❒ 2500

Comparative analysis of multiple classification models to improve PM10 prediction performance

Yong-Jin Jung, Kyoung-Woo Cho, Jong-Sung Lee, Chang-Heon Oh
Department of Electrical, Electronics and Communication Engineering, Korea University of Technology and Education
(KOREATECH), Korea

Article Info	ABSTRACT
	With the increasing requirement of high accuracy for particulate matter prediction, various attempts have been made to improve prediction accuracy by applying machine learning algorithms. However, the characteristics of particulate matter and the problem of the occurrence rate by concentration make it difficult to train prediction models, resulting in poor prediction. In order to solve this problem, in this paper, we proposed multiple classification models for predicting particulate matter concentrations required for prediction by dividing them into AQI-based classes. We designed multiple classification models using logistic regression, decision tree, SVM and ensemble among the various machine learning algorithms. The comparison results of the performance of the four classification models through error matrices confirmed the f-score of 0.82 or higher for all the models other than the logistic regression model. *This is an open access article under the <u>CC BY-SA</u> license.*

Corresponding Author:

Chang-Heon Oh
Department of Electrical, Electronics and Communication Engineering
Korea University of Technology and Education (KOREATECH)
1600, Chungjeol-ro, Byeongcheon-myeon, Dongnam-gu, Cheonan-si, Chungcheongnam-do, 31253
Republic of Korea
Email: choh@koreatech.ac.kr

1. INTRODUCTION

Particulate matter is a substance made up of various sizes, shapes, and ingredients. Particulate matter, divided into PM_{10}, $PM_{2.5}$ according to the size of $10_{\mu g}$, $2.5_{\mu g}$ or less, affects our health by causing some diseases such as cardiovascular, respiratory, and cerebrovascular diseases. Accordingly, particulate matter was classified as a dangerous substance, and it is analyzed as the cause of decreasing the vitality of society members [1-7]. In order to avoid such harmful effects of particulate matter as much as possible, it has become a routine practice to check the information provided based on the air quality index (AQI), which is divided into four categories: 'good', 'moderate', 'bad', and 'very bad'. Korea's particulate matter prediction accuracy was approximately 60% in 2015, and the Korea Meteorological Administration's prediction process has the annual predicton accuracy of approximately 80%. However, this is information reflecting on the weather forecaster's experience, and the actual particulate matter prediction model shows the accuracy of approximately 50% [8, 9].

Therefore, various attempts have been made to improve the prediction accuracy of particulate matter by applying machine learning prediction algorithms along with conventional statistical techniques [10, 11]. However, the characteristics of particulate matter arising from various external factors and the problem of the the occurrence rate by concentration make it difficult to effectively train prediction models. In order to improve the prediction accuracy of particulate matter concentrations, K. W. Cho *et al.,* in their study,

proposed a prediction model that separated and predicted them based on a specific concentration. By dividing the low and high concentrations based on the particulate matter concentration of $81_{\mu g}$, they compared the prediction performance through a deep neural network-based prediction model. The prediction results confirmed that the prediction performance of the low and high concentrations was improved, and especially it showed the performance improvement of 20.62% for the high concentrations [9]. The study by K. Kaya *et al.* proposed a solution to an unbalanced problem in order to address the prediction problem of the regression model due to the variation in the occurrence rate by particulate matter concentration. They confirmed the accuracy of approximately 80% in the entire data set by making the number of samples of the class the same for unbalanced data through the proposed upper sampling and down sampling [12].

In this paper, we propose data classification models by concentration to improve the performance of a particulate matter concentration prediction model. Of the machine learning classification models, we use the logistic regression, decision tree, support vector machine (SVM), and ensemble models. Based on the AQI, we configure multiple classification models by dividing particulate matter concentrations into 4 classes. In order to apply the optimal parameters to the models, we design the models by performing parameter search through grid search cross validation. We perform model evaluation using the error matrix.

2. DATA COLLECTION AND CONFIGURATION
2.1. Data collection and preprocessing
Particulate mattert is affected by various factors. Air pollutants and meteorological elements are typical, which are commonly applied to studies for predicting particulate matter concentrations [13-16]. Based on the studies, we selected the major data as shown in Table 1.

Table 1. Major data definition

Type	Name	Description
Air pollutants	PM_{10}	The average particulate matter($< 10\mu m$) per hour
	PM_{10h}	The average particulate matter($< 10\mu m$) of the previous 1 hour
	O_3	The average ozone of the previous 1 hour
	CO	The average carbon monoxide of the previous 1 hour
	NO_2	The average nitrogen dioxide of the previous 1 hour
	SO_2	The average sulfur dioxide of the previous 1 hour
Meteorological elements	Temperature	The average temperature of the previous 1 hour
	Humidity	The average humidity of the previous 1 hour
	Wind Speed	The average wind speed of the previous 1 hour
	Wind Direction	The most frequent wind direction of the previous 1 hour

According to the selected data, we collected the final confirmed data measured at an interval of an hour for 10 years from 2009 to 2018 at the measurement station around Cheonan in Korea. Air pollution data is composed of $PM_{10}, PM_{10h}, O_3, CO, NO_2$, and SO_2, and meteorological elements consist of temperature, humidity, wind speed, and wind direction. Since some data were missing due to the power outage and maintenance of measurement equipment, we removed all data of the same time when the missing data was present. Of the meteorological elements, the largest wind direction expressed in azimuth, that is, the 0° and 360°, which were often used with mixed notation, were unified to 360°.

There is a need for data preprocessing to perform classification through machine learning algorithms using the collected data. Since we used the classification algorithm based on supervised learning, we performed classification by separately dividing the data corresponding to independent variables and the data corresponding to dependent variables. The independent variable data, which includes PM_{10h}, O_3, CO, NO_2, SO_2, temperature, humidity, wind speed and wind direction, is used to predict the range of particulate matter concentrations based on the AQI. As for the wind direction, it is necessary to convert it to a vector form because it corresponds to categorical data expressed in 16 directions. Therefore, through one-hot encoding, we converted the categories corresponding to 16 directions to 16 vectors expressed in 0 and 1. For the remaining input sample data other than the wind direction, they are numerical data with different characteristics, and they were converted to a value between 0 and 1 through min max scaling in order to unify the range of numerical values expressed according to the data. The dependent variable data correspond to PM_{10}, and as shown in Table 2, based on the AQI used as a forecast by the Ministry of Environment, we divided PM_{10} into the sequential categories: 'good', 'moderate', 'bad', and 'very bad' , and expressed them as four classes of 0, 1, 2, and 3, respectively.

Table 2. The range of particulate matter concentrations based on AQI

Grade	Good	Moderate	Bad	Very Bad
$PM_{10}(\mu g/m^3)$	0 ~ 30	31 ~ 80	81 ~ 150	150 ~

2.2. Data configuration

The data used in the supervised learning model is mainly composed of a training set for learning and a test set for evaluating the trained model. The training set used to train the model is subdivided into a train set and a validation set because of the need to verify whether training is well completed. In this paper, we configure a training set of 75% and a test set of 25% with the preprocessed data. The training set is composed of a train set of 80% and a validation set of 20%. Figure 1 shows the structure of the final data used in the model, and Table 3 shows the configuration of the data set.
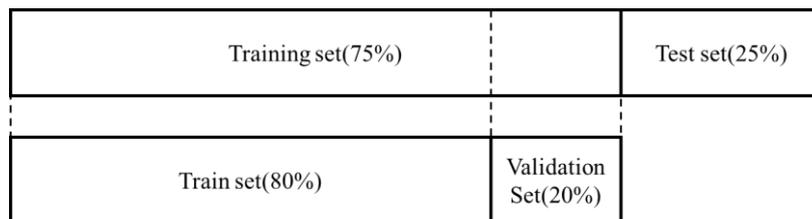

Figure 1. Structure of data set

Table 3. Data set configuration

Structure of data set		Samples
Training set	Train set	52,315
	Validation set	13,079
Test set		21,799
Total		87,193

3. CLASSIFICATION MODEL DESIGN
3.1. Logistic regression model design

Logistic regression is an algorithm used to predict the likelihood of an event using a linear combination of independent variables. As in general regression analysis, it is used in future prediction models by deriving a specific function through the relationship between dependent and independent variables. However, unlike linear regression, since the prediction result is classified as a specific category when the dependent variable is categorical data, it is used as a classification technique rather than a regression technique. It is divided into binomial or multinomial depending on the category characteristics of the dependent variable. The dependent variable for training the classification model of particulate matter concentrations has four categories. Accordingly, we built a model by applying a multinomial logistic regression method.

For a model predicting a certain result, overfitting or underfitting is contingent on the intensity of training. It is difficult for a model with overfitting to predict new data since it only focuses on training data. In the case of underfitting, there is a problem that the model does not predict most of the data since it does not identify the characteristics of the data due to simple training. To solve these problems, logistic regression basically uses L2 regularization parameter c [17].

Therefore, for better prediction performance, we performed the search for the optimal c value using grid search cross validation to find the value of c that fits the model. We set the range of c values to be searched to 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, and 1000. In order to select parameters with high generalization performance, we set the cv parameter of k-fold cross validation to 5. Accordingly, the c value was sequentially accessed to compare scores using the test set after 5 repetitive training runs and validations. For preprocessing of validation fold during cross validation, we searched the c values by building the pipeline of min max scaler and the model. Table 4 shows the mean test score and c values of the top 3 rankings in the cross validation results. The cross validation results showed that the mean test score was highest with 0.808958 when the c value was 10.0, thus we selected the c value to be applied to logistic regression as 10.0.

Table 4. Grid search cross validation results (logistic regression)

Rank	Mean test score	c
1	0.808958	10.0
2	0.808927	1000.0
3	0.806817	1.0

3.2. Decision tree model design

Decision tree is a widely used model for classification and regression. It is basically an algorithm that learns by continuously answering questions to approach a specific decision. With the increase in the number of leaf nodes, the accuracy of the training set increases but overfitting may occur [18]. One of the methods used to prevent overfitting is to stop the growth of the tree when the depth of the tree reaches a certain level. The parameter that limits the depth of a decision tree is max_depth, and we are able to improve the performance of the model by adjusting the depth.

Therefore, we performed the search for the optimal max_depth value using grid search cross validation. We set the range of max_depth values to be searched to 1~24, and performed the search by setting the cv parameter of k-fold cross validation to 5. Additionally, for preprocessing of validation fold during cross validation, we searched the max_depth values by building the pipeline of min max scaler and the model. Table 5 shows the mean test score and max_depth values of the top 3 rankings in the cross validation results. The cross validation results showed that the mean test score was highest with 0.85936013 when the max_depth value was 4, thus we selected the max_depth value to be applied to decision tree as 4.

Table 5. Grid search cross validation results (decision tree)

Rank	Mean test score	Max_depth
1	0.85936013	4
2	0.85896255	5
3	0.85856496	3

3.3. SVM model design

SVM is one of machine learning methods and is a supervised learning model for pattern recognition and data analysis. It is mainly used for classification and regression analysis. Given a set of data belonging to one of two categories, it generates a non-stochastic binary linear classification model that determines the category to which new data belongs based on a given data set. The generated classification model is expressed as a boundary in the space onto which the data is mapped. It is an algorithm to find the boundary with the largest width [19-21]. Therefore, SVM is a model that defines a baseline for classification between categories, which is expressed as a decision boundary.

In SVM, the difference in performance is determined depending on how the decision boundary is defined, and it is crucial to find the optimal decision boundary. The parameters applied to find the optimal decision boundary are c and gamma. C is a parameter that adjusts the allowable range of outliers by controlling the margin of the decision boundary, and gamma is a parameter that prevents overfitting of the model by controlling the flexibility of the decision boundary.

We performed the search for the optimal c and gamma values to find the optimal decision boundary using grid search cross validation. We set the range of c and gamma values to be searched to 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, and 1000, and performed the search by setting the cv parameter of k-fold cross validation to 5. Additionally, for preprocessing of validation fold during cross validation, we searched the c and gamma values by building the pipeline of min max scaler and the model. Table 6 shows the mean test score and relevant c and gamma values of the top 3 rankings in the cross validation results. The cross validation results showed that the mean test score was highest with 0.859238 when the c and gamma values were 1000 and 0.01, respectively, thus we selected the c value and the gamma value to be applied to SVM as 1000 and 0.01, respectively.

Table 6. Grid search cross validation results (SVM)

Rank	Mean test score	c	gamma
1	0.859238	1000	0.01
2	0.859177	1	1
3	0.859146	1	0.1

3.4. Ensemble model design

Ensemble is a technique that generates a powerful model by combining multiple models to achieve better prediction performance as compared with using an individual machine learning model. When multiple models are combined, the amount of calculation is generally increased, yet it prevents overfitting more effectively than using an individual model and it has the advantage of showing better performance than an individual model if the performance of an individual model is poor [22-24]. Ensemble is mainly divided into a collection methodology and a boosting methodology. The collection methodology has the predetermined set of models to be used, but the boosting methodology gradually increases the models to be used. In this study, we combined the logistic regression, decision tree, and SVM models previously designed, which corresponds to the collection methodology, to build an enemble model. Figure 2 shows the structure of the ensemble model.
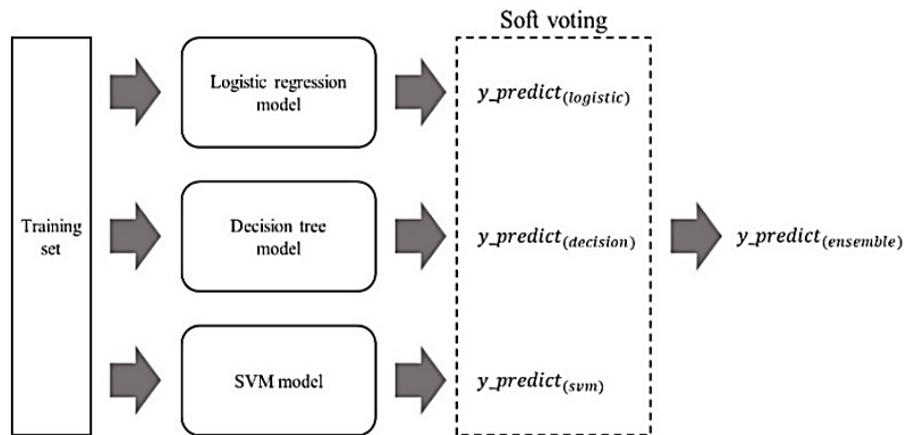

Figure 2. Structure of ensemble model

The training set data are used as an input variable to the combined logistic regression, decision tree, and SVM models, and the predicted results are outputted from an individual model. The final prediction results are generated by voting on the outputted results [25]. Voting is divided into hard and soft voting. Hard voting simply selects the final prediction based on the prediction results of an individual model. The voting method of the ensemble model designed in this study is soft voting, which selects the final prediction based on the sum of conditional probabilities of an individual model.

4. PERFORMANCE EVALUATION

We evaluated classification performance using the previously configured data set and designed the classification models. For performance evaluation, we used precision, recall, and f-score based on the error matrix. Figure 3 shows the error matrices created based on the classification results of the trained models. Table 7 shows the performance evaluation of the classification models calculated by referring to the error matrices.

When the logistic regression model predicted 'good', the precision was highest with 0.8685. When the prediction was performed based on the input data of 'moderate', the recall was highest with 0.9341. On the other hand, the classification did not work well for 'bad' and 'very bad'. Especially, the prediction was not made at all for 'very bad'. When the decision tree model predicted 'moderate', the precision was highest with 0.8977. When the prediction was performed based on the input data of 'moderate', the recall was highest with 0.9023. On the other hand, the precision and recall for 'bad' and 'very bad' showed relatively low values compared to 'good' and 'moderate'. The SVM model showed the highest precision and recall with 0.8997 and 0.8997, respectively, for 'moderate'. As in the decision tree model, the precision and recall for 'bad' and 'very bad' showed relatively low values compared to 'good' and 'moderate'. The ensemble model showed the highest precision and recall with 0.8997 and 0.8997 for 'moderate'. However, the precision and recall for 'bad' and 'very bad' showed relatively low values compared to 'good' and 'moderate', resulting in difficulty classifying the relevant classes. The analysis results based on the precision and recall showed that the precision and recall of 'good' and 'moderate' were relatively higher than those of 'bad' and 'very bad'. When analyzed through the error matrixes in Figure 3, it was confirmed that, of the input data, the proportion of

data corresponding to 'good' and 'moderate' used for classification was high. For an unbalanced model with the high proportion of a specific class, the generally used accuracy is meaningless. Therefore, we should evaluate the performance of the model by taking into account all classes with the same proportion, and used the mean of macro f-score. Other than the logistic regression model with an f-score of 0.4930, the other models showed the similar scores with 0.8264, 0.8277, and 0.8265.
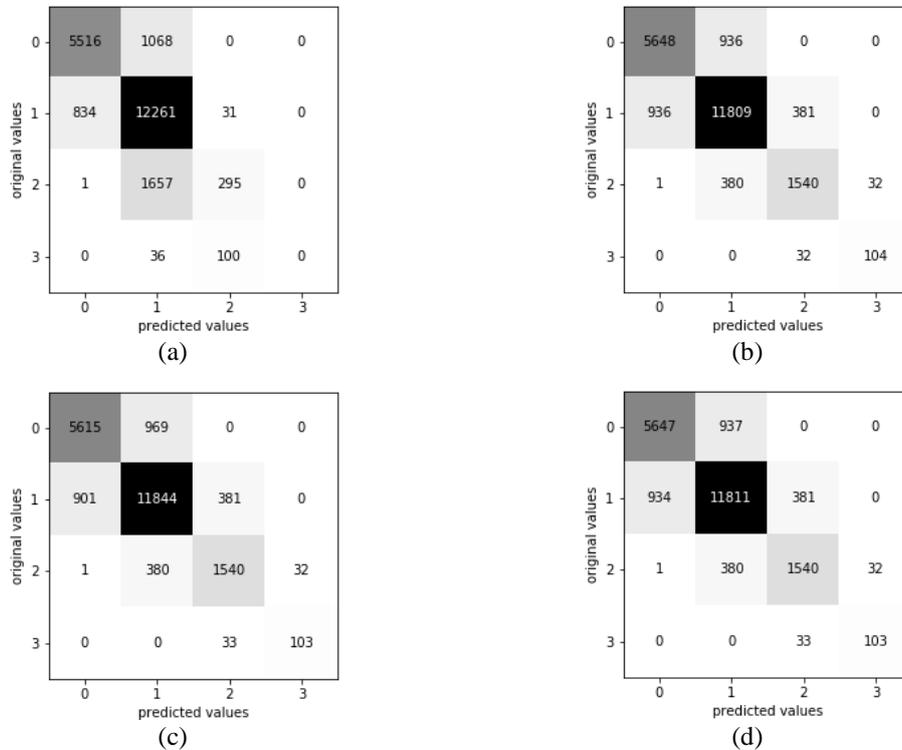

(a)

(b)

(c)

(d)

Figure 3. Confusion matrix; (a) Logistic regression model, (b) Decision tree model, (c) SVM model, (d) Ensemble model

Table 7. Classification performance evaluation by models

class	Logistic regression		Decision tree		SVM		Ensemble	
	precision	recall	precision	recall	precision	recall	precision	recall
0 (good)	0.8685	0.8378	0.8616	0.8528	0.8577	0.8578	0.8579	0.8577
1 (moderate)	0.8162	0.9341	0.8977	0.9023	0.8997	0.8997	0.8997	0.8998
2 (bad)	0.6925	0.1510	0.7881	0.7885	0.7885	0.7885	0.7881	0.7885
3 (very bad)	0.0000	0.0000	0.7630	0.7574	0.7647	0.7647	0.7630	0.7574
f-score (macro)	0.4930		0.8264		0.8277		0.8265	

5.   CONCLUSION

In predicting particulate matter concentrations, there is a problem of training particulate matter concentration prediction models because of the characteristics of particulate matter. In order to solve this problem, various studies have been underway such as performing prediction by dividing particulate matter concentrations based on a specific concentration. In this paper, to improve the performance of the particulate matter concentration prediction model, we proposed multiple classification models that provided particulate matter concentrations in four classes based on the AQI. To this end, we configured data sets by selecting air pollutant data and meteorological elements collected at an interval of an hour for 10 years around Cheonan. As the classification models in this study, we used the logistic regression, decision tree, SVM, ensemble. In order to apply optimal parameters to each model, we searched the parameters through grid search cross validation. We built the ensemble model by combining the logistic regression, decision tree, and SVM models into one. We used error matrixes to evaluate the performance of four multiple classification models. Logistic regression showed poor precision, recall, and f-score compared to other classification models.

Decision tree, SVM, and ensemble models all showed the precision and recall with 0.85 or higher for 'good' and 'moderate' based on the AQI, whereas they showed 0.75~0.79 for ''bad' and 'very bad'. We confirmed that this was because the particulate matter data used in the classification models were unbalanced data with the high proportion of a specific class. Accordingly, we verified the scores of the models by taking into account all classes with the same proportion, and found that the models other than the logistic regression model showed a score of 0.82 or higher. Of these models, the SVM model showed the best classification performance with 0.8277. In future, in order to address the problem of unbalanced data, we are going to compare classification performance through the algorithm changes of the classification models and design a particulate matter concentration prediction model based on the improved classification models.

REFERENCES
[1] G. W. Evans, "Air Pollution and Human Behavior," *Journal of Social,* vol. 37, no. 1, pp. 95-125, Apr. 2010.
[2] M. S. Seo, "The Impact of Particulate Matter on Economic Activity," *The Korean Women Economists Association,* vol. 12, pp. 75-100, Jun. 2015.
[3] A. Valavanidis, K. Fiotakis, and T. Vlachogianni., "Airborne Particulate Matter and Human Health: Toxicological Assessment and Importance of Size and Composition of Particles for Oxidative Damage and Carcinogenic Mechanisms," *Journal of Environmental Science and Health, Part C,* vol. 26, no. 4, pp. 339-362, 2008.
[4] J. O. Anderson, Josef G. Thundiyil and Andrew Stolbach, "Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health," *Journal of Medical Toxicology,* vol. 8, no. 2, pp. 166-175, 2012.
[5] K. H. Kim, E. Kabir and S. Kabir, "A Review on the Human Health Impact of Airborne Particulate Matter," *Environment International,* vol. 74, pp. 136-143, 2015.
[6] N. J. Hime, et al., "A Comparison of the Health Effects of Ambient Particulate Matter Air Pollution From Five Emission Sources," *International Journal of Environmental Research and Public Health,* vol. 15, no. 6, pp. 1-24, 2016, doi: 10.3390/ijerph15061206.
[7] World Health Organization (WHO), "Health effects of particulate matter. Policy implications for countries in eastern Europe, Caucasus and central Asia," *Regional Office for Europe,* 2013.
[8] Board of Adit and Inspection (BAI), "Weather forecast and earthquake notification system operation," *International THE Board of Audit and Inspection of KOREA,* 2017.
[9] K. W. Cho, et al., "Separation Prediction Model by Concentration based on Deep Neural Network for Improving PM10 Forecast Accuracy," *Journal of the Korea Institute of Information and Communication Engineering,* vol. 24, no. 1, pp. 8-14, 2020.
[10] J. W. Cha and J. Y. Kim, "Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model," *Journal of the Korea Institute of Information and Communication Engineering,* vol. 22, no. 4, pp. 595-601, 2018.
[11] A. Chaloulakou, G. Grivas, and N. Spyrellis, "Neural Network and Multiple Regression Models for PM10 Prediction in Athens: A Comparative Assessment," *Journal of the Air & Waste Management Association,* vol. 53, no. 10, pp. 1183-1190, 2003.
[12] K. Kaya and S. G. Oguducu, "A Binary Classification Model for PM10 Levels," *2018 3rd International Conference on Computer Science and Engineering (UBMK),* 2018, pp. 361-366.
[13] J. M. Han, Jae-Goo Kim, and Ki-Hyun Cho, "Verify a Causal Relationship between Fine Dust and Air Condition-Weather Data in Selected Area by Contamination Factors," *The journal of Bigdata,* vol. 2, no. 1, pp. 17-26, 2017.
[14] X. Zhao, et al., "A Deep Recurrent Neural Network for Air Quality Classification," *Journal of Information Hiding and Multimedia Signal Processing,* vol. 9, pp. 346-354, 2018.
[15] B. T. Ong, K. Sugiura, and K. Zettsu., "Dynamic pre-training of Deep Recurrent Neural Networks for predicting environmental monitoring data," *2014 IEEE International Conference on Big Data (Big Data),* 2014, pp. 760-765.
[16] X. Li, et al., "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation," *Environmental Pollution,* vol. 231, pp. 997-1004, 2017.
[17] S. H. Jeon and Y. S. Son, "Prediction of fine dust PM10 using a deep neural network model," *The Korean journal of applied statistics,* vol. 31, no. 2, pp. 265-285, 2018.
[18] R. S. Michalski, et al., "Learning Efficient Classification Procedures and Their Application to Chess End Games," *Machine Learning,* pp. 463-482, 1983.
[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.
[20] P. H. Huynh and Thanh-Nghi Do, "Enhancing gene expression classification of support vector machines with generative adversarial networks," *Journal of information and communication convergence engineering,* vol. 17, no. 1, pp. 14-20, 2019.
[21] G. Wang and S. Y. Shin, "An Improved Text Classification Method for Sentiment Classification," *Journal of information and communication convergence engineering,* vol. 17, no. 1, pp. 41-48, 2019.

[22] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review,* vol. 33, no. 1-2, pp. 1-39, 2010.
[23] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine,* vol. 6, no. 3, pp. 21-45, 2006.
[24] R. E. Sutanto and S. H. Lee, "Ensemble of Degraded Artificial Intelligence Modules Against Adversarial Attacks on Neural Networks," *Journal of information and communication convergence engineering,* vol. 16, no. 3, pp. 148-152, 2018.
[25] A. Husin and K. R. Ku-Mahamud, "Ant System and Weighted Voting Method for Multiple Classifier Systems," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 8, no. 6, pp. 4705-4712, 2018.

BIOGRAPHIES OF AUTHORS

Yong-Jin Jung, received his B.S. in Electronics Engineering from Kongju National University, Cheonan, South Korea, in 2014, and his M.S. in Electrical, Electronics and Communication Engineering from Korea University of Technology and Education (KOREATECH) in 2016. He is currently pursuing a Ph.D. in Electrical, Electronics and Communication Engineering at the Korea University of Technology and Education (KOREATECH). His research interests include machine learning, data analysis and deep learning.

Kyoung-Woo Cho, received his B.S. in Electronics Engineering from Kongju National University, Cheonan, South Korea, in 2013, and his M.S. in Electrical, Electronics and Communication Engineering from Korea University of Technology and Education (KOREATECH) in 2015. He is currently pursuing a Ph.D. in Electrical, Electronics and Communication Engineering at the Korea University of Technology and Education (KOREATECH). His research interests include machine learning, deep learning and particulate matters prediction.

Jong-Sung Lee, received his B.S in Information Communication from Korea Nazarene University, Cheonan, South Korea, in 2011, and his M.S in Electrical, Electronics and Communication Engineering from Korea University of Technology and Education (KOREATECH) in 2016. He is currently pursuing a Ph.D. in Electrical, Electronics and Communication Engineering at the Korea University of Technology and Education (KOREATECH). His research interests include 5G network, machine learning and deep learning.

Chang-Heon Oh, received the B.S. and M.S.E. degrees in telecommunication and information engineering from Korea Aerospace University, Kyunggi-Do, Korea, in 1988 and 1990, respectively. He received the Ph.D. degree in avionics engineering from Korea Aerospace University, Kyunggi-Do, Korea, in 1996. From February 1990 to August 1993, he was with Hanjin Electronics Co., where he was involved in the research and development of radio communication & monitoring systems. From October 1993 to February 1999, he was with the CDMA R&D center of Samsung Electronics Co., where he was involved in the design and development of CDMA cellular systems and CDMA PCS systems for successful commercial CDMA deployment in Korea. Since March 1999, he has been with the School of Electrical, Electronics and Communication Eng., Korea University of Technology and Education (KOREATECH), Cheonan, Korea, where he is currently a professor. His research interest is in the areas of wireless/mobile communication, wireless localization, IoT and engineering education.