Environmental Science and Pollution Research (2023) 30:37440–37448
https://doi.org/10.1007/s11356-022-24956-9

RESEARCH ARTICLE

LSTM model for predicting the daily number of asthma patients in Seoul, South Korea, using meteorological and air pollution data

Munyoung Chang[1,2] · **Yunseo Ku**[3]

Received: 29 August 2022 / Accepted: 20 December 2022 / Published online: 27 December 2022

Abstract

Asthma is a common respiratory disease that is affected by air pollutants and meteorological factors. In this study, we developed models that predict the daily number of patients receiving treatment for asthma using air pollution and meteorological data. A neural network with long short-term memory (LSTM) and fully connected (FC) layers was used. The daily number of asthma patients in the city of Seoul, the capital of South Korea, was collected from the National Health Insurance Service. The data from 2015 to 2018 were used as the training and validation datasets for model development. Unseen data from 2019 were used for testing. The daily number of asthma patients per 100,000 inhabitants was predicted. The LSTM-FC neural network model achieved a Pearson correlation coefficient of 0.984 ($P < 0.001$) and root mean square error of 3.472 between the predicted and original values on the unseen testing dataset. The factors that impacted the prediction were the number of asthma patients in the previous time step before the predicted date, type of day (regular day and day after a holiday), minimum temperature, SO_2, daily changes in the amount of cloud, and daily changes in diurnal temperature range. We successfully developed a neural network that predicts the onset and exacerbation of asthma, and we identified the crucial influencing air pollutants and meteorological factors. This study will help us to establish appropriate measures according to the daily predicted number of asthma patients and reduce the daily onset and exacerbation of asthma in the susceptible population.

Keywords Asthma · Long short-term memory · LSTM · Air pollution · Meteorological factor · Prediction

Introduction

Asthma is a common respiratory disease characterized by variable expiratory airflow limitation (Reddel et al. 2021). It presents with multiple time-varying respiratory symptoms, such as wheezing, coughing, shortness of breath, and chest tightness. Asthma is widespread and affects more than 300 million people in the world. Moreover, asthma is one of the leading causes of disability; it significantly lowers quality of life and places a great burden on society and the economy (Global Asthma Network 2018). In particular, a recently published study reported that the prevalence of asthma in South Korea is steadily increasing (Lee et al. 2020). In addition, it was reported that asthma-related medical service consumption and cost are increasing, and deaths due to asthma are also increasing. Therefore, reducing the onset and exacerbation of asthma would be of great benefit to health, society, and the economy.

Asthma affects the respiratory system, which is in direct contact with outside air; therefore, the external environment can cause both the onset and exacerbation of asthma. Many studies have reported that outdoor air pollution can exacerbate asthma (Guarnieri and Balmes 2014; Tiotiu et al. 2020). According to a meta-analysis, levels of nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), and particulate matter are reportedly significantly correlated with asthma exacerbation (Orellano et al. 2017). In addition, air pollution influences the onset of asthma (Guarnieri and Balmes 2014). Asthma is also influenced

Responsible Editor: Lotfi Aleya

✉ Munyoung Chang
cadu01@cau.ac.kr

[1] Department of Otorhinolaryngology-Head and Neck Surgery, Chung-Ang University College of Medicine, 84 Heukseok-Ro, Dongjak-Gu, 06974 Seoul, South Korea

[2] Department of Electrical and Computer Engineering, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, 08826 Seoul, South Korea

[3] Department of Biomedical Engineering, Chungnam National University College of Medicine, 99 Daehak-Ro, Yuseong-Gu, 34134 Daejeon, South Korea

by meteorological factors; for example, cold weather can worsen respiratory symptoms in patients who have asthma (Driessen et al. 2012; Hyrkäs et al. 2016; Koskela 2007). Meteorological factors influence the incidence of respiratory infections, which influence asthma exacerbation (Liu et al. 2017; Taussig et al. 2003).

In addition to appropriate medication, it is crucial to identify and adjust the modifiable risk factors of asthma to properly manage asthma (Reddel et al. 2021). Therefore, predicting the onset and exacerbation of asthma, as well as taking appropriate measures to prevent them, would be of considerable help in managing asthma. Recently, machine learning methods have been used in various fields because they make accurate predictions (Jordan and Mitchell 2015). In particular, in the medical field, machine learning has shown excellent performance in image-based diagnosis and disease progression prediction (Anderson et al. 2015; Esteva et al. 2017; Sidey-Gibbons and Sidey-Gibbons 2019). Therefore, in this study, we developed, using machine learning methods, models that predict the onset and exacerbation of asthma from air pollution and meteorological data.

Methods

Study design

The study period was from January 2015 to December 2019. The study area was limited to Seoul city to minimize the effect of regional differences in air pollutants and meteorological factors. Seoul, the capital of South Korea, is the largest city in South Korea and has a population of 10 million. The daily number of patients treated for asthma per 100,000 inhabitants of Seoul was calculated using the National Health Insurance Service (NHIS) data. We collected data on the daily air pollutants and meteorological factors in Seoul. Because the number of available medical institutions varies by the type of day (regular day, holiday, or day after a holiday), we included the type of day as an input feature using one-hot encoding. The analysis of variance (ANOVA) method was used to verify whether the daily number of patients treated for asthma significantly differed according to the type of day.

We developed machine learning models using long short-term memory (LSTM) and a fully connected (FC) neural network to predict the daily number of patients treated for asthma per 100,000 inhabitants of Seoul (Chen and Guestrin 2016; Hochreiter and Schmidhuber 1997). The prediction model development proceeded as follows: data preprocessing, feature selection (multicollinearity control, development of the baseline model, and stepwise feature selection), and development of the final model. The statistical analyses and the development of the machine learning models were performed using IBM SPSS Statistics 26 (IBM Corp., Armonk, NY, USA) and Google Colaboratory (Alphabet Inc., Mountain View, CA, USA).

Asthma cases

In South Korea, all citizens are enrolled in the NHIS, and whenever citizens use medical institutions, their diagnostic information is stored in the NHIS database. From this database, the number of patients receiving treatment for a specific disease can be obtained. The NHIS posts data for the daily number of patients treated for asthma on the public data portal website (https://www.data.go.kr/), with the primary diagnosis under the following Korean Classification of Diseases codes: asthma (J45) and status asthmaticus (J46). The annual population of Seoul is listed on the Seoul Metropolitan Government website (http://data.seoul.go.kr/dataList/419/S/2/datasetView.do). We used these data to calculate the daily number of patients treated for asthma per 100,000 inhabitants of Seoul. The number of patients treated for asthma was used as an indicator of asthma onset and exacerbation.

Meteorological and air pollution data

We obtained data for 15 daily meteorological factors (minimum temperature, maximum temperature, average temperature, diurnal temperature range (difference between the maximum and minimum temperatures within a day), average humidity, minimum humidity, daylight duration, sunshine duration, maximum solar insolation amount per hour, total solar insolation amount, precipitation, average wind speed, maximum wind speed, atmospheric pressure, and the amount of cloud) for Seoul from the Korea Meteorological Administration website (http://www.kma.go.kr/eng/index.jsp) and 6 daily air pollution factors (particulate matter smaller than 2.5 μm in aerodynamic diameter ($PM_{2.5}$), particulate matter smaller than 10 μm in aerodynamic diameter (PM_{10}), O_3, NO_2, CO, and SO_2) for Seoul from the Air Korea website (https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123). Daily changes in each of the air pollutants and meteorological factors were calculated as follows: current day level − previous day level (Supplementary Fig. 1).

Data preprocessing

The daily meteorological factor and air pollutant levels and their daily changes and the daily number of asthma patients in the previous time step before the predicted date were used to develop the prediction model. In addition, the type of day (regular day, holiday, or day after a holiday) was used. For the type of day, the data from the same period, including

the data of the day to be predicted, were used as input features because the type of day to be predicted was already known. For example, if the time step is 1 week, the number of asthma patients on January 12, 2015, was predicted using the daily data for air pollutants and meteorological factors and the daily number of patients treated for asthma from January 5, 2015, to January 11, 2015, and the daily data for the type of day from January 6, 2015, to January 12, 2015.

The data from 2015 to 2017 were used as the training dataset to develop the neural network model, whereas the unseen data from 2018 were used as the validation dataset to validate the trained model and to modify the network structure and hyperparameters. The unseen data from 2019 were used as the testing dataset to evaluate the performance of the trained models. Finally, min–max normalization was performed on all data using the maximum and minimum values of the training dataset to remove the influence of the different ranges of features. In the case of PM_{10}, the level on May 23, 2015, was 568.7 μg/m^3, which was unusually high compared with other days. With May 23, 2015, excluded, the range of PM_{10} levels was 4.7–248.5 μg/m^3. Therefore, in the case of PM_{10}, min–max normalization was performed by setting the next maximum value of 248.5 μg/m^3 as the maximum value instead of 568.7 μg/m^3.

Feature selection

Feature selection consisted of three steps: multicollinearity control, development of the baseline model, and stepwise feature selection. In the first step, we used the Mutual Information-Variance Inflation Factor (MI-VIF) method, which was proposed by Cheng et al. to control multicollinearity (Cheng et al. 2022). Briefly, the mutual information between an output feature and each input feature was calculated. The input features were arranged in order of highest to lowest mutual information. The candidate set was constructed by adding input features, one by one, in order. Whenever a candidate set was constructed, the VIFs of the included input features were calculated. If there was an input feature whose VIF was greater than the threshold value, the input feature last added to the candidate set was excluded. We set the VIF threshold at 5. In this way, the first feature set with a VIF of less than 5 for all included input features was created.

In the second step, a baseline model was constructed using the selected features from the previous step. A model was created using a combination of LSTM layers and FC layers (Yun 2020). The number of layers in the FC structure was set to two. The number of units in the last layer of the FC layers was set to one. To determine the most suitable network structure, 48 total conditions were defined that combined the learning rate (0.01, 0.001, and 0.0001), the length of the time step (1 week and 2, 3, 4, 5, 6, 7, and 8 weeks), and the number of layers in the LSTM structure (1 and 2). A framework called Optuna (Akiba et al. 2019) was used to find the best combination of the following hyperparameters in each condition: number of units (750 to 2000, step = 250), dropout rate (0.2 to 0.6, step = 0.1) in the LSTM and FC structures, and batch size (32 to 256, step = 32) (Winastwan 2021). Adam was used as the optimizer. The activation functions of the LSTM and FC layers were the tanh and rectified linear unit (ReLU) functions, respectively. Of these models, the model that exhibited the lowest root mean square error (RMSE) between the original and predicted values on the validation dataset was selected as the baseline model.

In the third step, stepwise feature selection was performed using the previously developed baseline model and the feature set obtained in the previous step. Forward and backward feature selection was repeated based on RMSE. The RMSE threshold for stopping the process was set to 0.002. The feature set exhibiting the lowest RMSE between the original and predicted values in the validation dataset was selected as the final feature set.

Development of final model using LSTM-FC neural network

A final model was developed using the previously selected final feature set. The neural network model comprised the LSTM layers and two FC layers. As before, the most suitable network structure and hyperparameters were determined using Optuna (Akiba et al. 2019; Winastwan 2021). The model that exhibited the lowest RMSE between the original and predicted values in the validation dataset was selected as the final model.

We applied the final model to the testing dataset to obtain the predicted values and to evaluate the performance by obtaining the RMSE and Pearson correlation coefficient between the predicted and original values. In addition, Shapley additive explanations (SHAP) were used to evaluate the contribution of the input features to the predictive model (Lundberg 2017). SHAP is based on Shapley values, which are the average expected marginal contributions of one feature after all possible combinations are considered (Shapley and Roth 1988).

Ethical statement

The study was reviewed and approved by the Institutional Review Board of Chung-Ang University Hospital (2111–050-19,393). The requirement for consent to participate was waived.

Results

Feature selection

The averages of the daily number of asthma patients and meteorological and air pollution factors are presented in

Table 1. The average of the daily number of asthma patients for all periods was 45.3 (95% confidence interval (CI) 44.3 to 46.3). The average of the daily number of asthma patients for regular day, holiday, and day after a holiday was 49.8 (95% CI 49.1 to 50.4), 7.0 (95% CI 6.3 to 7.7), and 71.3 (95% CI 69.5 to 73.2), respectively. The daily number of asthma patients differed significantly depending on the type of day ($P < 0.001$).

In the multicollinearity control step, we intended to create a feature set that was highly related to the output feature and that had a VIF of less than 5 for all included input features. Using the MI-VIF method, a feature set consisting of 18 features, including the output feature, was created: number of asthma patients, regular day, day after a holiday, minimum temperature, minimum humidity, daylight duration, atmospheric pressure, the amount of cloud, NO_2, SO_2, PM_{10}, and daily change in diurnal temperature range, total solar insolation amount, average wind speed, the amount of cloud, precipitation, NO_2, and $PM_{2.5}$.

A baseline model was developed with this feature set using the Optuna framework. The results for the Optuna framework are presented in Table S1. The final baseline model, which exhibited the lowest RMSE (0.03423) for the validation dataset, contained one layer in the LSTM and two layers in the FC structures. The number of units and dropout rates in the layers of the LSTM and FC structures were 1500 and 0.3 and 1500 and 0.2, respectively. The number of units in the last FC layer was fixed to one. The learning rate was 0.001, and the length of the time step was 7 weeks. Stepwise feature selection was performed using the final baseline model. The feature set that included number of asthma patients in the previous episode, regular day, day after a holiday, SO_2, minimum temperature, daily change in diurnal temperature range, and daily change in the amount of cloud showed the lowest RMSE (0.03041) on the validation dataset. This feature set was selected as the final feature set.

Development of final model using LSTM-FC neural network

The final model was determined with the final feature set using the Optuna framework. The results for the Optuna

Table 1 Number of patients treated for asthma, air pollutants, and meteorological factors

	Whole data set, 2015–2019 (95% CI)	Training set, 2015–2017 (95% CI)	Validation set, 2018 (95% CI)	Testing set, 2019 (95% CI)
Minimum temperature (°C)	9.2 (8.7, 9.7)	9.2 (8.6, 9.9)	8.9 (7.7, 10)	9.3 (8.3, 10.4)
Maximum temperature (°C)	18.4 (17.9, 18.9)	18.5 (17.8, 19.1)	18.0 (16.7, 19.1)	18.6 (17.5, 19.8)
Average temperature (°C)	13.4 (12.9, 13.9)	13.4 (12.8, 14.1)	13.0 (11.8, 14.1)	13.6 (12.6, 14.7)
Diurnal temperature range (°C)	9.2 (9.1, 9.4)	9.3 (9.1, 9.5)	9.1 (8.8, 9.4)	9.3 (9, 9.6)
Daylight duration (h)	12.2 (12.1, 12.3)	12.2 (12.1, 12.3)	12.2 (12, 12.4)	12.2 (12, 12.4)
Sunshine duration (h)	7.0 (6.9, 7.2)	7.0 (6.8, 7.3)	7.1 (6.7, 7.5)	7.0 (6.5, 7.4)
Maximum solar insolation amount per hour (MJ/m^2)	2.0 (2.0, 2.0)	1.9 (1.9, 1.9)	2.1 (2.0, 2.2)	2.1 (2.0, 2.2)
Total solar insolation amount (MJ/m^2)	13.1 (12.8, 13.4)	12.6 (12.2, 13)	13.9 (13.1, 14.7)	13.8 (13.1, 14.5)
Minimum humidity (%)	36.3 (35.6, 37)	37.1 (36.2, 38)	35.4 (33.8, 36.9)	34.8 (33.2, 36.3)
Average humidity (%)	58.2 (57.5, 58.9)	58.9 (58.1, 59.8)	57.5 (55.9, 59)	56.8 (55.2, 58.4)
Atmospheric pressure (hPa)	1006.2 (1005.8, 1006.5)	1006.2 (1005.7, 1006.6)	1006.5 (1005.7, 1007.4)	1006.0 (1005.2, 1006.8)
Maximum wind speed (m/s)	4.7 (4.6, 4.7)	5.1 (5, 5.1)	4.0 (3.9, 4.1)	4.3 (4.2, 4.4)
Average wind speed (m/s)	2.2 (2.1, 2.2)	2.4 (2.3, 2.4)	1.7 (1.7, 1.8)	2.0 (1.9, 2)
Cloud amount (/10)	4.7 (4.6, 4.9)	4.7 (4.5, 4.9)	4.7 (4.4, 5.1)	4.9 (4.6, 5.2)
Precipitation (mm)	2.8 (2.4, 3.4)	2.8 (2.1, 3.4)	3.5 (2.4, 4.8)	2.4 (1.7, 3.4)
$PM_{2.5}$ ($\mu g/m^3$)	24.3 (23.6, 24.9)	24.6 (23.9, 25.3)	22.8 (21.1, 24.5)	24.7 (22.9, 26.5)
PM_{10} ($\mu g/m^3$)	43.6 (42.4, 44.8)	45.6 (44, 47.4)	39.7 (37.2, 42.2)	41.4 (38.7, 44.2)
O_3 (ppm)	0.024 (0.023, 0.024)	0.024 (0.023, 0.024)	0.023 (0.022, 0.025)	0.025 (0.024, 0.027)
NO_2 (ppm)	0.030 (0.030, 0.030)	0.031 (0.030, 0.032)	0.028 (0.027, 0.030)	0.028 (0.027, 0.030)
CO (ppm)	0.519 (0.511, 0.526)	0.5214 (0.5111, 0.5322)	0.504 (0.487, 0.522)	0.525 (0.508, 0.543)
SO_2 (ppm)	0.005 (0.005, 0.005)	0.005 (0.005, 0.005)	0.004 (0.004, 0.005)	0.004 (0.004, 0.004)
No. of asthma pts. (/100,000)	45.3 (44.3, 46.3)	47.7 (46.3, 49.2)	43.1 (40.7, 45.4)	40.4 (38.4, 42.3)

CI, confidence interval; $PM_{2.5}$, particulate matter smaller than 2.5 μm in aerodynamic diameter; PM_{10}, particulate matter smaller than 10 μm in aerodynamic diameter; O_3, ozone; NO_2, nitrogen dioxide; *CO*, carbon monoxide; SO_2, sulfur dioxide; *pts*, patients

framework are presented in Table S2. The final model, which exhibited the lowest RMSE (0.03024) on the validation dataset, contained one layer in the LSTM and two layers in the FC structures. The number of units and dropout rates in the layers were 1000 and 0.2, respectively, for the LSTM structure, and 1250 and 0.4, respectively, for the FC structure. The number of units in the last FC layer was set to one. The learning rate was 0.001 and the length of the time step was 7 weeks.

Figure 1 shows the prediction results for the daily number of asthma patients in the unseen testing dataset. The RMSE between the original and predicted values in the form of min–max normalized values was 0.03037. After converting the min–max normalized values to the actual values, the RMSE between the original and predicted values for daily number of asthma patients per 100,000 inhabitants was 3.47190. The correlation coefficient between the original and predicted values was 0.98432 ($P < 0.001$).

The SHAP analysis revealed the contribution of each feature. The larger the bar size, the greater the contribution of the feature. As shown in Fig. 2, features were found to have effects on the prediction of the number of asthma patients in the order of regular day, day after a holiday, number of asthma patients in the previous time step, daily change in the amount of cloud, minimum temperature, SO_2, and daily change in diurnal temperature range. It was found that the number of asthma patients was high when it was the day after the holiday and when the number of asthma patients in the previous time step was large. It was found that the lower the daily change in the amount of cloud, minimum temperature, and daily change in diurnal temperature range, the greater the number of asthma patients.

Discussion

We developed a model for predicting the daily number of patients receiving treatment for asthma using machine learning techniques: the LSTM-FC neural network, a type of recurrent neural network (RNN). The model included number of asthma patients in the previous time step, the type of day (regular day and day after a holiday), SO_2, minimum temperature, daily change in diurnal temperature range, and daily change in the amount of cloud as input features. When this model was applied to the unseen dataset, it showed satisfactory results with an RMSE of 3.47190 and a correlation coefficient of 0.98432 ($P < 0.001$) between the original and predicted values.

An RNN stores information about the previous state in the form of memory and uses it to solve problems. This makes RNNs powerful for analyzing time-series data. However, using a state in the distant past is disadvantageous, as it leads to the vanishing gradient problem, and RNNs have difficulty managing long-term dependencies (Bengio et al. 1994). LSTMs are designed to solve these problems (Hochreiter and Schmidhuber 1997). They include an input, output, and forget gates that control the amount of exposed memory and better deal with long-term dependency problems. Many studies have predicted disease incidence using LSTMs (Tsan et al. 2022; Zhao et al. 2021). Tsan et al. (2022) developed a model to predict the incidence of influenza-like illnesses and respiratory diseases using LSTM and the autoregressive integrated moving average method. They reported that the performance of the model using LSTM was superior. Additionally, Zhao et al. (2021) developed a model to predict the incidence of bronchopneumonia in children using LSTM-FC, FC, regression of support vector machine, and random forest methods. They also reported that LSTM exhibited the best performance. We used an LSTM-FC neural network to predict the daily number of asthma patients, and we achieved successful results.

There is a close interaction between air pollutants and meteorological factors. Therefore, multicollinearity control is important in developing predictive models using air pollutants and meteorological factors. The MI-VIF method was used to preserve the input features that had a close relationship with the output feature while at the same time avoiding

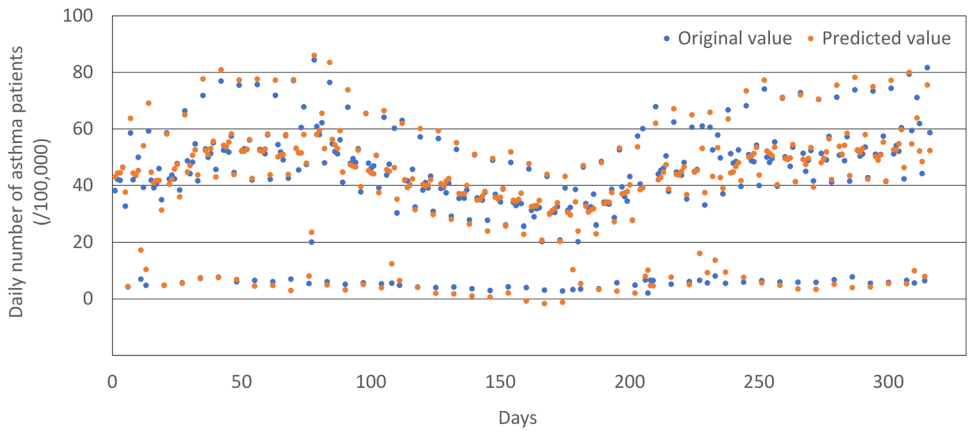

Fig. 1 Prediction results on unseen testing dataset for daily number of patients treated for asthma. The correlation coefficient between the original and predicted values is 0.98432 ($P < 0.001$). LSTM, long short-term memory; FC, fully connected; pts, patients

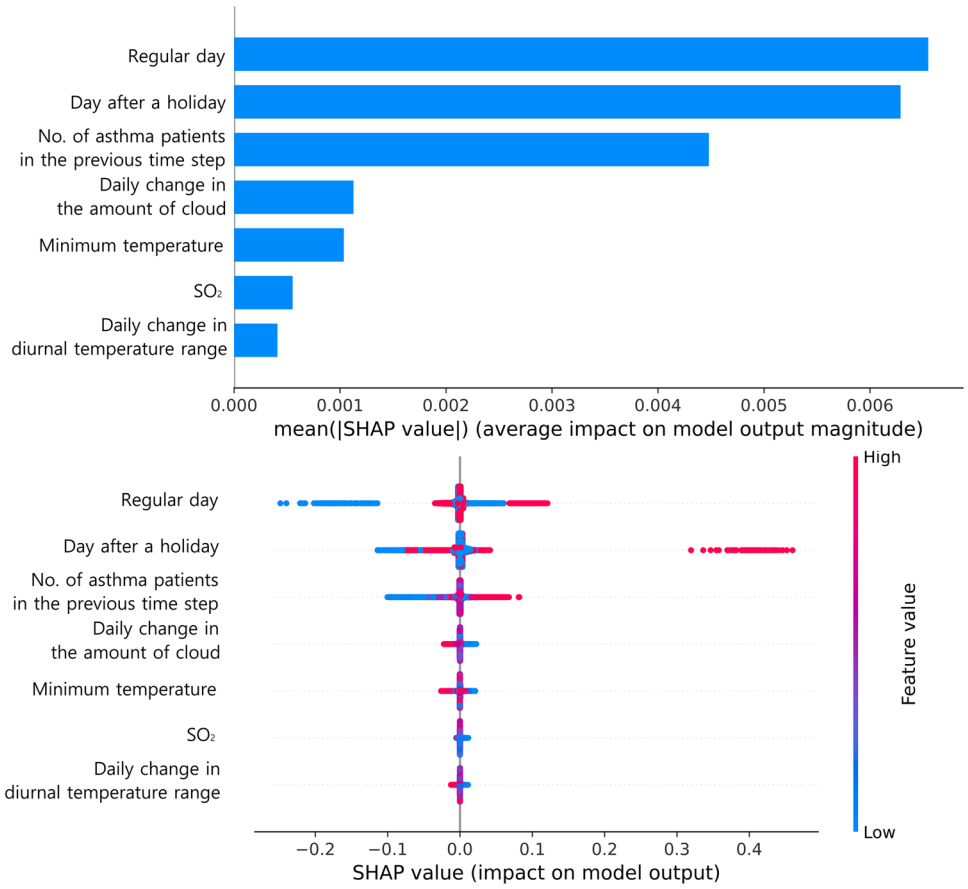

Fig. 2 Results of SHAP analysis in LSTM-FC neural network model. SHAP, Shapley additive explanations; LSTM, long short-term memory; FC, fully connected

features with a VIF of 5 or higher (Cheng et al. 2022). By including features that had a close relationship with the output feature in the feature set first, the probability that features having close relationships with the output feature were excluded from the feature set was reduced.

Stepwise feature selection was performed using the baseline model, which was developed with 18 features, including the number of asthma patients. As a result, the best performance was achieved when only 7 features, including the number of asthma patients, were used. When 18 features were used, the RMSE in the validation dataset was 0.03423, but when only 7 features were used, the RMSE in the validation dataset improved to 0.03041. By reducing the number of features used in the model, both the performance and interpretability of the model were improved, and the possibility of overfitting was reduced (Kuhn and Johnson 2019).

We used the Optuna framework to determine the best structure and hyperparameters for the LSTM-FC neural network. Because deep neural networks can have various structures, it is impossible to manually test all conditions. Therefore, platforms have been developed for hyperparameter optimization, e.g., Optuna, that enable the construction of organized experiments and provide several algorithms for the sampler and pruner (Akiba et al. 2019). Optuna has been used to optimize models for disease diagnosis and prediction (Barros et al. 2021; Krivorotko et al. 2022; Lacerda et al. 2021). In our study, 48 conditions were set according to the learning rate, number of layers, and time step to reduce the computational load of each session. The structures and hyperparameters that showed the best performance under each condition were determined.

The lagged effect of air pollutants and meteorological factors on asthma has been reported in several studies (Chen et al. 2022; Lam et al. 2016; Liu et al. 2019; Zhang et al. 2014). A recent study by Chen et al. reported that the lagged effects of meteorological factors on hospitalization for asthma persisted for 3–4 weeks (Chen et al. 2022). In consideration of the lagged effects, a model was tested by setting the time step variously to be reflected in the LSTM-FC neural network, that is, the period of the previous state to be reflected in the prediction. In our study, the optimal time step was determined to be 7 weeks, which was longer than in previous studies (Chen et al. 2022). This difference is presumed to have occurred because our model used the number of asthma patients in the previous period as well as air pollutants and meteorological factors to predict the number of asthma patients.

The SHAP analysis revealed the contribution of each feature. As expected, types of days (regular day and day after a holiday) and number of asthma patients in the previous

time step played an important role in predicting the number of asthma patients. It is estimated that the number of asthma patients on holidays decreases because the number of medical institutions providing treatment on holidays is small. As a result, it is estimated that the number of patients increases the day after a holiday and then maintains a normal level on regular days.

Among the meteorological factors, minimum temperature was selected as an important feature for predicting the number of asthma patients. Several studies have reported that cold weather can aggravate asthma symptoms (Hyrkäs et al. 2016, 2014; Näyhä et al. 2011). This is presumably caused by inhaling cold air, which causes adverse functional changes in the upper and lower airways (Cruz and Togias 2008; Koskela and Tukiainen 1995; Koskela 2007). It has also been reported that these changes may induce bronchoconstriction (Koskela 2007). Our model showed the same results, indicating that the lower the minimum temperature, the greater the number of asthma patients.

Of the air pollution factors, SO_2 was selected as an important feature for predicting the number of asthma patients. Several studies have reported that SO_2 penetrates the lungs and causes bronchoconstriction, eventually reducing lung function and exacerbating symptoms in asthma patients (Tiotiu et al. 2020). However, from an observation of the results of the SHAP analysis, the direction of the effect of SO_2 on the number of asthma patients in our model was not clear. This is presumably because SO_2 interacts with other air pollutants and meteorological factors.

In our study, daily changes (current day level – previous day level) in air pollutants and meteorological factors were used as input features. Several studies have reported the effect of daily change in meteorological factors on diseases (Guo et al. 2011; Kim et al. 2014; Lin et al. 2013; Wasilevich et al. 2012). Kim et al. reported that an absolute difference in average temperature between a current day and the previous day was associated with an increase in the number of emergency room visits due to asthma (Kim et al. 2014). This could be because the larger the daily change, the more difficult it is to adapt, which could exacerbate asthma symptoms. However, it was reported that the opposite trend was observed in winter. As the absolute difference of average temperature increased, the number of emergency room visits due to asthma decreased in winter. The authors explained that this phenomenon occurred because asthma patients mostly stayed indoors when the temperature dropped considerably. Similar results were obtained by our model. Daily changes in the diurnal temperature range and the amount of cloud were selected as important features for predicting the number of asthma patients. As daily changes in diurnal temperature range and the amount of cloud increased, the number of asthma patients decreased. It is presumed that this is because, when the diurnal temperature range or the amount of cloud increases compared with the previous day, outdoor activities decrease and allergen exposure decreases. In the case of diurnal temperature range and the amount of cloud, the daily change was found to play a more important role in the prediction of the number of asthma patients than corresponding daily levels. In future studies that use air pollutants and meteorological factors, daily changes in air pollutants and meteorological factors should be considered.

This study has some limitations. We obtained the daily number of patients treated for asthma from the NHIS data. These data included patients who visit the hospital for regular medication pickups. These visits might be unrelated to changes in the asthma conditions and relatively less affected by air pollutants and meteorological factors. Since we were unable to distinguish such visits from other data, we developed a prediction model using data that included such visits. However, because these visits for regular medication pickups might be uniformly distributed over the entire period, these data might not significantly affect the model performance. Additionally, some undiscovered features may be related to the number of patients with asthma, such as seasonal differences. Future studies are needed to enhance a better understanding of undiscovered features that could be related to the number of patients with asthma.

Conclusions

We used LSTM to successfully develop a model to predict the daily number of patients treated for asthma, and we identified the crucial influencing air pollutants and meteorological factors. This analysis will help us to establish appropriate measures according to the daily predicted number of asthma patients and to reduce the daily onset and exacerbation of asthma in the susceptible population.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-022-24956-9.

Author contribution Munyoung Chang: project administration, methodology, writing, investigation, writing—review and editing, supervision; Yunseo Ku: investigation, writing—review and editing. All authors have given approval to the final version of the manuscript.

Data availability The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical approval The study was reviewed and approved by the Institutional Review Board of Chung-Ang University Hospital (2111–050-19393).

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2623–2631

Anderson JP, Parikh JR, Shenfeld DK, Ivanov V, Marks C, Church BW, Laramie JM, Mardekian J, Piper BA, Willke RJ, Rublee DA (2015) Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. J Diabetes Sci Technol 10:6–18. https://doi.org/10.1177/1932296815620200

Barros B, Lacerda P, Albuquerque C, Conci A (2021) Pulmonary COVID-19: learning spatiotemporal features combining CNN and LSTM networks for lung ultrasound video classification. Sensors (Basel) 2110.3390/s21165486

Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5:157–166. https://doi.org/10.1109/72.279181

Chen Y, Kong D, Fu J, Zhang Y, Zhao Y, Liu Y, Chang Z, Liu Y, Liu X, Xu K, Jiang C, Fan Z (2022) Associations between ambient temperature and adult asthma hospitalizations in Beijing, China: a time-stratified case-crossover study. Respir Res 23:38. https://doi.org/10.1186/s12931-022-01960-8

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. arXiv 1603, 02754v02753

Cheng J, Sun J, Yao K, Xu M, Cao Y (2022) A variable selection method based on mutual information and variance inflation factor. Spectrochim Acta A Mol Biomol Spectrosc 268:120652. https://doi.org/10.1016/j.saa.2021.120652

Cruz AA, Togias A (2008) Upper airways reactions to cold air. Curr Allergy Asthma Rep 8:111–117. https://doi.org/10.1007/s11882-008-0020-z

Driessen JM, van der Palen J, van Aalderen WM, de Jongh FH, Thio BJ (2012) Inspiratory airflow limitation after exercise challenge in cold air in asthmatic children. Respir Med 106:1362–1368. https://doi.org/10.1016/j.rmed.2012.06.017

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056

Global Asthma Network (2018) The global asthma report 2018, Auckland, New Zealand

Guarnieri M, Balmes JR (2014) Outdoor air pollution and asthma. Lancet 383:1581–1592. https://doi.org/10.1016/s0140-6736(14)60617-6

Guo Y, Barnett AG, Yu W, Pan X, Ye X, Huang C, Tong S (2011) A large change in temperature between neighbouring days increases the risk of mortality. PLoS ONE 6:e16511. https://doi.org/10.1371/journal.pone.0016511

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyrkäs H, Jaakkola MS, Ikäheimo TM, Hugg TT, Jaakkola JJ (2014) Asthma and allergic rhinitis increase respiratory symptoms in cold weather among young adults. Respir Med 108:63–70. https://doi.org/10.1016/j.rmed.2013.10.019

Hyrkäs H, Ikäheimo TM, Jaakkola JJ, Jaakkola MS (2016) Asthma control and cold weather-related respiratory symptoms. Respir Med 113:1–7. https://doi.org/10.1016/j.rmed.2016.02.005

Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349:255–260. https://doi.org/10.1126/science.aaa8415

Kim J, Lim Y, Kim H (2014) Outdoor temperature changes and emergency department visits for asthma in Seoul, Korea: a time-series study. Environ Res 135:15–20. https://doi.org/10.1016/j.envres.2014.07.032

Koskela HO (2007) Cold air-provoked respiratory symptoms: the mechanisms and management. Int J Circumpolar Health 66:91–100. https://doi.org/10.3402/ijch.v66i2.18237

Koskela H, Tukiainen H (1995) Facial cooling, but not nasal breathing of cold air, induces bronchoconstriction: a study in asthmatic and healthy subjects. Eur Respir J 8:2088–2093. https://doi.org/10.1183/09031936.95.08122088

Krivorotko O, Sosnovskaia M, Vashchenko I, Kerr C, Lesnic D (2022) Agent-based modeling of COVID-19 outbreaks for New York state and UK: parameter identification algorithm. Infect Dis Model 7:30–44. https://doi.org/10.1016/j.idm.2021.11.004

Kuhn M, Johnson K (2019) Feature engineering and selection: a practical approach for predictive models. Chapman & Hall/CRC, London

Lacerda P, Barros B, Albuquerque C, Conci A (2021) Hyperparameter optimization for COVID-19 pneumonia diagnosis based on chest CT. Sensors (Basel) 2110.3390/s21062174

Lam HC, Li AM, Chan EY, Goggins WB 3rd (2016) The short-term association between asthma hospitalisations, ambient temperature, other meteorological factors and air pollutants in Hong Kong: a time-series study. Thorax 71:1097–1109. https://doi.org/10.1136/thoraxjnl-2015-208054

Lee E, Kim A, Ye YM, Choi SE, Park HS (2020) Increasing prevalence and mortality of asthma with age in Korea, 2002–2015: a nationwide, population-based study. Allergy Asthma Immunol Res 12:467–484. https://doi.org/10.4168/aair.2020.12.3.467

Lin H, Zhang Y, Xu Y, Xu X, Liu T, Luo Y, Xiao J, Wu W, Ma W (2013) Temperature changes between neighboring days and mortality in summer: a distributed lag non-linear time series analysis. PLoS ONE 8:e66403. https://doi.org/10.1371/journal.pone.0066403

Liu L, Pan Y, Zhu Y, Song Y, Su X, Yang L, Li M (2017) Association between rhinovirus wheezing illness and the development of childhood asthma: a meta-analysis. BMJ Open 7:e013034. https://doi.org/10.1136/bmjopen-2016-013034

Liu F, Qu F, Zhang H, Chao L, Li R, Yu F, Guan J, Yan X (2019) The effect and burden modification of heating on adult asthma hospitalizations in Shijiazhuang: a time-series analysis. Respir Res 20:122. https://doi.org/10.1186/s12931-019-1092-0

Lundberg SM LS-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30 [Internet], 4765–4774

Näyhä S, Hassi J, Jousilahti P, Laatikainen T, Ikäheimo TM (2011) Cold-related symptoms among the healthy and sick of the general population: national FINRISK study data, 2002. Public Health 125:380–388. https://doi.org/10.1016/j.puhe.2011.02.014

Orellano P, Quaranta N, Reynoso J, Balbi B, Vasquez J (2017) Effect of outdoor air pollution on asthma exacerbations in children and adults: systematic review and multilevel meta-analysis. PLoS ONE 12:e0174050. https://doi.org/10.1371/journal.pone.0174050

Reddel HK et al (2021) Global Initiative for Asthma (GINA) strategy 2021 - executive summary and rationale for key changes. Eur Respir J. https://doi.org/10.1183/13993003.02730-2021

Shapley LS, Roth AE (1988) The Shapley value : essays in honor of Lloyd S. Cambridge University Press, Cambridge Cambridgeshire. New York, Shapley

Sidey-Gibbons JAM, Sidey-Gibbons CJ (2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19:64. https://doi.org/10.1186/s12874-019-0681-4

Taussig LM, Wright AL, Holberg CJ, Halonen M, Morgan WJ, Martinez FD (2003) Tucson children's respiratory study: 1980 to present. J Allergy Clin Immunol 111, 661–675; quiz 676. https://doi.org/10.1067/mai.2003.162

Tiotiu AI, Novakova P, Nedeva D, Chong-Neto HJ, Novakova S, Steiropoulos P, Kowal K (2020) Impact of air pollution on asthma outcomes. Int J Environ Res Public Health 17:6212. https://doi.org/10.3390/ijerph17176212

Tsan YT, Chen DY, Liu PY, Kristiani E, Nguyen KLP, Yang CT (2022) The prediction of influenza-like illness and respiratory disease using LSTM and ARIMA. Int J Environ Res Public Health 19:1858. https://doi.org/10.3390/ijerph19031858

Wasilevich EA, Rabito F, Lefante J, Johnson E (2012) Short-term outdoor temperature change and emergency department visits for asthma among children: a case-crossover study. Am J Epidemiol 176(Suppl 7):S123-130. https://doi.org/10.1093/aje/kws326

Winastwan R (2021) Hyperparameter tuning of neural networks with Optuna and PyTorch. Towards Data Science. https://towardsdatascience.com/hyperparameter-tuning-of-neural-networks-with-optuna-and-pytorch-22e179efc837. Accessed 2 Aug 2022

Yun YS (2020) Time series prediction walking with deep learning. Stock price prediction practice with Python, Keras, and TensorFlow. BJPUBLIC, Seoul (in Korean)

Zhang Y, Peng L, Kan H, Xu J, Chen R, Liu Y, Wang W (2014) Effects of meteorological factors on daily hospital admissions for asthma in adults: a time-series analysis. PLoS ONE 9:e102475. https://doi.org/10.1371/journal.pone.0102475

Zhao D, Chen M, Shi K, Ma M, Huang Y, Shen J (2021) A long short-term memory-fully connected (LSTM-FC) neural network for predicting the incidence of bronchopneumonia in children. Environ Sci Pollut Res Int 28:56892–56905. https://doi.org/10.1007/s11356-021-14632-9