CEO Clinical and Experimental Otorhinolaryngology Vol. 15, No. 2: 168-176, May 2022
https://doi.org/10.21053/ceo.2021.01536
pISSN 1976-8710 eISSN 2005-0720

Original Article

Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors

Yunseo Ku[1,*] · Soon Bin Kwon[2,*] · Jeong-Hwa Yoon[3] · Seog-Kyun Mun[4] · Munyoung Chang[4]

[1]Department of Biomedical Engineering, Chungnam National University College of Medicine, Daejeon, Korea; [2]Department of Neurology, Columbia University, New York, NY, USA; [3]Institute of Health Policy and Management, Medical Research Center, Seoul National University, Seoul; [4]Department of Otorhinolaryngology-Head and Neck Surgery, Chung-Ang University College of Medicine, Seoul, Korea

Objectives. Because climatic and air-pollution factors are known to influence the occurrence of respiratory diseases, we used these factors to develop machine learning models for predicting the occurrence of respiratory diseases.

Methods. We obtained the daily number of respiratory disease patients in Seoul. We used climatic and air-pollution factors to predict the daily number of patients treated for respiratory diseases per 10,000 inhabitants. We applied the relief-based feature selection algorithm to evaluate the importance of feature selection. We used the gradient boosting and Gaussian process regression (GPR) methods, respectively, to develop two different prediction models. We also employed the holdout cross-validation method, in which 75% of the data was used to train the model, and the remaining 25% was used to test the trained model. We determined the estimated number of respiratory disease patients by applying the developed prediction models to the test set. To evaluate the performance of each model, we calculated the coefficient of determination (R^2) and the root mean square error (RMSE) between the original and estimated numbers of respiratory disease patients. We used the Shapley Additive exPlanations (SHAP) approach to interpret the estimated output of each machine learning model.

Results. Features with negative weights in the relief-based algorithm were excluded. When applying gradient boosting to unseen test data, R^2 and RMSE were 0.68 and 13.8, respectively. For GPR, the R^2 and RMSE were 0.67 and 13.9, respectively. SHAP analysis showed that reductions in average temperature, daylight duration, average humidity, sulfur dioxide (SO_2), total solar insolation amount, and temperature difference increased the number of respiratory disease patients, whereas increases in atmospheric pressure, carbon monoxide (CO), and particulate matter ≤2.5 μm in aerodynamic diameter ($PM_{2.5}$) increased the number of respiratory disease patients.

Conclusion. We successfully developed models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. These models could evolve into public warning systems.

Keywords. *Machine Learning; Respiratory Diseases; Climate; Air Pollution; Gradient Boosting; Gaussian Process Regression*

• Received August 6, 2021
Revised September 17, 2021
Accepted September 19, 2021

• Corresponding author: **Munyoung Chang**
Department of Otorhinolaryngology-Head and Neck Surgery, Chung-Ang University Hospital, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea
Tel: +82-2-6299-1781, Fax: +82-2-825-1765, E-mail: cadu01@cau.ac.kr

• Co-corresponding author: **Seog-Kyun Mun**
Department of Otorhinolaryngology-Head and Neck Surgery, Chung-Ang University Hospital, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea
Tel: +82-2-6299-3129, Fax: +82-2-825-1765, E-mail: entdoctor@cau.ac.kr

*These authors contributed equally to this study.

INTRODUCTION

The unprecedented pandemic of coronavirus disease 2019 (COVID-19), a respiratory disease, continues to challenge the human race. Owing to the significant impact of COVID-19 on human health and social economies across the globe, research on respiratory diseases is becoming more important. Numerous studies have explored the influence of climatic and air-pollution factors on the occurrence of respiratory diseases. However, studies on methods of predicting the occurrence of respiratory diseases that integrate climatic and air-pollution factors are rare. Identifying the factors related to the occurrence of respiratory diseases and predicting the occurrence of respiratory diseases makes it possible to control respiratory disease risk factors or formulate respiratory disease preventive measures. For instance, reducing outdoor activities or wearing masks may be suggested to the public as approaches for reducing the occurrence of respiratory diseases during high-risk periods, thereby significantly contributing to the improvement of human health and global socio-economic conditions. The respiratory tract contains organs that are in direct contact with the atmosphere and are affected by climatic and air-pollution factors, which are known to be closely related to the occurrence of respiratory diseases because they affect the respiratory system or the survival and transmission of pathogens that cause respiratory infections [1,2]. Therefore, in this study, we hypothesized that climatic and air-pollution factors can be used to develop machine learning models for predicting the occurrence of respiratory diseases.

Machine learning models have shown the potential to predict the occurrence or the prognosis of clinical diseases, such as influenza-like illnesses or stroke [3,4]. For the occurrence of respiratory diseases, there exist a few previous studies that applied machine learning to produce forecasting models using air-pollution factors. Long short-term memory, which is a type of artificial recurrent neural network, was applied to analyze the lag effect of fine particles (particulate matter ≤2.5 μm in aerodynamic diameter [$PM_{2.5}$]) on the frequency of hospital emergency visits for respiratory diseases [5]. A multilayer perceptron using levels of particulate matter ($PM_{2.5}$ and PM_{10}) was also proposed for predicting outpatient visits for upper respiratory tract infections [6]. However, those models only considered particulate pollution without climatic and other air-pollution factors, such as nitrogen dioxide (NO_2), carbon monoxide (CO), or sulfur dioxide (SO_2). In this study, climatic factors, such as temperature or humidity, and more air-pollution data were applied as the input features of the models. Gradient boosting and Gaussian process regression (GPR) models, which have been successfully applied to forecasting in time series analyses using multivariate data, were adopted to predict the occurrence of respiratory diseases [7,8]. Moreover, the positive or negative contribution of each input to the model's predicted outcome was analyzed using the Shapley Additive exPlanations (SHAP) approach, which is commonly used as a method for interpreting machine learning models [9].

The National Health Insurance Service (NHIS) of South Korea is available to all citizens. When citizens enrolled in the NHIS receive medical treatment, diagnosis-related information is stored in a central database. Therefore, this information can be used to estimate the occurrence of diseases. The NHIS provides data on the daily number of patients treated for respiratory diseases, which are common in South Korea. We used this information to determine the climatic and air-pollution factors that influence the occurrence of respiratory diseases, after which we developed models for predicting the occurrence of respiratory diseases. To analyze an area affected by similar climatic and air-pollution factors, the study area was limited to Seoul, the capital city of South Korea, with a population of 10 million. We obtained the daily number of patients treated for respiratory diseases in Seoul and the levels of climatic and air-pollution factors from 2014 to 2019. Considering that very few countries have such insurance systems, our research is significantly valuable.

HIGHLIGHTS

- We developed machine-learning models to predict the occurrence of respiratory diseases.
- Climatic and air-pollution factors were used as the input features of the models.
- The models were developed using gradient boosting and Gaussian process regression (GPR) methods.
- For gradient boosting, the R^2 and root mean square error values were 0.68 and 13.8, respectively.
- For GPR, the R^2 and root mean square error values were 0.67 and 13.9, respectively.

MATERIALS AND METHODS

Study population and preprocessing

This study was approved by the Institutional Review Board of Chung-Ang University Hospital (IRB No. 2140-001-457). The written informed consent requirement was waived for this study because anonymized data were used. All research was performed in accordance with the tenets of the Declaration of Helsinki. The study period was from January 1, 2014, to December 31, 2019. Daily numbers of patients treated for respiratory diseases per 10,000 inhabitants in Seoul during the study period were collected from the NHIS data, as shown in Fig. 1. Seoul's daily climatic and air-pollution factors during the study period were collected.

Because the daily number of patients treated for respiratory diseases was analyzed in this study, it was expected that the number of patients would decrease on holidays, including Sundays and public holidays, when most medical institutions are not in

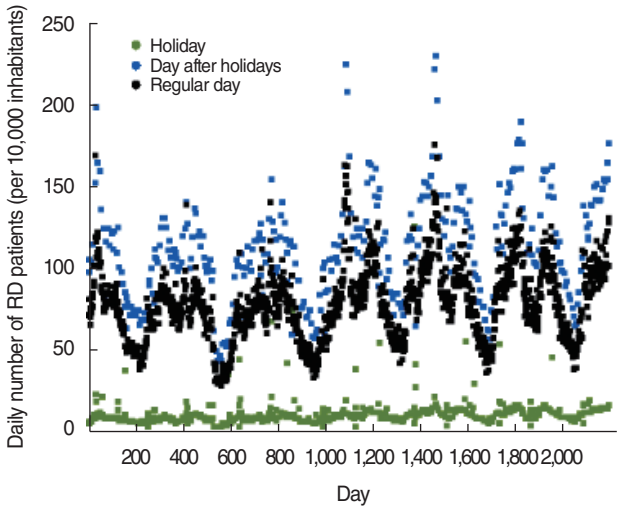

Fig. 1. Daily numbers of patients treated for respiratory disease (RD) per 10,000 inhabitants in Seoul from January 1, 2014, to December 31, 2019. The green dots indicate holidays, the blue dots, the days after holidays, and black dots, regular weekdays.

operation. On the other hand, it was expected that the number of patients would increase on the days immediately after holidays. If this hypothesis is correct, the numbers of patients on holidays and on the day immediately after holidays should be excluded to assess the actual impact of climatic and air-pollution factors on respiratory disease occurrence. To confirm this hypothesis, the number of patients treated for respiratory diseases per 10,000 inhabitants was divided into three groups: holidays, days after holidays, and regular days. The three groups were then compared to each other using one-way analysis of variance (ANOVA) followed by the Bonferroni *post-hoc* test. After this step, the 7-day moving averaging of input features was applied to reflect the cumulative effect of climatic and air-pollution factors on respiratory diseases [10]. Statistical analyses were performed using the IBM SPSS ver. 21.0 (IBM Corp., Armonk, NY, USA).

Respiratory diseases

In South Korea, after a patient is treated in a hospital, Korean Classification of Diseases (KCD) codes are assigned to each patient according to the diagnosis. KCD codes are based on the tenth edition of the International Classification of Diseases. By analyzing the KCD codes, the number of patients treated for a specific disease in South Korea can be identified. NHIS provides data on the daily number of patients treated for the following respiratory diseases, with the primary diagnosis under the following KCD codes, which are diagnosis codes for common respiratory diseases in South Korea: acute nasopharyngitis (J00), acute sinusitis (J01), acute pharyngitis (J02), acute tonsillitis (J03), acute laryngitis or tracheitis (J04), acute upper respiratory infections (J06), bronchopneumonia (J18), acute bronchitis (J20), acute bronchiolitis (J21), acute lower respiratory infections (J22), vasomotor or allergic rhinitis (J30), chronic rhinitis, nasopharyngitis, or pharyngitis (J31), chronic sinusitis (J32), disorders of nose and nasal sinuses (J34), peritonsillar abscess (J36), and bronchitis (J40). These data are posted on the Public Data Portal website operated by the Korea Information Society Agency (https://www.data.go.kr/). The total populations in Seoul by year were obtained from the Seoul Metropolitan Government website (http://data.seoul.go.kr/dataList/419/S/2/datasetView.do). The daily numbers of patients treated for the respiratory diseases mentioned above per 10,000 inhabitants were calculated from January 1, 2014, to December 31, 2019.

Climatic and air-pollution factors

Climatic factors, including average temperature (°C), temperature difference (°C), average humidity (%), daylight duration (hr), sunshine duration (hr), total solar insolation amount (MJ/m^2), atmospheric pressure (hPa), precipitation (mm), and cloud amount (cloud covered area, expressed as 10 fractions of the entire sky), in Seoul were collected daily from the Korea Meteorological Administration Weather Data Service website, which is operated by the Korea Meteorological Administration (http://www.kma.go.kr/eng/index.jsp). Seoul's daily air-pollution indicators, including the levels of $PM_{2.5}$, particulate matter ≤10 μm in aerodynamic diameter (PM_{10}), ozone (O_3), NO_2, CO, and SO_2, were also extracted throughout the study period from the Seoul atmospheric environment information website operated by the Seoul Metropolitan Government (https://cleanair.seoul.go.kr/2020/statistics/dayAverage). As the levels of O_3, NO_2, CO, and SO_2 were lower than 1 ppm, a unit of ppb was used to facilitate the determination of the effect of air-pollution factors on the number of respiratory disease patients.

Machine learning model development

Fig. 2 shows the overall procedure for the development of the prediction models in this study. First, the relief-based algorithm for regression was applied to evaluate the importance of feature selection. The relief algorithm is based on an instance-based learning approach, and it provides statistical relevance of input features to the target response [11]. Features with negative weights were excluded in this step.

Next, two different prediction models were developed using gradient boosting and GPR methods, respectively. Gradient boosting is a supervised machine learning technique based on decision trees, and it combines weak prediction models into a single strong learner in an iterative framework. At every step, the algorithm fits the difference between the observed response and the aggregated prediction of all previous learners by minimizing the mean-squared error. In this study, the minimum leaf size and the number of learners were optimized during model training. This boosting method provides not only the accuracy comparable to other classical methods, such as support vector machine, but also the interpretability of a machine learning model [12].

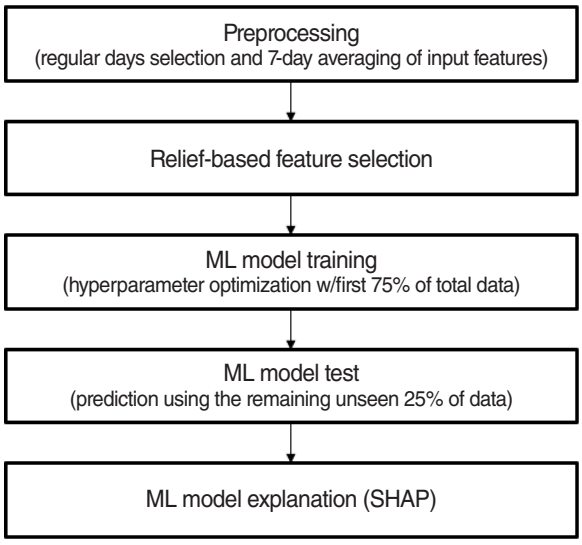

Fig. 2. The overall procedure for the development of the machine learning (ML) prediction models. The training and test sessions with hyperparameter optimization were performed after data preprocessing and feature selection, after which Shapley Additive exPlanations (SHAP)-based interpretation for the developed models was performed.

GPR is a non-parametric Bayesian algorithm that is popularly applied to prediction in time series data [13]. The probabilistic GPR model is a predictive variance under Gaussian assumptions, which is that data points with similar input values tend to be close in the output space, i.e., similarity [14]. Thus, the GPR model is defined using the covariance function, i.e., kernel, and its hyperparameters, which specify the effect of input changes on the output [15]. In other words, the kernel determines how the response at one point is affected by responses at other points. In this study, the exponential kernel containing the signal standard deviation (σ_f) and the characteristic length scale (σ_l) was optimized during model training. The characteristic length scale can be different for each predictor during the optimization process. For two different predictor values x_i and x_j ($i \neq j$, i=1, 2, ..., n), the basic exponential kernel with parametrization vector θ is defined as follows:

$$k(x_i, x_j|\theta) = \sigma_f^2 \exp\left(-\frac{r}{\sigma_l}\right), \quad (1)$$

where r represents the Euclidean distance between x_i and x_j.

The sparse GPR using a regressor approximation subset was also evaluated to investigate change in prediction performance. The temporal validation splitting data into temporal folds was performed, which has generally been applied to time series forecasting [14,16,17]. In the present study, the first 75% of the data from 2014 to the first half of 2018 was used as the training data for model development. In this model training session, holdout validation was performed using 30% of the training data. Then the trained model was prospectively evaluated with the more recent remaining 25% of data, which is the unseen data between the latter part of 2018 and 2019.

Bayesian optimization was applied to select the hyperparameters of the machine learning models. The estimated numbers of respiratory disease patients were calculated by applying the developed prediction models to the test set. Machine learning model development was performed using MATLAB R2020a (Mathworks, Natick, MA, USA). The coefficient of determination (R^2) and root mean square error (RMSE) between the original and estimated numbers of respiratory disease patients were calculated for each model to evaluate the performance. The R^2 and RMSE are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (3)$$

where y_i represents the predicted number of respiratory disease patients, $\hat{y}_i$ represents the actual number of respiratory disease patients, and $\bar{y}$ represents the mean value of the actual number of respiratory disease patients.

Finally, SHAP was used to interpret the estimated output of each machine learning model [9]. SHAP is an approach for explaining the contribution of a specific input to the prediction. This method is based on the Shapley values from coalitional game theory, which is the average marginal contribution across all possible coalitions [18]. In this study, a SHAP Python package using the TreeSHAP algorithm for decision tree-based models, such as random forests and gradient boosted trees, was adopted along with XGBoost, which is an efficient implementation of a gradient boosting model for Python [19].

RESULTS

The daily numbers of patients treated for respiratory diseases as well as the climatic and air-pollution factors are listed in Supplementary Table 1. In the preprocessing step, one-way ANOVA confirmed significant differences between the three different day groups (F(2, 2243)=1,287.4, P<0.001). Bonferroni *post hoc* testing revealed that the number of respiratory disease patients on holidays (12.8±10.1) was significantly lower than that on regular days (78.3±22.9, P<0.001). Conversely, the number of respiratory disease patients on days after holidays (112.5±32.0) was significantly higher than that on regular days (P<0.001). Thus, data from holidays and the days after holidays were excluded from further analysis.

Fig. 3 shows the results of the relief-based feature selection algorithm after applying the 7-day moving average to the input features. Three climatic factors (cloud amount, precipitation, and sunshine duration) and one air-pollution factor (O_3) that showed

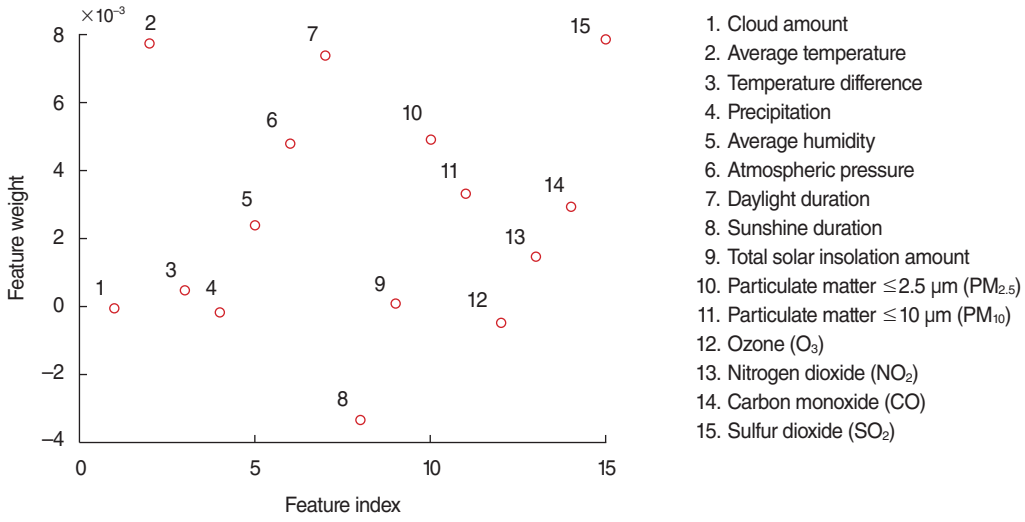

Fig. 3. The results of the relief-based feature selection algorithm for 15 climatic and air-pollution factors. Higher feature weights indicate higher importance for the target response.

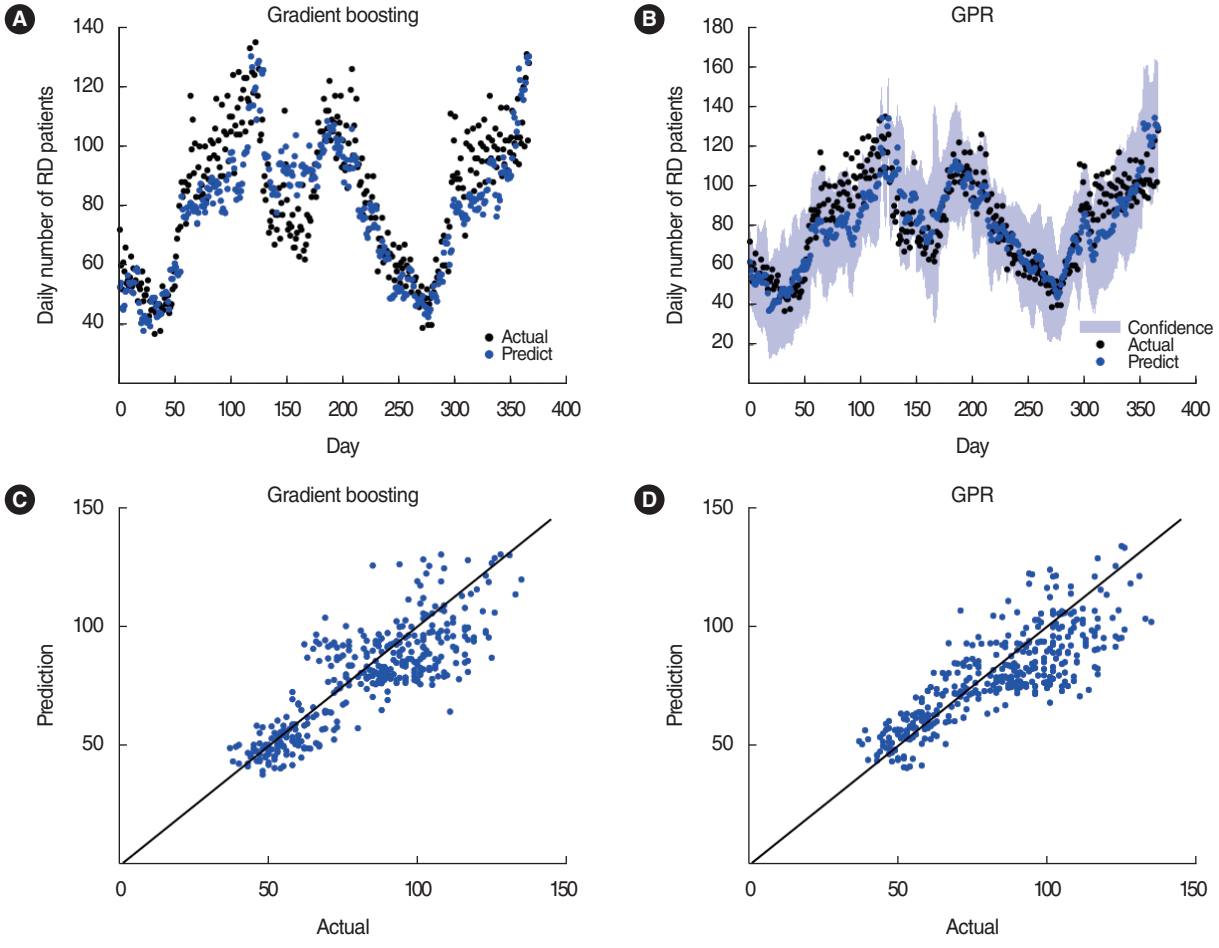

Fig. 4. The prediction results of the daily number of respiratory disease (RD) patients using unseen test data (the latter part of 2018 and 2019). (A) The prediction results using the developed gradient boosting model. (B) The prediction results using the developed Gaussian process regression (GPR) model. The black and blue dots indicate the actual and predicted daily numbers of RD patients per 10,000 inhabitants, respectively. The shaded area represents the 95% confidence interval. (C, D) Scatter plots between the actual and predicted RD patients for the developed gradient boosting and GPR models, respectively. The solid black line represents the Y=X line.

negative weights were excluded from the input features during the model development process.

Fig. 4 shows the prediction results of the daily number of respiratory disease patients using unseen test data. The gradient boosting and GPR methods demonstrated similar performance after hyperparameter optimization. When applying gradient boosting with a minimum leaf size of 30 and the number of learning cycles at 100, the R^2 and RMSE values were 0.68 and 13.8, respectively. For GPR using an exponential kernel with a signal standard deviation of 13 and a customized length scale for each predictor, the R^2 and RMSE values were 0.67 and 13.9, respectively. For sparse GPR using a regressor approximation subset, the R^2 and RMSE values were 0.64 and 14.3, respectively.

Fig. 5 shows the results of SHAP analysis. The value for each input feature represents the summation of absolute Shapley values across the data. Larger values indicate higher global importance in terms of feature contribution. As shown in Fig. 5A, the top four features with stronger influences on the occurrence of respiratory diseases among patients were all climatic factors: average temperature, daylight duration, average humidity, and atmospheric pressure. They were followed by SO_2, total solar insolation amount, CO, $PM_{2.5}$, and others. Fig. 5B shows the SHAP

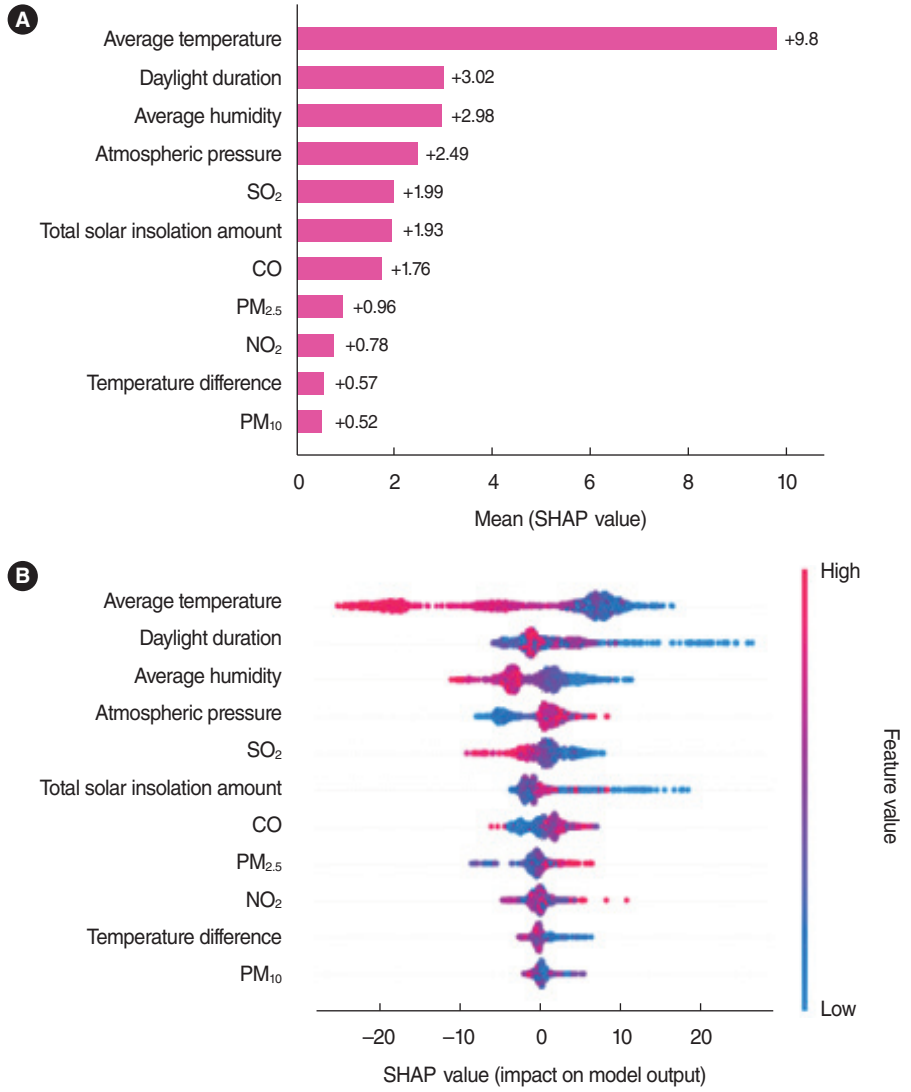

Fig. 5. Shapley Additive exPlanations (SHAP) feature importance (A) and summary plot (B). The SHAP feature importance (i.e., the mean absolute Shapley values) for the gradient boosting model. In the SHAP summary plot, the features on the Y-axis are ordered based on their importance. The color bars indicate the amplitudes of feature values from low to high. Overlapping points are stacked in the Y-axis directions of both images to show the distribution of the Shapley values for each feature. Reductions in average temperature, daylight duration, average humidity, sulfur dioxide (SO_2), total solar insolation amount, and temperature difference increased the number of respiratory disease patients, whereas increases in atmospheric pressure, carbon monoxide (CO), and particulate matter ≤2.5 μm in aerodynamic diameter ($PM_{2.5}$) increased the number of respiratory disease patients. NO_2, nitrogen dioxide; PM_{10}, particulate matter ≤10 μm in aerodynamic diameter.

summary plot, which explains the effect of each feature on the prediction results. Each point represents a Shapley value for a feature and an instance, and its color indicates the amplitude from low to high. According to the relationships demonstrated by the SHAP summary plot, reductions in average temperature, daylight duration, average humidity, SO_2, total solar insolation amount, and temperature difference increased the number of respiratory disease patients, whereas increases in atmospheric pressure, CO, and $PM_{2.5}$ increased the number of respiratory disease patients. No distinct trend was established for PM_{10} and NO_2.

DISCUSSION

In this study, we proposed gradient boosting- and GPR-based machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. Both models demonstrated competitive prediction performance, with R^2 values of over 0.67 and RMSE values below 13.9 per 10,000 inhabitants. It is noteworthy that the prediction performance was evaluated using totally unseen data (the latter part of 2018 and 2019) that were not used in the training stage.

Regarding the effect of each factor on the occurrence of respiratory diseases, reductions in average temperature, average daylight duration, total solar insolation amount, humidity, SO_2, and temperature difference increased the number of respiratory disease patients. In contrast, increases in atmospheric pressure, CO, and $PM_{2.5}$ increased the number of respiratory disease patients. Most of the respiratory diseases included in this study were infectious diseases, although some allergic diseases were included as well. In the case of infectious diseases, climatic and air-pollution factors influence the occurrence of respiratory infections in two major ways. First, they can influence the incidence of respiratory infections by affecting the survival, proliferation, and dissemination of pathogens that cause respiratory infections. Average temperature and humidity can influence the incidence of respiratory infections in this way. For instance, respiratory syncytial virus, which is a virus that induces respiratory infections, is significantly active at low temperatures because low temperatures make the lipid envelope of the virus highly resistant to degradation, thereby making the virus more stable in secretions through which it is transmitted [1,20]. In addition, as temperature decreases, the indoor life increases, thereby making virus transmission easier [20]. Therefore, reductions in average temperature would increase the incidence of respiratory infections. Humidity has also been reported to affect the transmission of pathogens. Ward et al. [21] reported that the lower the humidity, the smaller the pathogenic droplets discharged during coughing or sneezing, and the smaller the droplet size, the longer a pathogen remains in the air, making it easier for the pathogen to spread to the surroundings. These patterns are consistent with our findings.

Second, climatic and air-pollution factors can influence the development of respiratory infections by affecting the respiratory defense system [22]. Climatic factors, such as humidity, average daylight duration, and total solar insolation amount, can influence the occurrence of respiratory infections in a similar manner. Dry air reduces the ability of respiratory epithelial cells to repel viral particles, thereby suppressing the body's defense against pathogens [23]. Therefore, the lower the humidity, the more susceptible the body is to respiratory infections. This pattern is consistent with our findings. There exist few studies on the association between sunlight and respiratory infections. Ferrari et al. [24] reported that sunlight prevents the deterioration of chronic obstructive pulmonary disease. Schwarz and Schwarz [25] suggested that sunlight can protect the respiratory tract by reducing inflammatory responses in the respiratory tract. Through this mechanism, sunlight may have a protective effect on the occurrence of respiratory infections, which is consistent with our results. Air pollutants are also known to weaken the body's defense against pathogens, which has been reported in South Korea [26,27]. Exposure to air pollutants results in free radicals in the respiratory tract. Free radicals damage the respiratory tract and weaken the body's defense against pathogens [22]. The effect of $PM_{2.5}$ on the occurrence of respiratory infections has been reported in many studies [28,29]. Croft et al. [28] studied 500,000 adults diagnosed with influenza, bacterial pneumonia, or culture-negative pneumonia in New York and reported that the incidence of culture-negative pneumonia and influenza was associated with an increase in $PM_{2.5}$ concentrations over the previous weeks. These results are consistent with our findings. The effect of CO on the occurrence of respiratory infections has also been reported in several studies [30]. In these studies, the researchers reported that increased concentrations of CO resulted in an increase in the occurrence of respiratory infections. These results are also consistent with our findings.

Climatic and air-pollution factors are known to influence respiratory allergic diseases, including allergic rhinitis. In particular, it has been reported that cold weather worsens respiratory symptoms in allergic rhinitis [31-33]. Among air pollutants, $PM_{2.5}$ has been reported to increase the prevalence of allergic rhinitis [34,35]. These results are also consistent with our findings.

In this study, we established that the occurrence of respiratory diseases decreased as the levels of SO_2 increased. This is a different trend from that reported in previous studies, in which an increase in the concentration of SO_2 levels resulted in an increased occurrence or exacerbation of respiratory diseases [30]. However, trends similar to our findings were observed in some previous studies. Nhung et al. [36] reported that an increase in SO_2 concentration levels was associated with a shorter hospital stay among children with acute lower respiratory tract infections in single-pollutant models. Zhu et al. [37] investigated the relationship between SO_2 concentration levels and daily confirmed cases of COVID-19 and reported that SO_2 levels were negatively correlated with the number of confirmed cases per day. This is pre-

sumably a result of the interaction between SO_2 and other air pollutants as well as the differences in SO_2 concentration levels depending on the study areas. For the SO_2 concentration levels in Seoul, which are covered in our study, the maximum concentration level was 14 ppb. This is much lower than 50 ppb, which is the limit of SO_2 concentration allowed by the South Korean government. Therefore, it is difficult to evaluate the effect of SO_2 concentration levels on the occurrence of respiratory diseases using our data.

The effect of atmospheric pressure on the occurrence of respiratory diseases has yet to be studied extensively. However, a few studies have reported a positive correlation between hospitalizations resulting from lower respiratory tract infections and atmospheric pressure [38,39]. This pattern is consistent with our findings. However, the mechanism behind this phenomenon remains to be elucidated, and further research is required. In this study, we established that the smaller the temperature difference, the higher the incidence of respiratory diseases. However, looking at the results of the SHAP analysis, it is difficult to consider this a meaningful result because the effect of the temperature difference was insignificant.

There are several limitations of our study. First, as this study used the daily number of respiratory disease patients provided by NHIS, personal information such as age, sex, body mass index, and past medical history were not considered. Second, this study only used data from the urban area of Seoul for analysis. Different regions have different climatic factors and degrees of air pollution. Therefore, we determined that there are differences in the main factors affecting the occurrence of respiratory diseases. Because the respiratory disease prediction model developed in this study used data from Seoul, which is a large city, we believe that there is a limit to its application in rural areas or areas with different climatic characteristics. Further studies on more regions, including rural areas or areas with different climatic characteristics, are required. Regarding the non-parametric GPR employed in this study, the expensive computation due to matrix inversions is a major limitation. The current number of data points (less than 2,000) in this study does not cause serious computational constraints. However, as data accumulates and a large dataset is established, computational bottlenecks can occur. To address this problem, compressive methods using only a subset of the regression model or dimensional reduction of the data could be applied to larger datasets [40]. Moreover, the prediction performance could be improved by addressing the unrealistic assumptions of GPR. The combinatorial method of covariance functions using both spatial and temporal data inputs has been proposed as a possible solution when real-world data is used [14]. Similarly, non-stationary and periodic behaviors of the time series of respiratory diseases and the dependence on climate and air-pollution factors can be combined to improve the prediction performance.

We successfully developed machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. In the future, these models could evolve into systems for warning the public by predicting the occurrence of respiratory diseases. In addition, these models could also be used to prevent the occurrence of respiratory diseases by aiding in the control of risk factors.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

This study used NHIS data made by National Health Insurance Service (NHIS), which are posted on the Public Data Portal website operated by the Korea Information Society Agency: https://www.data.go.kr/. The author(s) declare no conflict of interest with NHIS.

This study was supported by the research fund of Chungnam National University, Daejeon, Korea.

ORCID

Yunseo Ku https://orcid.org/0000-0003-2737-4427
Soon Bin Kwon https://orcid.org/0000-0001-5076-0743
Jeong-Hwa Yoon https://orcid.org/0000-0002-9150-3732
Seog-Kyun Mun https://orcid.org/0000-0001-8624-2964
Munyoung Chang https://orcid.org/0000-0003-0136-3893

AUTHOR CONTRIBUTIONS

Conceptualization: YK, SKM, MC. Data curation: all authors. Formal analysis: YK, SBK, MC. Funding acquisition: YK, MC. Methodology: all authors. Project administration: YK, SKM, MC. Visualization: YK, SBK. Writing–original draft: YK. SBK, MC. Writing–review & editing: all authors.

SUPPLEMENTARY MATERIALS

Supplementary materials can be found via https://doi.org/10.21053/ceo.2021.01536.

REFERENCES

1. Tang JW, Loh TP. Correlations between climate factors and incidence:

a contributor to RSV seasonality. Rev Med Virol. 2014 Jan;24(1): 15-34.
2. Vandini S, Corvaglia L, Alessandroni R, Aquilano G, Marsico C, Spinelli M, et al. Respiratory syncytial virus infection in infants and correlation with meteorological factors and air pollutants. Ital J Pediatr. 2013 Jan;39(1):1.
3. Sohn J, Jung IY, Ku Y, Kim Y. Machine-learning-based rehabilitation prognosis prediction in patients with ischemic stroke using brainstem auditory evoked potential. Diagnostics (Basel). 2021 Apr;11(4):673.
4. Volkova S, Ayton E, Porterfield K, Corley CD. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. PLoS One. 2017 Dec;12(12):e0188941.
5. Lu J, Bu P, Xia X, Yao L, Zhang Z, Tan Y. A new deep learning algorithm for detecting the lag effect of fine particles on hospital emergency visits for respiratory diseases. IEEE Access. 2020;8:145593-600.
6. Yang PH, Hsieh MT, Lin GM, Chen MJ, Yeh CH, Huang ZX, et al. Prediction of outpatient visits for upper respiratory tract infections by machine learning of PM2.5 and PM10 levels in Taiwan. In Proceedings of the 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW); 2018.
7. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. J Med Internet Res. 2021 Feb;23(2):e24246.
8. Dürichen R, Pimentel MA, Clifton L, Schweikard A, Clifton DA. Multitask Gaussian processes for multivariate physiological time-series analysis. IEEE Trans Biomed Eng. 2015 Jan;62(1):314-22.
9. Lundberg S, Lee SI. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017.
10. Kim H, Kim Y, Hong YC. The lag-effect pattern in the relationship of particulate air pollution to daily mortality in Seoul, Korea. Int J Biometeorol. 2003 Sep;48(1):25-30.
11. Robnik-Sikonja M, Kononenko I. An adaptation of Relief for attribute estimation in regression. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97); 1997.
12. Liu L, Yu Y, Fei Z, Li M, Wu FX, Li HD, et al. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. BMC Syst Biol. 2018 Nov;12(Suppl 6):105.
13. Liu H, Ong YS, Shen X, Cai J. When Gaussian process meets big data: a review of scalable GPs. IEEE Trans Neural Netw Learn Syst. 2020 Nov;31(11):4405-23.
14. Chen S, Xu J, Wu Y, Wang X, Fang S, Cheng J, et al. Predicting temporal propagation of seasonal influenza using improved gaussian process model. J Biomed Inform. 2019 May;93:103144.
15. Caywood MS, Roberts DM, Colombe JB, Greenwald HS, Weiland MZ. Gaussian process regression for predictive but interpretable machine learning models: an example of predicting mental workload across tasks. Front Hum Neurosci. 2017 Jan;10:647.
16. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput Biol. 2015 Oct; 11(10):e1004513.
17. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ Digit Med. 2021 May; 4(1):87.
18. Shapley LS, Roth AE. The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge: Cambridge University Press; 1988.
19. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020 Jan;2(1):56-67.
20. Vandini S, Bottau P, Faldella G, Lanari M. Immunological, viral, environmental, and individual factors modulating lung immune response to respiratory syncytial virus. Biomed Res Int. 2015;2015:875723.
21. Ward MP, Xiao S, Zhang Z. Humidity is a consistent climatic factor contributing to SARS-CoV-2 transmission. Transbound Emerg Dis. 2020 Nov;67(6):3069-74.
22. Ciencewicki J, Jaspers I. Air pollution and respiratory viral infection. Inhal Toxicol. 2007 Nov;19(14):1135-46.
23. Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of respiratory viral infections. Annu Rev Virol. 2020 Sep;7(1):83-101.
24. Ferrari U, Exner T, Wanka ER, Bergemann C, Meyer-Arnek J, Hildenbrand B, et al. Influence of air pressure, humidity, solar radiation, temperature, and wind speed on ambulatory visits due to chronic obstructive pulmonary disease in Bavaria, Germany. Int J Biometeorol. 2012 Jan;56(1):137-43.
25. Schwarz T, Schwarz A. Molecular mechanisms of ultraviolet radiation-induced immunosuppression. Eur J Cell Biol. 2011 Jun-Jul;90(6-7): 560-4.
26. Kim SY, Kong IG, Min C, Choi HG. Association of air pollution with increased risk of peritonsillar abscess formation. JAMA Otolaryngol Head Neck Surg. 2019 Jun;145(6):530-5.
27. Kim SY, Min C, Yoo DM, Park B, Choi HG. Short-term exposure to air pollution and epiglottitis: a nested case-control study. Laryngoscope. 2021 Nov;131(11):2483-9.
28. Croft DP, Zhang W, Lin S, Thurston SW, Hopke PK, Masiol M, et al. The association between respiratory infection and air pollution in the setting of air quality policy and economic change. Ann Am Thorac Soc. 2019 Mar;16(3):321-30.
29. Horne BD, Joy EA, Hofmann MG, Gesteland PH, Cannon JB, Lefler JS, et al. Short-term elevation of fine particulate matter air pollution and acute lower respiratory infection. Am J Respir Crit Care Med. 2018 Sep;198(6):759-66.
30. Su W, Wu X, Geng X, Zhao X, Liu Q, Liu T. The short-term effects of air pollutants on influenza-like illness in Jinan, China. BMC Public Health. 2019 Oct;19(1):1319.
31. Hyrkas H, Ikaheimo TM, Jaakkola JJ, Jaakkola MS. Asthma control and cold weather-related respiratory symptoms. Respir Med. 2016 Apr;113:1-7.
32. Hyrkas H, Jaakkola MS, Ikaheimo TM, Hugg TT, Jaakkola JJ. Asthma and allergic rhinitis increase respiratory symptoms in cold weather among young adults. Respir Med. 2014 Jan;108(1):63-70.
33. Koskela HO. Cold air-provoked respiratory symptoms: the mechanisms and management. Int J Circumpolar Health. 2007 Apr;66(2):91-100.
34. Lin L, Li T, Sun M, Liang Q, Ma Y, Wang F, et al. Effect of particulate matter exposure on the prevalence of allergic rhinitis in children: a systematic review and meta-analysis. Chemosphere. 2021 Apr;268: 128841.
35. Zou QY, Shen Y, Ke X, Hong SL, Kang HY. Exposure to air pollution and risk of prevalence of childhood allergic rhinitis: a meta-analysis. Int J Pediatr Otorhinolaryngol. 2018 Sep;112:82-90.
36. Nhung NT, Schindler C, Dien TM, Probst-Hensch N, Kunzli N. Association of ambient air pollution with lengths of hospital stay for hanoi children with acute lower-respiratory infection, 2007-2016. Environ Pollut. 2019 Apr;247:752-62.
37. Zhu Y, Xie J, Huang F, Cao L. Association between short-term exposure to air pollution and COVID-19 infection: evidence from China. Sci Total Environ. 2020 Jul;727:138704.
38. Liu Y, Liu J, Chen F, Shamsi BH, Wang Q, Jiao F, et al. Impact of meteorological factors on lower respiratory tract infections in children. J Int Med Res. 2016 Feb;44(1):30-41.
39. Tasci SS, Kavalci C, Kayipmaz AE. Relationship of meteorological and air pollution parameters with pneumonia in elderly patients. Emerg Med Int. 2018 Mar;2018:4183203.
40. Banerjee A, Dunson DB, Tokdar ST. Efficient Gaussian process regression for large datasets. Biometrika. 2013 Mar;100(1):75-89.