Forecasting of non-accidental, cardiovascular, and respiratory mortality with environmental exposures adopting machine learning approaches

Woojoo Lee1 · Youn-Hee Lim2 · Eunhee Ha3 · Yoenjin Kim1 · Won Kyung Lee4

Received: 17 January 2022 / Accepted: 27 June 2022 / Published online: 14 July 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Environmental exposure constantly changes with time and various interactions that can affect health outcomes. Machine learning (ML) or deep learning (DL) algorithms have been used to solve complex problems, such as multiple exposures and their interactions. This study developed predictive models for cause-specific mortality using ML and DL algorithms with the daily or hourly measured meteorological and air pollution data. The ML algorithm improved the performance compared to the conventional methods, even though the optimal algorithm depended on the adverse health outcomes. The best algorithms were extreme gradient boosting, ridge, and elastic net, respectively, for non-accidental, cardiovascular, and respiratory mortality with daily measurement; they were superior to the generalized additive model reducing a mean absolute error by 4.7%, 4.9%, and 16.8%, respectively. With hourly measurements, the ML model tended to outperform the conventional models, even though hourly data, instead of daily data, did not enhance the performance in some models. The proposed model allows a better understanding and development of robust predictive models for health outcomes using multiple environmental exposures.

Keywords Cardiovascular diseases · Deep learning · Environmental exposures · Machine learning · Respiratory tract diseases

Introduction
Environmental exposures are constantly changing, but they have been simplified as a representative value when used in statistical models. For example, the mean temperature is commonly used as a representative temperature parameter in studies of mortality from meteorological factors. On the other hand, the mean temperature could not reflect all the information on the waveform temperature during a day. Therefore, previous research has found that in addition to the mean temperature, other values could also contribute to mortality. Hence, they included the maximum and minimum temperature, diurnal temperature range (DTR), and temperature variability in their models (Guo et al. 2016, Lee et al. 2017). On the other hand, no previous study has considered all the points of the waveform of exposures, even though meteorological stations measure the hourly meteorological data.

The health effects of environmental factors are complicated, and they interact with one another (Al Ahad et al. 2020). Air pollutants and meteorological factors have been suggested to have synergistic effects on health (Al Ahad et al. 2020, Li et al. 2017a). Recently, some studies reported that air pollutants have synergistic interactions with the temperature and season (Anenberg et al. 2020, Bergmann et al. 2020, Lee et al. 2019). They reported that particulate matter with...
an aerodynamic diameter of ≤10 μm (PM<sub>10</sub>) had a stronger influence on cardiovascular and non-accidental mortality in hot temperatures than in the median temperature (Li et al., 2017a). Another study reported that the season significantly altered the effects of sulfur dioxide (SO<sub>2</sub>) on cardiovascular disease and ozone (O<sub>3</sub>) on stroke and pneumonia (Bergmann et al. 2020). Hence, the performance in predicting the environmental effects on public health may be improved if the complex relationships between environmental exposures can be included in a model.

Machine learning (ML) is a better alternative to handle more covariates and their complex relationships between the correlated covariates (Lin et al. 2021). Although it interrupts the interpretation, ML can reflect these relationships when applying the ML algorithm to a prediction model. Although the ML algorithm has been applied increasingly in environmental research, there are limitations in specific topics of air pollution epidemiology (Bellinger et al. 2017, Benke and Benke, 2018, Li et al. 2017b, Zhang et al. 2020). A recent systematic review reported that most ML-based models were on the clustering and prediction of the air pollution distribution (Bellinger et al. 2017). The researchers suggested that deep learning (DL) might be a future direction to disentangle complex problems in environmental epidemiology.

In this study, ML-based predictive models were constructed to consider the interaction and hourly measurements of meteorological and air pollution data and compare the performance of these models with the corresponding models including daily measurements and the traditional models using the generalized linear model (GLM) and generalized additive model (GAM), including daily measurements.

**Materials and methods**

**Study population**

The inclusion criteria were residents in Seoul who died from non-accidental causes between 2015 and 2019. The deaths from cardiovascular and respiratory diseases were extracted from the non-accidental deaths for subgroup analysis. Deaths from intentional and unintentional injuries were excluded. The mortality data in Seoul were obtained from the Death Statistics Database of the Korean National Statistical Office. Seoul is the capital city and the largest metropolis of South Korea, with a population of 9.8 million people in 2021. Deaths from non-accidental causes, cardiovascular diseases, and respiratory diseases are defined as A00-R99, I00-I99, and J00-J99 using the International Classification of Disease, 10th revision codes. The Death Statistics Database includes information on the residential address, sex, and age. The daily frequency of mortality was constructed during the study period according to three causes of death. The current study was exempted from IRB review (2021-06-006) by the Inha University Hospital Institutional Review Board because it used de-identified data accessible to the public.

**Exposure assessment**

Meteorological data, including the ambient temperature and relative humidity, were extracted from the Korea Meteorological Administration. In Seoul, two meteorological stations measured the hourly meteorological data. Separate temperature and humidity variables were used in the model, which allowed flexibility and performed better than the apparent temperature in predicting mortality (Rodopoulou et al. 2015).

Air pollutants were provided by the Korean National Institute of Environmental Research: particulate matter with a diameter of ≤2.5 μm (PM<sub>2.5</sub>), PM<sub>10</sub>, nitrogen dioxide (NO<sub>2</sub>), SO<sub>2</sub>, O<sub>3</sub>, and carbon monoxide (CO). The hourly concentrations of air pollutants were measured at 40 monitors in Seoul. The 24-hour concentrations were summarized by averaging the monitor-specific concentrations within the city.

**Machine learning algorithms**

ML algorithms, such as the Random Forest (RF), extreme gradient boosting (XGBoost), logistic regressions regularized with an L1 penalty (lasso regression), an L2 penalty (ridge regression), and a net elastic penalty (elastic net), were applied to develop prediction models for non-accidental, cardiovascular, and respiratory mortality, respectively. Considering the time-series structure of the data, 10-fold cross-validation on a rolling basis was used to tune the hyperparameters for each prediction model with a mean absolute error (MAE) and the root mean square error (RMSE) as the evaluation criteria (Schnaubelt 2019). The hyperparameters in the RF were tuned by searching for all combinations of the number of trees (250, 500, 750, and 1,000) and the number of variables in a tree (p/6 to p/2 by 1, p = the number of input variables depending on the length of the time lags). For XGBoost, the current study searched for all combinations of the number of boosting iterations (100, 200, and 300) and the learning rate (0.05, 0.1, and 0.15), the minimum loss reduction required for an additional node partition (0, 0.05, and 0.1), subsampling ratio (0.5 and 0.75), the fraction of input variables to be randomly sampled (0.5, 0.66, and 0.75), the minimum sum of weights in the child (0, 1), and the maximum depth of the tree (3 to 10 by 1). Furthermore, lasso, ridge, and elastic net were tuned by λ = 10<sup>k</sup> (k moves from –5 to 2 by 1/10) for the daily measurements of exposure and (k moves from –3 to 3 by 1/15) for the hourly measurements.
The ML-based models considered single-day lags, the cumulative lags from zero to seven days, and the moving average lags for up to seven days as predictors after standardization. For each ML algorithm, the best model, whose cross-validated RMSE was the smallest, was selected among the various lag models.

Deep learning algorithms

Predictive models were also constructed using deep learning algorithms for time-series data: long short-term memory (LSTM) and stacked LSTM. The LSTM could handle long-term dependency problems using cell state \( c_t \), input gate \( i_t \), forget gate \( f_t \), and output gate \( o_t \) in the cell (Supplementary Fig. 1). The cell state plays a role in remembering the information over arbitrary time intervals. The three gates decide how much of the past information is remembered and how much of the new information is forgotten (Gu et al. 2019, Tian et al. 2018). There are three sigmoid functions, which are denoted by \( \sigma \) in Supplementary Fig. 1, used as the gating functions. They decide what information will be discarded from the cell state, what information will be saved in the cell state, and how much of the final cell state value to take.

The current study set the number of hidden layers to one and two for LSTM and stacked LSTM. The hyperparameters in the LSTM and stacked LSTM were tuned separately by searching for all combinations of the number of nodes (16, 32, 64, 128, 256, and 512), learning rate (0.001 and 0.005), drop-out rate (0.2, 0.3, and 0.5), batch size (100, 150, and 200), epoch (300), and sliding window size (from 2 to 15 by 1). The ADAM optimizer was used to minimize the RMSE.

In the LSTM and stacked LSTM, the lagged variables of environmental exposures were not included in the model because these deep learning architectures automatically discover the relevant time-lagged dependency between the environmental exposures and the outcome.

GAM and GLM

The Poisson GAM was used to predict mortality using meteorological and air pollution factors as predictors. The daily mean value of meteorological and air pollution factors, as well as the day of the week and seasonality, were included in the analysis. REML was used to select the optimal smoothing parameters to estimate the nonparametric smoothing functions. The time-lagged effects of the air pollutants on mortality were considered by fitting the models separately with single-day lags and the cumulative lags from zero to seven days and moving the average lags for up to seven days (Kim and Lee, 2019, Qiu et al. 2020). The Poisson GAMs can be described as follows:

\[
\log(\mu_t) = \beta_0 + s(\text{temp}_t) + s(\text{hum}_t) \\
+ (PM_{2.5}, \text{lagged terms}) \\
+ (PM_{10}, \text{lagged terms}) \\
+ (CO, \text{lagged terms}) \\
+ (SO_{2}, \text{lagged terms}) \\
+ (NO_{2}, \text{lagged terms}) \\
+ (O_{3}, \text{lagged terms}) \\
+ (\text{HPA}_t, \text{lagged terms}) \\
+ \text{factor(\text{DOW}_t)} + s(\text{time}_t)
\]

where \( \mu_t \) denotes the mean of the outcome, and \( s() \) denotes the nonparametric smoothing terms.

A further simplified predictive model was constructed using GLM by replacing the nonparametric smoothing terms in the Poisson GAM with the corresponding linear parametric terms. Similar to the ML-based models, the best model, whose cross-validated RMSE was the smallest, was selected among the various lag models.

Statistical analysis

Before model fitting, the data were split into a training set and a test set. The training set, which was used for model fitting, included the dataset from January 2015 to December 2018. The test set, which was used to compare the predicted mortality from the fitted model to the observed mortality, comprised the data from January 2019 to December 2019. Model selection was made using a 10-fold growing window and forward validation of the training set. The performance of the prediction model was evaluated using the test data and suggested by MAE and RMSE.

Statistical analyses were performed using R software (R version 4.0.3; The R Foundation for Statistical Computing, Vienna, Austria) and TensorFlow (TensorFlow: large-scale machine learning on heterogeneous systems, 2015). This study used the mgcv (Wood 2020) and caret packages in R for the GAM and GLM (Kuhn 2021), regularized linear regression, RF, and XGBoost, and the Tensorflow for LSTM and stacked LSTM.

Results

Study characteristics

Table 1 lists the mortality according to the cause of death and non-accidental causes, environmental exposures to meteorological conditions, and air pollutants. The number of deaths in 2015–2019 due to non-accidental causes, cardiovascular disease, and respiratory disease was 197,884 (daily mean 108.4), 43,696 (daily mean 23.9), and 21,182
(daily mean 11.6), respectively. During the study period, the daily mean concentrations of PM\textsubscript{2.5}, PM\textsubscript{10}, NO\textsubscript{2}, SO\textsubscript{2}, CO, and O\textsubscript{3} were 24.5 μg/m\textsuperscript{3}, 45.9 μg/m\textsuperscript{3}, 4.8 ppb, 35.3 ppb, 0.56 ppm, and 21.1 ppb, respectively. The mean temperature and humidity were 13.4 °C and 58.2 %, respectively.

Time-series plots illustrate the variation in the daily non-accidental mortality counts, daily mean values of temperature, and PM\textsubscript{2.5} (Fig. 1). The daily mean temperature and non-accidental mortality showed distinct seasonality, whereas PM\textsubscript{2.5} had less prominent seasonality, but it tended to be lower in summer.

**Comparison of the predictive models, including the daily measurements with various algorithms**

The ML-based model suggested better performance than GLM and GAM in predicting the non-accidental, cardiovascular, and respiratory mortality with the daily measurements of environmental exposure (Table 2). Elastic net performed best in predicting the non-accidental mortality, while XGBoost and Ridge regression were the best models for cardiovascular and respiratory mortality, respectively. The ML algorithms tended to predict the non-accidental mortality better than GLM and GAM, which showed better performance than the LSTM when the daily data of the environmental exposures were included in the model. In predicting the non-accidental mortality, the MAE and RMSE were improved by 4.7% and 4.0% in elastic net, respectively, compared to GAM. For cardiovascular mortality, XGBoost enhanced the performance by 4.9% in terms of the MAE and by 3.5% in terms of the RMSE relative to GAM. The relative improvement in the MAE and RMSE was 16.8% and 12.2% compared to GAM in the respiratory mortality when daily data was used.

**Comparison of the predictive models, including the hourly measurements with various algorithms**

Similarly, ML proposed better performance in predicting non-accidental, cardiovascular, and respiratory mortality with hourly measurements than GLM and GAM (Table 3). When the hourly measurements of environmental exposures were included, the most accurate models were those with lasso and elastic net regression; the MAE and the RMSE were decreased by 20.9% and 17.4%, respectively, compared to GAM. For cardiovascular mortality, the relative improvement in the MAE and RMSE was 16.8% and 12.2% compared to GAM in the respiratory mortality when daily data was used.
Visualization of the predictive models with various algorithms

Figure 2 shows the expected and observed non-accidental, cardiovascular, and respiratory mortality using the daily and hourly measurements. The blue and red lines represent the observed and predicted frequency of death in 2019 according to the causes of mortality when the best-performing algorithm was applied: elastic net for the model using the daily and hourly data in predicting the non-accidental mortality; XGBoost and RF for the model using the daily and hourly data, respectively, in predicting cardiovascular mortality; ridge regression for the model using the daily and hourly data in predicting respiratory mortality. These plots showed that the optimal models expressed the seasonal and nonlinear patterns. All the fitted models forecasted non-accidental, cardiovascular, and respiratory diseases (Supplementary Figs. 2, 3, and 4). When the $R^2$ between the expected and observed mortality was calculated, the highest value in the ML and GAM was 0.142 and 0.109, respectively, for non-accidental disease; 0.141 and 0.080, respectively, for cardiovascular disease; and 0.049 and 0.059, respectively, for respiratory disease.

Performance of the predictive models, including hourly measurements vs. daily measurements

Figure 3 compares the relative improvement in MAE and RMSE when hourly measurements were utilized instead of daily measurements. Although hourly measurements of environmental exposures did not enhance the performance in the GLM and GAM, they enhanced the LSTM, stacked LSTM, and RF in predicting non-accidental mortality. Similarly, for cardiovascular diseases, ridge, lasso, elastic net, and RF showed better performance when the hourly measurements were used, whereas the performance was worse in the GLM, GAM, and other ML-based models. In contrast, the hourly measurements did not enhance the performance in predicting respiratory mortality except for GAM. These were similar to the comparison of the RMSE between the models, including daily and hourly measurements.

Discussion

This study developed various predictive models using ML and DL algorithms, including hourly and daily measurements of environmental exposures, and compared their performance with that of the traditional models: the GLM and GAM. Although the algorithm showing the best performance was not uniform in the causes of death analyzed, the ML algorithm improved the performance in predicting cardiovascular, respiratory, and non-accidental diseases compared to the traditional method. On the other hand, hourly measurements of the environmental exposures failed to enhance the performance in some predictive models, whereas it enhanced the performance in other models. These findings could improve the understanding of using modeling specifications of ML algorithms and environmental exposure.

The predictive model developed with the ML algorithm outperformed the traditional models with GLM and GAM. Before the ML algorithm, GAM was commonly used in the forecast model of environmental epidemiology as the flexible and best model. GAM is an extension of the GLM and can allow for the nonparametric and nonlinear relationship of environmental exposures with the health outcomes and confounding effects of seasonality and trends (Hastie and Tibshirani, 1990, Sun et al. 2018). In the current study, ML-based models were superior to the GAM and GLM, whereas the best-performing algorithm was different depending on the outcomes; regularized linear regression had the smallest MAE and RMSE for respiratory and non-accidental disease, whereas the RF and XGBoost the smallest MAE and RMSE for cardiovascular disease. This result may be due to the different interactions between covariates in each model. In a meta-analysis, Bergmann et al. reported that the cold season was related to higher morbidity of cardiovascular disease than the warm season with an increase in SO$_2$ and PM$_{10}$, whereas there was no interaction between air pollutants and season in respiratory disease (Bergmann et al. 2020). In another meta-analysis, the study showed that PM$_{10}$ and O$_3$ were likely to have a higher risk of cardiovascular mortality at higher temperatures than non-accidental mortality (Li et al. 2017a). The changes in per 10 μg/m$^3$ of PM$_{10}$ were 1.28% (95%CI: 0.66–1.91) and 0.60% (95%CI: 0.30–0.90) for the cardiovascular and non-accidental mortality, respectively, and the changes per 10 μg/m$^3$ of O$_3$ were 1.63% (95%CI: 1.14–2.13) and 0.50% (95%CI: 0.30–0.60) for cardiovascular and non-accidental mortality, respectively. On the other hand, the ML algorithm, which had the best performance in the predictive models, was ephemeral. No specific ML algorithms consistently performed best in predicting cardiovascular, respiratory, and non-accidental mortality.

Bellinger et al. expected that DL could suggest the best prediction because it could reflect more complex relationships among the covariates than the other ML algorithms. They suggested that DL could be a solution to complex problems (Bellinger et al. 2017). Contrary to expectations, the DL-based models did not outperform the less complex models with other ML algorithms in this study (regularized linear regression, RF, and XGBoost). Previous research reported the successful development of a model with a good performance by adopting the LSTM (specialized recurrent...
(A) Atmospheric temperature

(B) PM2.5 distribution

(C) Daily Frequency of non-accidental death
neural network to learn patterns from time-series data) to predict the occurrence of COVID-19, hand-foot-mouth disease, cardiopulmonary disease, asthma attack, and cardiopulmonary hospitalization (Devaraj et al. 2021, Gu et al. 2019, Kim et al. 2020, Usmani et al. 2021, Wang et al. 2020). On the other hand, these studies did not compare the performance of DL-based models directly with other models with less complex ML algorithms. Therefore, they reached no clear conclusion and did not demonstrate that DL was superior to other ML-based models. Using LSTM, the $R^2$ ranged from 0.39 to 0.71 in predicting hand-foot-mouth disease based on meteorological conditions (Gu et al. 2019). Moreover, a well-trained LSTM model achieved optimal performance for predicting cardiopulmonary diseases with a 0.921% bias (Wang et al. 2020). Interestingly, regularized linear regression, which discouraged a more complex or flexible model and avoided the risk of overfitting by constraining the coefficient estimates toward zero, was superior to the LSTM and stacked LSTM. Therefore, the

Table 2 Comparison of the predictive models using machine learning algorithms with the generalized linear model and generalized additive model, including the daily measurements of air pollutants and meteorological factors for non-accidental, cardiovascular, and respiratory mortality

| Model               | MAE Compared to GAM | MAE Compared to GLM | RMSE Compared to GAM | RMSE Compared to GLM |
|---------------------|---------------------|---------------------|----------------------|----------------------|
| **Non-accidental mortality** |                     |                     |                      |                      |
| GLM                 | 9.113               | 11.781              | 9.131                | 11.652               |
| GAM                 | 9.131               | 11.652              | 9.131                | 11.652               |
| Ridge               | 8.809               | 11.299              | 8.809                | 11.299               |
| Lasso               | 8.740               | 11.257              | 8.740                | 11.257               |
| Elastic net         | **8.700**           | **11.182**          | **8.700**            | **11.182**           |
| Random Forest       | 9.212               | 11.769              | 9.212                | 11.769               |
| XGBoost             | 9.192               | 11.919              | 9.192                | 11.919               |
| LSTM                | 10.906              | 13.874              | 10.906               | 13.874               |
| Stacked LSTM        | 9.854               | 12.537              | 9.854                | 12.537               |
| **Cardiovascular mortality** |                     |                     |                      |                      |
| GLM                 | 4.327               | 5.396               | 4.327                | 5.396                |
| GAM                 | 4.329               | 5.422               | 4.329                | 5.422                |
| Ridge               | 4.242               | 5.301               | 4.242                | 5.301                |
| Lasso               | 4.133               | 5.240               | 4.133                | 5.240                |
| Elastic net         | 4.163               | 5.246               | 4.163                | 5.246                |
| Random Forest       | 4.181               | 5.267               | 4.181                | 5.267                |
| XGBoost             | **4.117**           | **5.231**           | **4.117**            | **5.231**            |
| LSTM                | 4.314               | 5.392               | 4.314                | 5.392                |
| Stacked LSTM        | 4.302               | 5.359               | 4.302                | 5.359                |
| **Respiratory mortality** |                     |                     |                      |                      |
| GLM                 | 3.314               | 4.031               | 3.314                | 4.031                |
| GAM                 | 3.346               | 4.060               | 3.346                | 4.060                |
| Ridge               | **2.785**           | **3.564**           | **2.785**            | **3.564**            |
| Lasso               | 2.800               | 3.588               | 2.800                | 3.588                |
| Elastic net         | 2.790               | 3.582               | 2.790                | 3.582                |
| Random Forest       | 2.886               | 3.573               | 2.886                | 3.573                |
| XGBoost             | 2.971               | 3.693               | 2.971                | 3.693                |
| LSTM                | 2.892               | 3.680               | 2.892                | 3.680                |
| Stacked LSTM        | 2.880               | 3.664               | 2.880                | 3.664                |

MAE, mean absolute error; RMSE, root mean square error; GLM, Generalized Linear Model; GAM, Generalized Additive Model; Ridge, logistic regression regularized with an L2 penalty; Lasso, logistic regression regularized with an L1 penalty; Elastic net, logistic regression regularized with an elastic net penalty; RF, Random Forest; XGBoost, extreme gradient boosting; LSTM, long short-term memory
The current study suggests that the best-performing algorithm should be identified and validated by comparing multiple ML algorithms when a new prediction model is constructed, or a new dataset is used. These results may be supported partially by previous research showing that classic algorithms performed as well as more advanced algorithms, even when disadvantaged by assuming linearity in the predictors (Christodoulou et al. 2019, Lynam et al. 2020).

The hourly measurement instead of daily data increased the performance in some models inconsistently. Hence, it might not be preferable to have more predictors in ML models: redundancy, overfitting, productivity, and understandability. In addition to low productivity and understandability, many predictors are discouraged because redundancy, which could disturb the performance, occurs even when many relevant predictors were included in the model. This can occur in a model that does not include feature selection in the algorithm. In this study, more predictors induced poorer performance in GLM and some GAM models. Furthermore, the error in the test data increased, while that in the training data decreased as the model became too complex. From the point of view of overfitting, the number of predictors contributes to model complexity, which could reduce the generalization and model performance. Experts suggested that a strong signal-to-noise ratio might be important in determining the success of an ML.

### Table 3 Comparison of the predictive models using machine learning algorithms with the generalized linear model and generalized additive model, including hourly measurement of air pollutants and meteorological factors for non-accidental, cardiovascular, and respiratory mortality

| Model                  | MAE  | Relative improvement Compared to GAM | RMSE  | Relative improvement Compared to GAM |
|------------------------|------|-------------------------------------|-------|--------------------------------------|
| **Non-accidental mortality** |      |                                     |       |                                      |
| GLM                    | 9.761|                                     | 12.545|                                      |
| GAM                    | 11.002|                                     | 13.558|                                      |
| Ridge                  | 8.708| 20.8%                               | 10.8% | 11.220                                | 17.2% | 10.6% |
| Lasso                  | 8.697| 20.9%                               | 10.9% | **11.192**                            | 17.4% | 10.8% |
| Elastic net            | 8.697| 20.9%                               | 10.9% | 11.210                                | 17.3% | 10.6% |
| Random Forest          | 9.058| 17.7%                               | 2.2%  | 11.660                                | 14.0% | 7.1%  |
| XGBoost                | 9.242| 16.0%                               | 5.3%  | 11.903                                | 12.2% | 5.1%  |
| LSTM                   | 9.441| 14.2%                               | 3.3%  | 12.109                                | 10.7% | 3.5%  |
| Stacked LSTM           | 9.550| 13.2%                               | 2.2%  | 12.408                                | 8.5%  | 1.1%  |
| **Cardiovascular mortality** |      |                                     |       |                                      |
| GLM                    | 4.529|                                     | 5.655 |                                      |
| GAM                    | 4.647|                                     | 5.797 |                                      |
| Ridge                  | 4.139| 10.9%                               | 8.6%  | 5.208                                 | 10.2% | 7.9%  |
| Lasso                  | 4.114| 11.5%                               | 9.2%  | 5.207                                 | 10.2% | 7.9%  |
| Elastic net            | 4.134| 11.0%                               | 8.7%  | 5.214                                 | 10.1% | 7.8%  |
| Random Forest          | **4.078**| 12.3%                               | 10.0% | **5.173**                            | 10.8% | 8.5%  |
| XGBoost                | 4.183| 10.0%                               | 7.6%  | 5.349                                 | 7.7%  | 5.4%  |
| LSTM                   | 4.372| 5.9%                                | 3.5%  | 5.460                                 | 5.8%  | 3.4%  |
| Stacked LSTM           | 4.316| 7.1%                                | 4.7%  | 5.402                                 | 6.8%  | 4.5%  |
| **Respiratory mortality** |      |                                     |       |                                      |
| GLM                    | 3.522|                                     | 4.357 |                                      |
| GAM                    | 3.076|                                     | 3.834 |                                      |
| Ridge                  | **2.803**| 8.9%                               | 20.4% | **3.609**                            | 5.9%  | 17.2% |
| Lasso                  | 2.813| 8.5%                                | 20.1% | 3.594                                 | 6.3%  | 17.5% |
| Elastic net            | 2.806| 8.8%                                | 20.3% | 3.596                                 | 6.2%  | 17.5% |
| Random Forest          | 2.991| 2.8%                                | 15.1% | 3.755                                 | 2.1%  | 13.8% |
| XGBoost                | 3.014| 2.0%                                | 14.4% | 3.820                                 | 0.4%  | 12.3% |
| LSTM                   | 2.912| 5.4%                                | 17.3% | 3.703                                 | 3.4%  | 15.0% |
| Stacked LSTM           | 2.915| 5.2%                                | 17.2% | 3.701                                 | 3.5%  | 15.1% |

*MAE*, mean absolute error; *RMSE*, root mean square error; *GLM*, Generalized Linear Model; *GAM*, Generalized Additive Model; *Ridge*, logistic regression regularized with an L2 penalty; *Lasso*, logistic regression regularized with an L1 penalty; *Elastic net*, logistic regression regularized with an elastic net penalty; *RF*, Random Forest; *XGBoost*, extreme gradient boosting; *LSTM*, long short-term memory.
Fig. 2 Plots of the predicted and observed cardiovascular, respiratory, and non-accidental mortality using the daily and hourly measurement of air pollutants and meteorological factors. The blue and red lines denote the observed and predicted frequency of deaths in 2019 when the best-performing algorithm was applied: elastic net (non-accidental mortality), XGBoost, and RF, respectively, for the daily and hourly data (cardiovascular mortality), and ridge regression (respiratory mortality).
algorithm (Christodoulou et al. 2019). In contrast, clinical prediction problems often have poor signal-to-noise ratios. Therefore, feature selection and feature engineering may help improve the model performance of ML-based models (Christodoulou et al. 2019).

This study had some limitations. Only meteorological and air pollution factors were considered, and socio-economic factors and individual-level health risks, such as underlying disease, lifestyle, and medication, were not included. This may lead to inaccurate predictions because meteorological...
factors were not strong determinants of mortality (Gu et al. 2019). Taking other potential predictors into the model development would improve the prediction performance. Second, only the effects of short-term exposure to environmental factors on adverse health outcomes were considered despite their long-term exposure (Amini et al. 2020).

Furthermore, measurement errors occurred in estimating the environmental exposures because air pollutant concentration and meteorological factors were collected from the fixed-site stations. These may not represent the actual exposure at an individual level. The above three points are limitations in all the models regardless of the algorithms. Lastly, the ML algorithm was less likely to be explainable and intuitive than the GLM and GAM models. Therefore, this study could not define how much each predictor and interaction contributed to the adverse health outcomes. Recently, explainable artificial intelligence has been developing, and the ML and DL algorithm may be applied in a clear box.

Conclusion

This study proposed a method for predicting mortality from cardiovascular, respiratory, and non-accidental diseases using a machine learning algorithm. These methods improved the performance compared to the traditional method. On the other hand, hourly measurements of environmental exposures showed inconsistent results in enhancing the performance of the models, which depended on the analytic methods and health outcomes. These findings add information for future epidemiological and ML-based modeling considering health impact predictions, such as hospital admission, healthcare utilization, and prevention measures.

Abbreviations PM: particulate matter; PM$_{2.5}$: particles smaller than 2.5 μm; PM$_{10}$: particles smaller than 10 μm; NO$_2$: nitrogen dioxide; SO$_2$: sulfur dioxide; O$_3$: ozone; ML: machine learning; RF: Random Forest; XGBoost: extreme gradient boosting; Lasso: logistic regression regularized with an L1 penalty; Ridge: logistic regression regularized with an L2 penalty; Elastic net: logistic regression regularized with an elastic net penalty; LSTM: long short-term memory; GLM: generalized linear model; GAM: generalized additive model; MAE: mean absolute error; RMSE: root mean square error

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-022-21768-9.

Author contribution All authors contributed to the study’s conception and design. Material preparation and data extraction were performed by Won Kyung Lee, and data analysis and visualization were performed by Woosoo Lee and Yoonjin Kim. Won Kyung Lee wrote the first draft of the manuscript, and the previous versions of the manuscript were supervised and commented on by Youn-Hee Lim and Eunhee Ha. All authors read and approved the final manuscript.

Funding The study was supported by the Bio & Medical Technology Development Program of the National Research Foundation funded by the Korean government (MSIT) (2019M3E5D1A0206962012). The funders had no role in study design, data collection, analysis, decision to publish, or manuscript preparation.

Data Availability The data was obtained from the Death Statistics Database of the Korean National Statistical Office, and air pollutants were provided by the Korean National Institute of Environmental Research. These datasets are available to the public.

Declarations

Ethics approval This research was exempted from IRB review (identification number: 2021-06-006) by the Inha University Hospital Institutional Review Board because it used de-identified data accessible to the public.

Consent to participate The consent to participate was not necessary because the research is using secondary datasets.

Consent for publication The work described was original research that has not been published previously and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Competing interests The authors declare no competing interests.

References

Al Ahad MA, Sullivan F, Demšar U, Melhem M, Kulu H (2020) The effect of air-pollution and weather exposure on mortality and hospital admission and implications for further research: a systematic scoping review. PLoS One 15:1–24

Amini H et al. (2020): Long-term exposure to air pollution and stroke incidence: a Danish nurse cohort study. Environ Int 142

Anenberg SC, Haines S, Wang E, Nassikas N, Kinney PL (2020) Synergistic health effects of air pollution, temperature, and pollen exposure: a systematic review of epidemiological evidence. Environ Health Global Access Sci Source 19

Bellinger C, Mohomed Jabbar MS, Zaione O, Osornio-Vargas A (2017) A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health 17:1–19

Benke K, Benke G (2018): Artificial intelligence and big data in public health. Int J Environ Res Pub Health 15

Bergmann S, Li B, Pilot E, Chen R, Wang B, Yang J (2020) Effect modification of the short-term effects of air pollution on morbidity by season: a systematic review and meta-analysis. Sci Total Environ 716:136985–136985

Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 110:12–22

Devaraj J, Madurai Elavarasan R, Pugazhendhi R, Shafullah GM, Ganesan S, Jeyssree AK, Khan IA, Hossain E (2021) Forecasting of COVID-19 cases using deep learning models: is it reliable and practically significant? Results Phys 21:103817–103817

Gu J, Liang L, Song H, Kong Y, Ma R, Hou Y, Zhao J, Liu J, He N, Zhang Y (2019) A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China. Sci Rep 9:17928–17928
Guo Y et al (2016) Temperature variability and mortality: a multi-country study. Environ Health Perspect 124:1554–1559

Hastie TJ, Tibshirani RJ (1990) Generalized additive models, 43. CRC press

Kim D, Cho S, Tamil L, Song DJ, Seo S (2020) Predicting asthma attacks: effects of indoor PM concentrations on peak expiratory flow rates of asthmatic children. IEEE Access 8:8791–8797

Kim H, Lee JT (2019) On inferences about lag effects using lag models in air pollution time-series studies. Environ Res 171:134–144

Kuhn M (2021) caret: classification and regression training

Lee W, Choi HM, Kim D, Honda Y, Leon Guo YL, Kim H (2019) Synergic effect between high temperature and air pollution on mortality in Northeast Asia. Environ Res 178:108735–108735

Lee WH, Lim YH, Dang TN, Seposo X, Honda Y, Guo YLL, Jang HM, Kim H (2017) An investigation on attributes of ambient temperature and diurnal temperature range on mortality in five East-Asian countries. Sci Rep 7:1–9

Li J, Woodward A, Hou XY, Zhu T, Zhang J, Brown H, Yang J, Qin R, Gao J, Gu S, Li J, Xu L, Liu X, Liu Q (2017a) Modification of the effects of air pollutants on mortality by temperature: a systematic review and meta-analysis. Sci Total Environ 575:1556–1570

Li X, Peng L, Yao X, Cui S, Hu Y, You C, Chi T (2017b) Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ Pollut 231:997–1004

Lin Z, Lin S, Neamtiu IA, Ye B, Csobod E, Fazakas E, Gurzau E (2021) Predicting environmental risk factors in relation to health outcomes among school children from Romania using random forest model – an analysis of data from the SINPHONIE project. Sci Total Environ 784:147145

Lynam AL, Dennis JM, Owen KR, Oram RA, Jones AG, Shields BM, Ferrat LA (2020) Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. Diag Prog Res 4:0–9

Qiu H, Luo L, Su Z, Zhou L, Wang L, Chen Y (2020) Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. BMC Med Inf Decision Making 20:1–11

Rodopoulou S, Samoli E, Analitis A, Atkinson RW, de Donato FK, Katsouyanni K (2015) Searching for the best modeling specification for assessing the effects of temperature and humidity on health: a time series analysis in three European cities. Int J Biometeorol 59:1585–1596

Schnaibelt M (2019): A comparison of machine learning model validation schemes for non-stationary time series data. FAU Discussion Papers in Economics, Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics

Sun JM, Lu L, Liu KK, Yang J, WuHX, Liu QY (2018) Forecast of severe fever with thrombocytopenia syndrome incidence with meteorological factors. Sci Total Environ 626:1188–1192

Tian L, Liang F, Guo Q, Chen S, Xiao S, Wu Z, Jin X, Pan X (2018) The effects of interaction between particulate matter and temperature on mortality in Beijing, China. Environ Sci Process Impacts 20:395–405

Usmani RSA, Pillai TR, Hashem IAT, Marjani M, Shaharudin R, Latif MT (2021) Air pollution and cardiorespiratory hospitalization, predictive modeling, and analysis using artificial intelligence techniques. Environ Sci Pollut Res Int 28:56759–56771

Wang C, Qi Y, Zhu G (2020): Deep learning for predicting the occurrence of cardiopulmonary diseases in Nanjing, China. Chemosphere 257

Wood S (2020) MGCV: mixed GAM computation vehicle with automatic smoothness estimation

Zhang Q, Fu F, Tian R (2020) A deep learning and image-based model for air quality estimation. Sci Total Environ 724:138178–138178

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.