmed the possibility of using it as an independent variable through further research. Zhang et al. [6] surveyed Asia's total emissions and found that

DOI: 10.24507/icicelb.14.10.1021

China's emissions are significantly larger than those of other countries. They emphasized emissions from China because the emissions from China dominate the Asia pollutant outflow to the Pacific, and the increase of emissions from China is of great concern. Kumar and Pande [7] presented machine learning models to predict air pollutants in 23 Indian cities. The Gaussian Naïve Bayes model achieved the highest accuracy, while the support vector machine model exhibited the lowest accuracy. Hansun et al. [8] presented two hybrid moving average methods, namely, B-WEMA and H-WEMA, to predict the air quality index. Yang et al. [9] predicted hourly pollutant concentrations using lightweight gradient boosting model shallow machine-learning and long short-term memory neural network.

This study investigates the correlation between air pollutants in China and Korea to confirm the influence of pollutants in the air between Korea and China. We analyze the air pollution impact in Korea according to the level of air pollution in major regions of mainland China. This study builds machine-learning models that predict the number of pollutants in a unit space of the atmosphere according to meteorological conditions. We analyze the feature importance in the model's predictive model and present improved prediction models over existing techniques by considering multiple variables. We evaluate and compare the results of regression models using performance measures of R2-Score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The prediction model in this study has a high R2-Score of about 0.92.

2. **Data and Attribute.** The data used in this study primarily include air pollutants and meteorological data in Korea and China. In addition to wind speed, we add wind direction to account for China's pollution in Korea. We obtain Korean air pollutant data from the 'Public Data Portal' [10], collecting daily data from a municipality in Seoul and China air pollutant data from the 'World Air Quality Index Project' [11] for five years from 2016 to 2020. We collect China's air pollutant data from the three administrative regions of Jiangsu, Shandong, and Hebei provinces, close to Korea, with high air pollution emissions. Meteorological data are collected from the 'Meteorological Data Open Portal' from 2016 to 2020 [12] for five years. We consider sulfur dioxide (ppm), carbon monoxide (ppm), ozone (ppm), nitrogen dioxide (ppm), particulate matter ($\mu g/m^3$), ultra particulate matter ($\mu g/m^3$), average temperature (°C), daily precipitation (mm), average wind speed (m/s), and Wind Direction at Maximum wind Speed (WDMS) (deg) as input features. The target feature is six air pollutants over the Korean peninsula. The average wind speed (m/s) is the distance moved by the atmosphere in unit time, and WDMS (deg) represents the wind direction of the wind speed that blew the hardest on average for any 10 minutes of the day.

This study analyzes correlations between input features after removing missing values and outliers from all data. We present two types of predictive models with different independent features. Type 1 uses 21 independent features: 17 air pollutant data from Shandong, Jiangsu, and Hebei provinces and 4 weather data from Korea, and model 1 does not consider carbon monoxide in Shandong. Type 2 uses 10 independent features, including 6 average values of air pollutant data from three administrative districts in China and 4 types of meteorological data in Korea. We use 11687 data in Type 1 and 28633 in Type 2 after removing outliers and missing values. In this study, we use correlation coefficients to analyze whether there is a correlation between the air pollutant features of Korea and China for predictive models.

In Figure 1, the horizontal axis represents six air pollutants in the administrative region of China, and the vertical axis does six air pollutants in Korea. Figure 1 shows that the correlation coefficient between ozone in Korea and China is high at 0.6, 0.57, and 0.44 in the order of Hebei, Shandong, and Jiangsu. Particulate matter in Korea positively correlates with particulate matter, sulfur dioxide, and carbon monoxide in each administrative

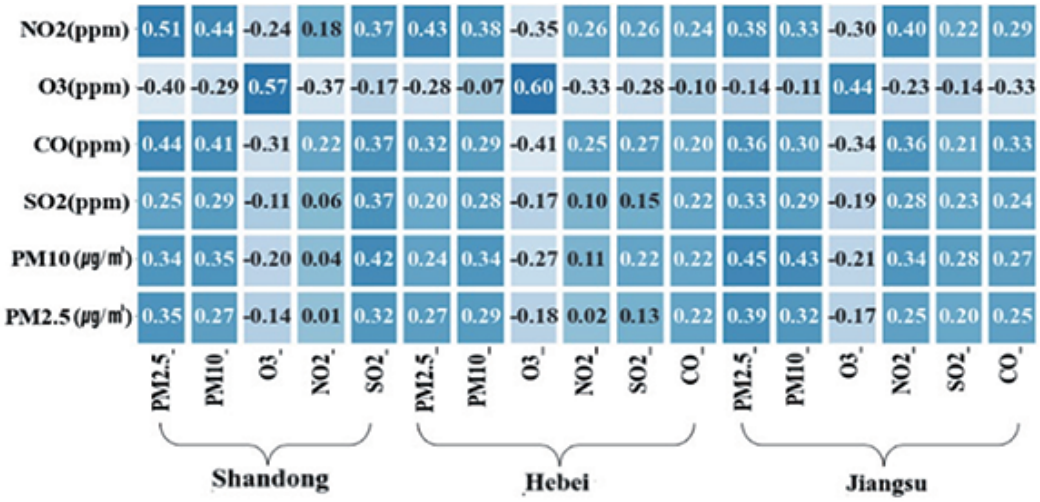

FIGURE 1. Correlation for Type 1

district in China. Figure 1 shows that overall, average air pollutants in China and Korea are correlated.

In Figure 2, the horizontal axis represents the average air pollutant of China's three administrative regions, and the vertical axis shows the air pollutant of Korea. Figure 2 shows the correlation coefficient between average ozone densities in China and Korea. The highest value is 0.58, and the correlation coefficient for particulate matter is 0.4 in both countries. Both data types showed a significant correlation between air pollutants in the two countries. We identify the feature importance of each air pollutant using the XGBoost model in Figure 3. Feature importance is used in tree-based models and determines how much a particular feature contributes to splitting the tree. Figure 3 shows the feature importance for Type 1, and Figure 4 shows the importance of China's average air pollutants as features for Type 2.

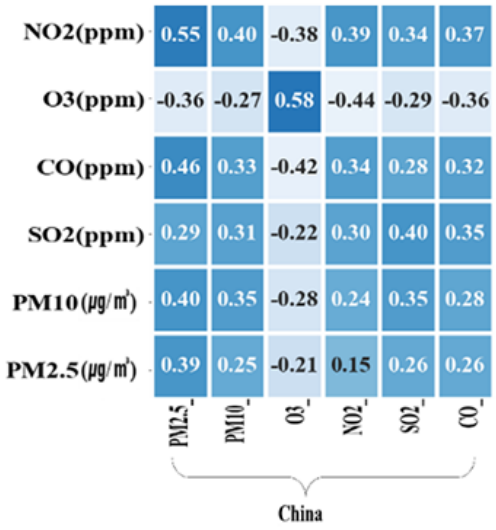

FIGURE 2. Correlation for Type 2

3. **Algorithms.** To present and compare predictive models, this study uses machine learning techniques such as multiple regression, k-nearest neighbor (k-nn) regressor, bagging, AdaBoost, gradient boost, random forest, and XGBoost. We use the k-nn regressor to find the nearest k neighbors through the distance formula due to simplicity and efficiency in the solution procedure. This study uses ensemble techniques that combine the

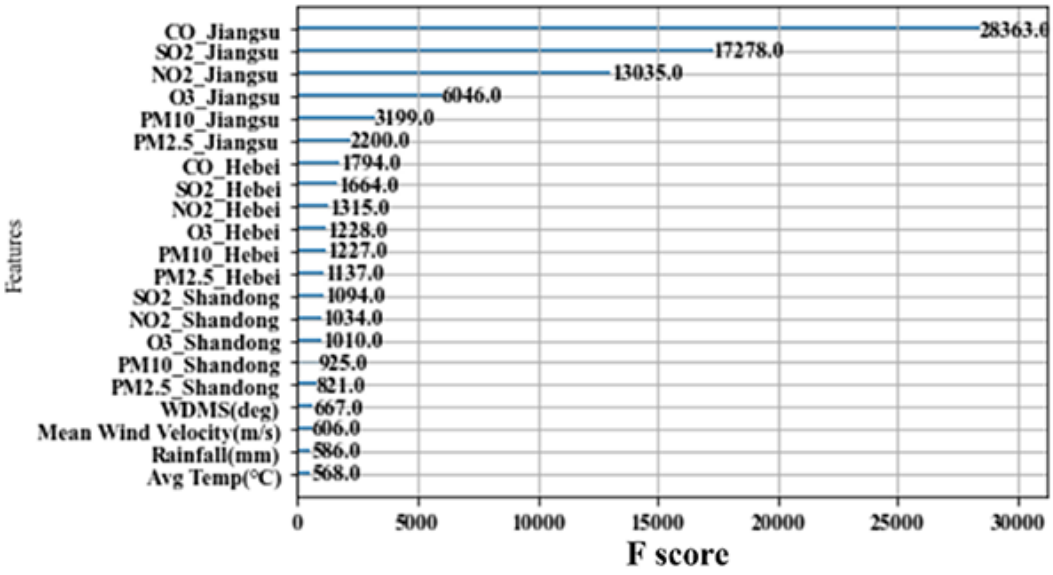

FIGURE 3. Feature importance for Type 1

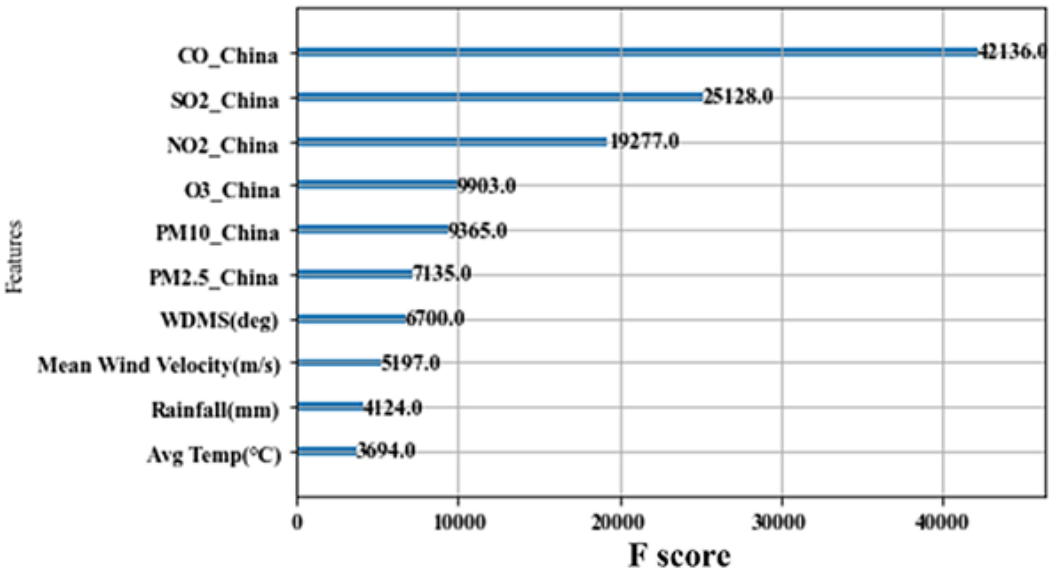

FIGURE 4. Feature importance for Type 2

predictions of several base estimators built with a given learning algorithm to improve generalizability/robustness over a single estimator [13]. We use a bagging regressor as a bootstrap aggregation that takes multiple samples and trains each model to aggregate the results. It is a method of learning the same algorithm-based decision tree in parallel based on each sample data and then combining the learning results of each model [14]. An important hyperparameter in bagging is that bagging prevents this risk through iterative training and model combination based on slightly different datasets. AdaBoost is an algorithm that sets the initial model as a weak model, uses weights every step to sequentially fit a new model that compensates for the weaknesses of the previous model, and finally creates a model obtained by linearly combining them. However, it has the disadvantage of being sensitive to outliers [15]. The learning rate is considered under the same condition as bagging for hyperparameters. Gradient boost is an ensemble technique that performs well in predicting tabular format data. Tabular format data refers to data in X-Y Grid. The gradient algorithm recognizes the weaknesses of the classifiers learned

so far and focuses on the weaknesses to compensate for them in the following learning step [16].

Random forest is a type of ensemble learning method used in regression analysis, which works by outputting average predictions from multiple decision trees constructed during training. Random forests have the advantages of reducing prediction variability, preventing overfitting, and exhibiting high accuracy [17]. We use hyperparameters to improve accuracy and prevent overfitting. XGBoost is one of the ensemble techniques that use several weak decision trees by combination. XGBoost learns quickly through parallel processing and can prevent overfitting of the model itself [18].

4. **Analysis of Models.** In this study, we build, test, and analyze the predictive models in Anaconda's jupyter notebook with version 3.7 using Intel core i5 CPU. We split the original data set into 70% of the train set and 30% of the test set for training and validation of the predictive model. In Figure 5, we analyze seven prediction models to predict four types of air pollutants. Figure 5 shows the prediction results for four types of air pollutants where the XGBoost model has the highest R2-Score for nitrogen dioxide and sulfur dioxide. However, the R2-Scores for AdaBoost and linear regression models are low, 0.006 and 0.392. For particulate matter (PM10; fine particles with a diameter of 10 μm or less) and ultra particulate matter (PM2.5; fine particles with a diameter of 2.5 μm or less), excluding linear regression and AdaBoost, the average prediction performance in the R2-Score is 0.89, and the XGBoost model is 0.92.

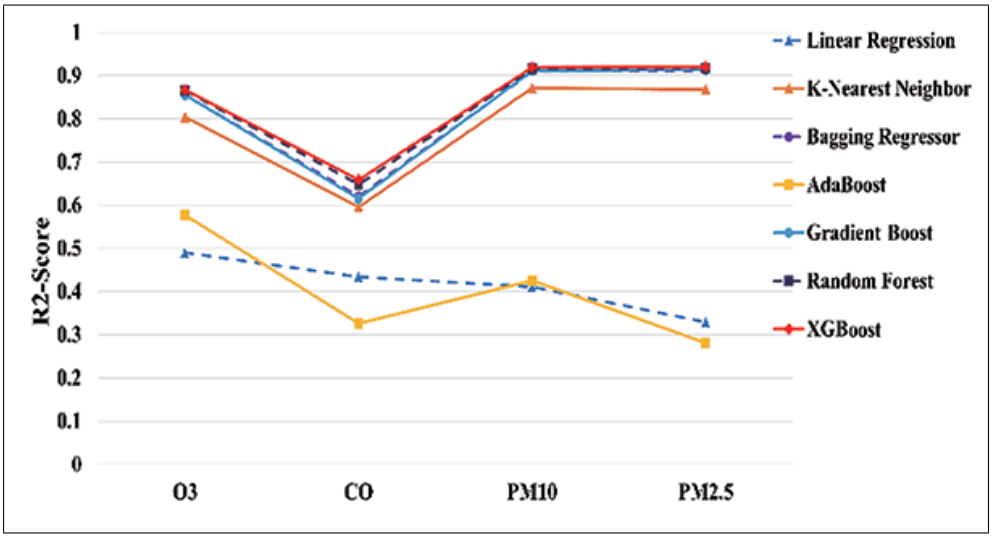

FIGURE 5. R2-Scores for Type 1

Figure 6 shows the R2-Score of seven models for Type 2. In Types 1 and 2, R2-Scores of particulate matter (PM10) and ultra-particulate matter (PM2.5) are 0.92 and 0.89, respectively. However, the R2-Scores of nitrogen dioxide and sulfur dioxide are 0.02 and 0.38, respectively, showing significantly lower prediction rates. The XGBoost model has the highest accuracy.

This study evaluates the seven predictive models using the regression models' evaluation measures, including R2-Score, MSE, RMSE, and MAE. MSE is the average of the squared values of the error, and the coefficient of determination (R2-Score) is an indicator of performance. Table 1 shows seven models' evaluation measure values for predicting particulate matter (PM10). XGBoost shows the highest R2-Score of 0.919 and has the best results for MSE, RMSE, and MAE. Table 2 also shows the values of evaluation indicators of seven models that predict the particulate matter (PM10; fine particles with a diameter of 10 μm or less) of air pollutants for Type 2. For Type 2, XGBoost shows the highest prediction accuracy, and MSE, RMSE, and MAE also show good results.

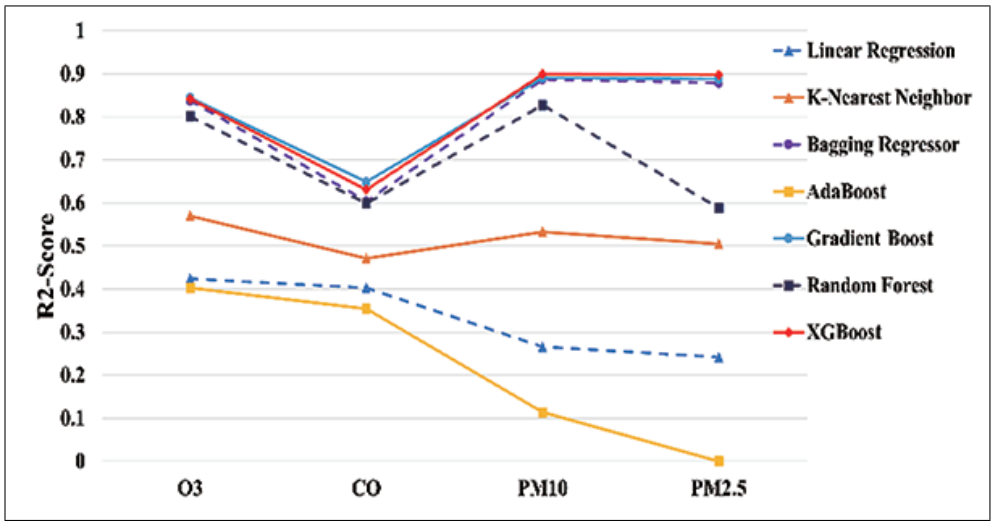

FIGURE 6. R2-Scores for Type 2

TABLE 1. Comparison of prediction results for Type 1

Measure / Algorithm	MSE	RMSE	MAE	R2-Score
Linear regression	315.071	17.750	12.937	0.411
k-nn	68.828	8.296	5.693	0.871
Random forest	45.212	6.724	4.755	0.916
Bagging	48.014	6.930	4.948	0.910
XGBoost	43.368	6.586	4.685	0.919
Gradient boost	47.750	6.910	4.932	0.911
AdaBoost	306.664	17.512	14.812	0.426

TABLE 2. Comparison of prediction results for Type 2

Measure / Algorithm	MSE	RMSE	MAE	R2-Score
Linear regression	328.2726	18.1183	13.4937	0.2663
k-nn	209.113	14.461	10.090	0.533
Random forest	78.755	8.874	6.476	0.824
Bagging	51.576	7.182	5.072	0.885
XGBoost	49.882	7.063	5.123	0.889
Gradient boost	78.481	8.859	6.537	0.825
AdaBoost	396.284	19.907	16.774	0.114

Figure 7 shows a bar graph of each model's learning time. The ensemble model's learning time varies depending on the number of weak learners. In both Type 1 and Type 2, we use 500 learners for AdaBoost, ten models for bagging, 100 trees for the random forest, 500 trees for gradient boost, and 300 estimators for XGBoost. Comparing the seven models, k-nn and linear regression show a relatively short learning time, and AdaBoost, bagging, and XGBoost models show around one second of learning time. We see that XGBoost is the best model regarding accuracy and speed.

5. **Discussion.** This study intended to provide a basic framework for predicting and analyzing the air pollution status of the Korean peninsula according to the atmospheric conditions and air pollution conditions of neighboring countries. Among the predictive

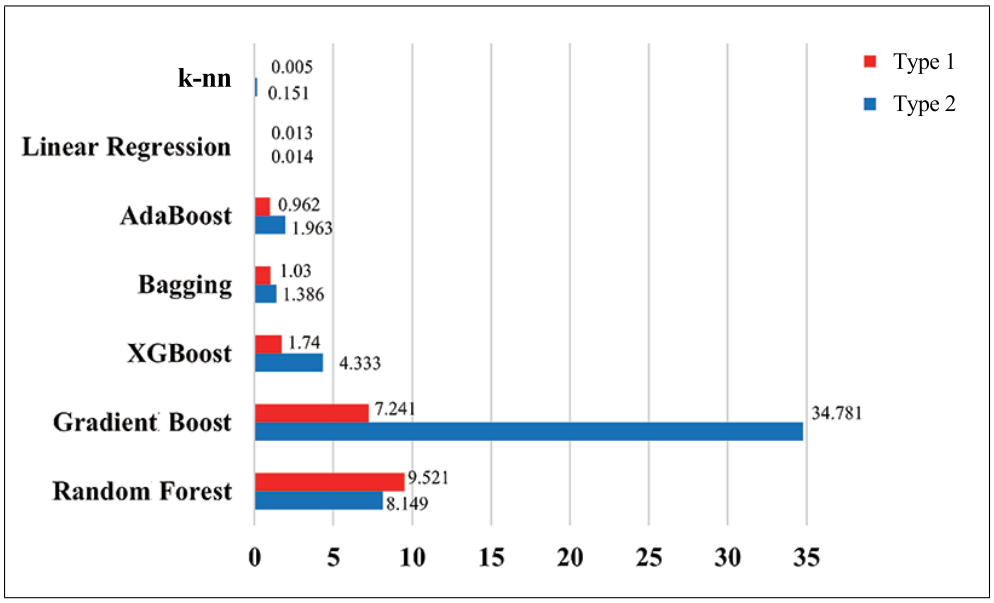

FIGURE 7. Learning time (sec.) of models

models, XGBoost had the highest prediction rate. The XGBoost model is a robust predictive model using the ensemble technique using a combination of several weak decision trees. Due to its regulation of overfitting and its excellent predictive performance in classification and regression, we concluded that the XGBoost model is the most suitable. Based on the prediction model presented in this study, implementing application software on a mobile device will be possible to provide real-time pollutant information to chronic patients suffering from air pollutants. By using the prediction system in this study, it will be possible to manage air pollutants seasonally, which is always intensive management of the high-concentration particulate matter. In addition, if high-concentration particulate matter persists for a certain period in a metropolitan area, such as Seoul, it will be possible to plan and implement measures against air pollution, such as emergency measures for pollutants, in advance through daily forecasts.

6. **Conclusions.** This study shows a high correlation between air pollutants in China and Korea and presented predictive models to forecast Korea's air pollutants using air pollutants from three administrative regions of China as independent variables. We build an air pollution prediction model in Korea based on the fact that air pollution in major regions of mainland China affects air pollution in Korea. The prediction model proposed is based on the meteorological characteristics of Korea and the characteristics of air pollutants in China. It is possible to effectively predict the number of air pollutants in the atmosphere, such as particulate matter and ultrafine particulate matter. Evaluating and comparing the results of the six regression models using performance measures of R2-Score, MSE, RMSE, and MAE, the predictive model in this study achieved a high R2-Score of approximately 0.92.

Acknowledgment. Soongsil University supports this work.

REFERENCES

[1] *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*, https://www.who.int/publications/i/item/9789240034228?ua=1, Accessed on June 8, 2022.

[2] K. S. Harishkumar, K. M. Yogesh and I. Gad, Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models, *Procedia Computer Science*, vol.171, pp.2057-2066, 2020.

[3] A. Shin, M. Shima and K. Yamamoto, Spatiotemporal land use random forest model for estimating metropolitan NO_2 exposure in Japan, *Science of the Total Environment*, vol.634, pp.1269-1277, 2018.
[4] T. M. Chiwewe and J. Ditsela, Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations, *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, pp.58-63, DOI: 10.1109/INDIN.2016.7819134, 2016.
[5] S. Park and H. Shin, Analysis of the factors influencing PM2.5 in Korea: Focusing on seasonal factors, *Journal of Environmental Policy and Administration*, vol.2, no.1, pp.227-248, DOI: 10.15301/jepa.2017.25.1.227, 2017.
[6] Q. Zhang, D. G. Streets, G. R. Carmichael, K. B. He, H. Huo, A. Kannari, Z. Klimont, I. S. Park, S. Reddy, J. S. Fu, D. Chen, L. Duan, Y. Lei, L. T. Wang and Z. L. Yao, Asian emissions in 2006 for the NASA INTEX-B mission, *Atmospheric Chemistry and Physics*, vol.9, no.14, pp.5131-5153, 2009.
[7] K. Kumar and B. P. Pande, Air pollution prediction with machine learning: A case study of Indian cities, *International Journal of Environmental Science and Technology*, DOI: 10.1007/s13762-022-04241-5, 2022.
[8] S. Hansun, A. Wicaksana and M. B. Kristanda, Prediction of Jakarta City air quality index: Modified double exponential smoothing approaches, *International Journal of Innovative Computing, Information and Control*, vol.17, no.4, pp.1363-1371, DOI: 10.24507/ijicic.17.04.1363, 2021.
[9] R. Yang, L. Yin, X. Hao, L. Liu, C. Wang, X. Li and Q. Liu, Identifying a suitable model for predicting hourly pollutant concentrations by using low-cost microstation data and machine learning, *Scientific Reports*, vol.12, 19949, DOI: 10.1038/s41598-022-24470-5, 2022.
[10] *Public Data Portal*, https://www.data.go.kr/data/15083697/fileData.do, Accessed on June 11, 2022.
[11] *World's Air Pollution: Real-Time Air Quality Index*, https://waqi.info/, Accessed on May 9, 2022.
[12] *Korea Meteorological Administration Meteorological Data Open Portal*, https://data.kma.go.kr, Accessed on May 4, 2022.
[13] *Supervised Learning*, https://scikit-learn.org/stable/supervised_learning.html#supervised-learning, Accessed on November 1, 2022.
[14] L. Breiman, Bagging predictors, *Machine Learning*, vol.24, no.2, pp.123-140, 1996.
[15] L. Hao and G. Huang, An improved AdaBoost algorithm for identification of lung cancer based on electronic nose, *Heliyon*, vol.9, no.3, 2023.
[16] Y. Shi, G. Ke, Z. Chen, S. Zheng and T.-Y. Liu, Quantized training of gradient boosting decision trees, *arXiv Preprint*, DOI: 10.48550/arXiv.2207.09682, 2022.
[17] L. Breiman, Random forests, *Machine Learning*, vol.45, pp.5-32, 2001.
[18] R. Mitchell and E. Frank, Accelerating the XGBoost algorithm using GPU computing, *Peer J. Computer Science*, vol.3, e127, DOI: 10.7717/peerj-cs.127, 2017.