Comparative Analysis of Predictive Models for Fine Particulate Matter in Daejeon, South Korea

Tserenpurev Chuluunsaihan 1,2, Menghok Heak 2, Aziz Nasridinov 1,* and Sanghyun Choi 3,3,*

1 Department of Computer Science, Chungbuk National University, Cheongju 28644, Korea; teo@chungbuk.ac.kr
2 Department of Management Information Systems, Chungbuk National University, Cheongju 28644, Korea; menghok@chungbuk.ac.kr
3 Department of Bigdata, Chungbuk National University, Cheongju 28644, Korea
* Correspondence: aziz@chungbuk.ac.kr (A.N.); chois@chungbuk.ac.kr (S.C.)
† Co-first authors, these authors contributed equally to this work.

Abstract: Air pollution is a critical problem that is of major concern worldwide. South Korea is one of the countries most affected by air pollution. Rapid urbanization and industrialization in South Korea have induced air pollution in multiple forms, such as smoke from factories and exhaust from vehicles. In this paper, we perform a comparative analysis of predictive models for fine particulate matter in Daejeon, the fifth largest city in South Korea. This study is conducted for three purposes. The first purpose is to determine the factors that may cause air pollution. Two main factors are considered: meteorological and traffic. The second purpose is to find an optimal predictive model for air pollutant concentration. We apply machine learning and deep learning models to the collected dataset to predict hourly air pollutant concentrations. The accuracy of the deep learning models is better than that of the machine learning models. The third purpose is to analyze the influence of road conditions on predicting air pollutant concentration. Experimental results demonstrate that considering wind direction and wind speed could significantly decrease the error rate of the predictive models.

Keywords: air pollution; deep learning; fine particulate matter; machine learning; predictive models

1. Introduction

Air pollution is a major issue in numerous countries worldwide because it causes harmful diseases, including physical and mental illnesses [1-3]. A World Health Organization report states that air pollution causes approximately 1/8 of premature deaths annually, which is estimated to be 6.5 million people [4]. Industrial emissions, vehicle engine emissions, and meteorological factors are considered to be the root causes of air pollution [5]. The air quality index (AQI) represents the pollution caused by six primary air pollutants: particulate matter (PM), ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), and sulfur dioxide (SO2). Among these, fine PM is a major air pollutant. PM10 refers to PM with a diameter of 10 μm or less, and PM2.5 refers to PM with a diameter of less than 2.5 μm. PM includes the waste generated by combustion engines, solid fuel, energy production, and other activities.

According to an air quality map obtained using a NASA satellite, South Korea is severely affected by air pollution [6-8]. The transportation system in South Korea has grown significantly because of rapid urbanization and industrialization. Even though South Korea has one of the world’s most modern transportation systems, most people still use personal vehicles. With a population of approximately 51 million, South Korea had approximately 23 million on-road motor vehicles registered as of 2018 [9]. The large

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Citation: Chuluunsaihan, T.; Heak, M.; Nasridinov, A.; Choi, S. Comparative Analysis of Predictive Models for Fine Particulate Matter in Daejeon, South Korea. Atmosphere 2021, 12, 1295. https://doi.org/10.3390/atmos12101295

Received: 26 July 2021
Accepted: 30 September 2021
Published: 5 October 2021

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
number of personal vehicles has caused the problems of traffic jams and pollution due to vehicle emissions [10]. Air pollutants not only directly affect human health but also have long-term effects on atmospheric air quality. Daejeon is the fifth-largest metropolitan city in the country, with a population of 1.45 million as of 2020 [11]. Air pollution is prevalent in Daejeon [12-14]. For example, according to the data for one month between February 10 and March 11, 2021, the AQI based on PM$_{2.5}$ was good, moderate, and unhealthy for 7, 19, and 4 days, respectively.

Several authors have proposed machine learning-based and deep learning-based models for predicting the AQI using meteorological data in South Korea. For example, Jeong et al. [15] used a well-known machine learning model, Random Forest (RF), to predict PM$_{10}$ concentration using meteorological data, such as air temperature, relative humidity, and wind speed. A similar study was conducted by Park et al. [16], who predicted PM$_{10}$ and PM$_{2.5}$ concentrations in Seoul using several deep learning models. Numerous researchers have proposed approaches for determining the relationship between air quality and traffic in South Korea. For example, Kim et al. [17] and Eum [18] proposed approaches to predict air pollution using various geographic variables, such as traffic and land use. Jang et al. [19] predicted air pollution concentration in four different sites (traffic, urban background, commercial, and rural background) of Busan using a combination of meteorological and traffic data.

This paper proposes a comparative analysis of the predictive models for PM$_{2.5}$ and PM$_{10}$ concentrations in Daejeon. This study has three objectives. The first is to determine the factors (i.e., meteorological or traffic) that affect air quality in Daejeon. The second is to find an accurate predictive model for air quality. Specifically, we apply machine learning and deep learning models to predict hourly PM$_{2.5}$ and PM$_{10}$ concentrations. The third is to analyze whether road conditions influence the prediction of PM$_{2.5}$ and PM$_{10}$ concentrations. More specifically, the contributions of this study are as follows:

- First, we collected meteorological data from 11 air pollution measurement stations and traffic data from eight roads in Daejeon from January 1, 2018, to December 31, 2018. Then, we preprocessed the datasets to obtain a final dataset for our prediction models. The preprocessing consisted of the following steps: (1) consolidating the datasets, (2) cleaning invalid data, and (3) filling in missing data.
- Furthermore, we evaluated the performance of several machine learning and deep learning models for predicting the PM concentration. We selected the RF, gradient boosting (GB), and light gradient boosting (LGBM) machine learning models. In addition, we selected the gated recurrent unit (GRU) and long short-term memory (LSTM) deep learning models. We determined the optimal accuracy of each model by selecting the best parameters using a cross-validation technique. Experimental evaluations showed that the deep learning models outperformed the machine learning models in predicting PM concentrations in Daejeon.
- Finally, we measured the influence of the road conditions on the prediction of PM concentrations. Specifically, we developed a method that set road weights on the basis of the stations, road locations, wind direction, and wind speed. An air pollution measurement station surrounded by eight roads was selected for this purpose. Experimental results demonstrated that the proposed method of using road weights decreased the error rates of the predictive models by up to 21% and 33% for PM$_{10}$ and PM$_{2.5}$, respectively.

The rest of this paper is organized as follows: Section 2 discusses related studies on the prediction of PM concentrations on the basis of meteorological and traffic data. Section 3 describes the materials and methods used in this study. Section 4 presents the results of performance evaluation. Section 5 summarizes and concludes the study.
2. Related Work

Various studies have been conducted on the harmful effects of air pollution. We classify these studies into the following three categories: 1) studies that use only meteorological data, 2) studies that use only traffic data, and 3) studies that use meteorological and traffic data. The subsequent sections discuss each category in detail.

2.1. Prediction of AQI Using Meteorological Data

Several authors have proposed machine learning-based and deep learning-based methods for predicting the AQI using meteorological data [16, 20-24]. For example, Park et al. [16] predicted PM2.5 concentrations on the basis of meteorological features, including temperature, humidity, wind direction, and wind speed. A dataset was collected from two areas in Seoul, South Korea. The study proposed the LSTM and artificial neural network (ANN) models to predict PM concentrations after a certain time. The authors proposed an algorithm that selected the LSTM or ANN model on an hourly basis. The accuracy of the proposed model was higher than that of the LSTM and ANN models. Lee et al. [20] predicted PM2.5 concentrations in Taiwan using the GB model. They used a dataset consisting of hourly measurements obtained over one year from 77 air monitoring stations and 580 meteorological stations in Taiwan. Experimental results indicated that the model provided accurate 24-h predictions at most air stations. Chang et al. [21] used the RF model to predict PM2.5 concentrations on the basis of meteorological features such as wind direction, wind speed, temperature, humidity, and rainfall. The authors compared the proposed model with two other time-series data analysis models: logistic regression and linear discriminant. Experimental results demonstrated that the RF model was the most accurate for predicting PM2.5 concentrations. Choubin et al. [22] assessed the spatial hazard of PM10 concentrations using three machine learning models: RF, bagged cart, and mixture discriminant analysis. The study area was selected from Barcelona, which is an urban and industrial area in Western Europe. The authors assembled a dataset that included PM concentrations (PM10, PM2.5, PM1, and others) and meteorological features (wind speed, wind direction, etc.). In addition, the features that affected PM modeling were identified by a feature selection approach referred to as simulated annealing. Experimental results demonstrated that the accuracies of all three machine learning models were higher than 87% for predicting PM10 concentrations.

A few studies have used deep learning approaches to predict the AQI. For example, Qadeer et al. [23] predicted hourly PM2.5 concentrations in two large South Korean cities (Seoul and Gwangju), along with various pollutants and meteorological features. The pollutant features consisted of PM2.5, PM10, SO2, O3, NO2, and CO concentrations. The meteorological features consisted of temperature, wind speed, relative humidity, surface roughness, planetary boundary layer, and precipitation. Experimental results showed that the LSTM model outperformed the XGBoost, LGBM, recurrent neural network (RNN), and convolutional neural network models in predicting hourly PM2.5 concentrations. Xayasouk et al. [24] applied the LSTM and deep autoencoder (DAE) models to predict hourly PM2.5 and PM10 concentrations in Seoul, South Korea. The authors used the AQI data for 2015–2018 and various meteorological features, such as humidity, rain, wind speed, wind direction, temperature, and atmospheric conditions. Experimental results showed that the performance of the LSTM model was slightly better than that of the DAE model in terms of the root mean square error (RMSE).

2.2. Prediction of AQI Using Traffic Data

Numerous researchers have proposed approaches for determining the relationship between air quality and traffic [25-27]. For example, Comert et al. [25] studied the impact of traffic volume on air quality in South Carolina, United States. They predicted O3 and PM2.5 concentrations on the basis of the annual average daily traffic (AADT) by obtaining historical traffic volume and air quality data between 2006 and 2016 from monitoring
stations. Experimental results showed that air quality worsened when the AADT increased. Adams et al. [26] examined the PM$_{2.5}$ concentration caused by vehicles in schools, particularly in the morning when parents dropped their children off. A dataset was obtained from a study of 23–116 personal vehicles at 25 schools, which had 160–765 students. The dataset was fit to predict the PM$_{2.5}$ concentration using a linear regression model. The PM$_{2.5}$ concentration was 10–50 μg/m$^3$ in the morning at the drop-off locations. This study concluded that the use of private vehicles could significantly deteriorate air quality. Askariyeh et al. [27] studied PM$_{2.5}$ concentrations on the basis of traffic on highways and arterial roads. Near-road PM$_{2.5}$ concentrations depended on the road type, vehicle weight, traffic volume, and other features. A dataset was collected from a hotspot in Dallas, Texas, by the U.S. Environmental Protection Agency (EPA). The authors proposed a traffic-related PM$_{2.5}$ concentration model using emission modeling based on MOVES Vehicle Emission Simulator (MOVES) and dispersion modeling based on the American Meteorological Society/Environmental Protection Agency Regulatory Model (AERMOD). The MOVES model required traffic-related variables, including exhaust, brake, and tire wear. AERMOD required emissions and meteorological features. Experimental results revealed that emission and dispersion modeling increased the prediction accuracy of near-road PM$_{2.5}$ concentrations by up to 74%.

2.3. Prediction of AQI Using Meteorological and Traffic Data

Studies have used a combination of meteorological and traffic data [28-32] to improve the accuracy of AQI prediction models. For example, Rossi et al. [28] studied the effect of road traffic flows on air pollution. The dataset of the study was collected in Padova, Italy, during the COVID-19 lockdown. The authors analyzed pollutant concentrations (NO, NO$_2$, NO$_x$, and PM$_{10}$) with vehicle counts and meteorology. Statistical tests, correlation analyses, and multivariate linear regression models were applied to investigate the effect of traffic on air pollution. Experimental results indicated that PM$_{10}$ concentrations were not primarily affected by local traffic. However, vehicle flows significantly affected NO, NO$_2$, and NO$_x$ concentrations. Lešnik et al. [29] performed a predictive analysis of PM$_{10}$ concentrations using meteorological and detailed traffic data. They used a dataset consisting of wind direction, atmospheric pressure, wind speed, rainfall, ambient temperature, relative humidity, vehicle speed, and traffic volume. They proposed a genetic algorithm to perform multiple regression analysis. Experimental results showed that the proposed genetic algorithm was more accurate than the present state-of-the-art algorithms. Wei et al. [30] proposed a framework to explore the relationship between roadside PM$_{2.5}$ concentrations and traffic volume. They collected three types of data, i.e., meteorological, traffic volume, and PM$_{2.5}$ concentrations, from Beijing, China. Their framework utilized data characteristics using a wavelet transform, which divided the data into different frequency components. The framework demonstrated two microscale rules: 1) the characteristic period of PM$_{2.5}$ concentrations; 2) the delay of 0.3–0.9 min between PM$_{2.5}$ concentrations and traffic volume. Catalano et al. [31] predicted peak air pollution episodes using an ANN. The study area was Marylebone Road in London, which consists of three lanes on each side. The dataset used in the study contained traffic volume, meteorological conditions, and air quality data obtained over ten years (1998–2007). The authors compared the ANN and autoregressive integrated moving average with an exogenous variable (ARIMAX) in terms of the mean absolute percent error. Experimental results showed that the ANN produced 2% fewer errors compared to the ARIMAX model. Askariyeh et al. [32] predicted near-road PM$_{2.5}$ concentrations using wind speed and wind direction. The EPA has installed monitors in near-road environments in Houston, Texas. The monitors collect PM$_{2.5}$ concentrations and meteorological data. The authors created a multiple linear regression model to predict 24-h PM$_{2.5}$ concentrations. The results indicated that wind speed and wind direction affected near-road PM$_{2.5}$ concentrations.
3. Materials and Methods

3.1. Overview

Figure 1 shows the overall flow of the proposed method. It consists of the following steps: data acquisition, data preprocessing, model training, and evaluation. Our main objective is to predict PM$_{10}$ and PM$_{2.5}$ concentrations on the basis of meteorological and traffic features using machine learning and deep learning models. First, we collected data from various governmental online resources via web crawling. Then, we integrated the collected data into a raw dataset and preprocessed it using several data-cleaning techniques. Finally, we applied machine learning and deep learning models to predict PM$_{10}$ and PM$_{2.5}$ concentrations and analyzed the prediction results. We have described each step in detail in the following subsections.

![Figure 1. Overall flow of the proposed method.](image)

3.2. Study Area

The study area was Daejeon, which is located in the central area of the Korean Peninsula. Daejeon experiences severe air pollution owing to the high usage of personal vehicles and proximity of power plants. There are 11 air pollution measurement stations in the five districts of Daejeon, as shown in Figure 2(a). These stations measure the city’s AQI for six different pollutants (PM$_{2.5}$, PM$_{10}$, O$_3$, NO$_2$, CO, and SO$_2$) every hour. We selected eight roads for our study on the basis of traffic congestion, i.e., Gyeryong-ro, Daedeok-daero, Dunsan-daero, Munye-ro, Mun-jeong-ro, Wolpyeong-ro, Cheongsaseo-ro, and Hanbat-daero, as shown in Figure 2(b).
3.3. Data Collection

All datasets used in this study were retrieved from South Korea’s open government data portals. Air quality data were obtained from AirKorea [33], which is operated by the Korean Ministry of Environment and the Korea Environment Corporation, and meteorological data were obtained from the Korea Meteorological Administration [34]. Traffic data were collected from the Daejeon Transportation Data Warehouse system [35], which provides road traffic information such as travel speed and traffic volume. The data were collected using web crawling techniques, which access web pages over the HTTP protocol to retrieve and extract data in the HTML or JSON format.

We collected hourly time-series data between January 1, 2018, and December 31, 2018. We concatenated the collected datasets into one dataset on the basis of the DateTime index. The final dataset consisted of 8,760 observations. Figure 3 shows the distribution of the AQI by the (a) DateTime index, (b) month, and (c) hour. The AQI is relatively better from July to September compared to the other months. There are no major differences between the hourly distribution of the AQI. However, the AQI worsens from 10 a.m. to 1 p.m.
3.4. Competing Models

Several models were used to predict air pollutant concentrations in Daejeon. Specifically, we fitted the data using ensemble machine learning models (RF, GB, and LGBM) and deep learning models (GRU and LSTM). This subsection provides a detailed description of these models and their mathematical foundations.

The RF [36], GB [37], and LGBM [38] models are ensemble machine learning algorithms, which are widely used for classification and regression tasks. The RF and GB models use a combination of single decision tree models to create an ensemble model. The main differences between the RF and GB models are in the manner in which they create and train a set of decision trees. The RF model creates each tree independently and combines the results at the end of the process, whereas the GB model creates one tree at a time and combines the results during the process. The RF model uses the bagging technique, which is expressed by Equation (1). Here, \( N \) represents the number of training subsets, \( h_t(x) \) represents a single prediction model with \( t \) training subsets, and \( H(x) \) is the final ensemble model that predicts values on the basis of the mean of \( n \) single prediction models. The GB model uses the boosting technique, which is expressed by Equation (2). Here, \( M \) and \( m \) represent the total number of iterations and the iteration number, respectively. \( H_m(x) \) is the final model at each iteration. \( y_m \) represents the weights calculated on the basis of errors. Therefore, the calculated weights are added to the next model (\( h_m(x) \)).

\[
H(x) = \{h_t(x), t = 1, \ldots, N\} \tag{1}
\]

\[
H_m(x) = \sum_{m=1}^{M} y_m h_m(x) \tag{2}
\]

The LGBM model extends the GB model with the automatic feature selection. Specifically, it reduces the number of features by identifying the features that can be merged. This increases the speed of the model without decreasing accuracy.

An RNN is a deep learning model for analyzing sequential data such as text, audio, video, and time series. However, RNNs have a limitation referred to as the short-term memory problem. An RNN predicts the current value by looping past information. This is the main reason for the decrease in the accuracy of the RNN when there is a large gap between past information and the current value. The GRU [39] and LSTM [40] models overcome the limitation of RNNs by utilizing additional gates to pass information in long sequences. The GRU cell uses two gates: an update gate and a reset gate. The update gate determines whether to update a cell. The reset gate determines whether the previous cell state is important. The LSTM cell uses three gates: an insert gate, a forget gate, and an output gate. The insert gate is the same as the update gate of the GRU model. The forget gate removes the information that is no longer required. The output gate returns the output to the next cell states. The GRU and LSTM models are expressed by Equations (3) and (4), respectively.

The following notations are used in these equations:

- \( t \): Time steps.
- \( \tilde{c}^t, c^t \): Candidate cell and final cell state at time step \( t \). The candidate cell state is also referred to as the hidden state.
- \( W \): Weight matrix.
- \( b \): Bias vectors.
- \( u, r, i, f, o \): Update gate, reset gate, insert gate, forget gate, and output gate, respectively.
- \( a^t \): Activation functions.
\[ \hat{C}^t = \tanh(W_c[a^{t-1}, X^t] + b_c) \]
\[ u^t = \sigma(W_u[C^{t-1}, X^t] + b_u) \]
\[ r^t = \sigma(W_r[C^{t-1}, X^t] + b_r) \]
\[ C^t = u^t \cdot \hat{C}^t + (1 - u^t) \cdot C^{t-1} \]
\[ a^t = c^t \]
\[ \hat{C}^t = \tanh(W_c[a^{t-1}, X^t] + b_c) \]
\[ i^t = \sigma(W_i[a^{t-1}, X^t] + b_i) \]
\[ f^t = \sigma(W_f[a^{t-1}, X^t] + b_f) \]
\[ o^t = \sigma(W_o[a^{t-1}, X^t] + b_o) \]
\[ C^t = u^t \cdot \hat{C}^t + f^t \cdot c^{t-1} \]
\[ a^t = o^t \cdot (C^t) \]

3.5. Evaluation Metrics

The models are evaluated to study their prediction accuracy and determine which model should be used. Three of the most frequently used parameters for evaluating models are the coefficient of determination (\(R^2\)), RMSE, and mean absolute error (MAE). The RMSE measures the square root of the average of the squared distance between actual and predicted values. As errors are squared before calculating the average, the RMSE increases exponentially if the variance of errors is large.

The \(R^2\), RMSE, and MAE are expressed by Equations (5), (6), and (7), respectively. Here, \(N\) represents the number of samples, \(y\) represents an actual value, \(\hat{y}\) represents a predicted value, and \(\bar{y}\) represents the mean of observations. The main metric is the distance between \(y\) and \(\hat{y}\), i.e., the error or residual. The accuracy of a model is considered to improve as these two values become closer.

\[ R^2 = 100 \times (1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}) \]  
\[ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \]  
\[ \text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \]

4. Results

4.1. Preprocessing

The datasets used in this study consisted of hourly air quality, meteorology, and traffic data observations. The blank cells in the datasets represented a value of zero for wind direction and snow depth. When the cells for wind direction were blank, the wind was not notable (the wind speed was zero or almost zero). Furthermore, the cells for snow depth were blank on non-snow days. Hence, they were replaced by zero. The seasonal factor was extracted from the DateTime column of the datasets. A new column, i.e., month, was used to represent the month in which an observation was obtained. The column consisted of 12 values (Jan–Dec). The wind direction column was converted from the numerical value in degrees (0°–360°) into five categorical values. The wind direction
at 0° was labeled N/A, indicating that no critical wind was detected. The wind direction from 1°–90° was labeled as northeast (NE), 91°–180° as southeast (SE), 181°–270° as southwest (SW), and 271° or more as northwest (NW). The average traffic speed was calculated and binned. The binning size was set as 10 (unit: km/h) because the minimum average speed was approximately 25 and the maximum was approximately 60. Subsequently, the binned values were divided into four groups. The average speeds in the first, second, third, and fourth groups were 25–35 km/h, 36–45 km/h, 46–55 km/h, and more than 55 km/h, respectively.

The datasets were combined into one dataset, as shown in Table 1. A few observations in this dataset were missing or invalid. Missing values were treated as types of data errors, in which the values of observations could not be found. The occurrence of missing data in a dataset can cause errors or failure in the model-building process. Thus, in the preprocessing stage, we replaced the missing values with logically estimated values. The following three techniques were considered for filling the missing values:

Table 1. Description of integrated dataset.

| Variable Name | Count | Mean       | Min | Max          | Std   | Missing Value |
|---------------|-------|------------|-----|--------------|-------|---------------|
| PM₂.₅        | 8342  | 20.185447  | 2   | 145          | 15.808386 | 418            |
| PM₁₀         | 8760  | 35.118607  | 0   | 296          | 23.372221 | 0              |
| TEMPERATURE  | 8756  | 13.593     | -16 | 39.3         | 11.593 | 4              |
| WIND_SPEED   | 8760  | 1.552      | 0   | 8.3          | 1.16   | 0              |
| WIND_DIRECTION | 8760   | 201.705   | 0   | 360          | 124.023 | 0              |
| HUMIDITY     | 8746  | 68.954     | 14  | 98           | 19.777 | 14             |
| AIR_PRESSURE | 8760  | 1008.918   | 979.6| 1030.7       | 8.129  | 0              |
| SNOW_DEPTH   | 270   | 3.088      | 0   | 7.9          | 2.015  | 8490           |
| ROAD_1       | 8328  | 38.275     | 0   | 58.489       | 9.614  | 432 (N/A) + 501 (Zero) |
| ROAD_2       | 8328  | 52.994     | 0   | 75.691       | 10.1   | 432 (N/A) + 501 (Zero) |
| ROAD_3       | 8328  | 39.371     | 0   | 62.828       | 11.078 | 432 (N/A) + 501 (Zero) |
| ROAD_4       | 8328  | 43.682     | 0   | 64.895       | 10.66  | 432 (N/A) + 501 (Zero) |
| ROAD_5       | 8328  | 41.353     | 0   | 68.33        | 12.375 | 432 (N/A) + 501 (Zero) |
| ROAD_6       | 8328  | 41.063     | 0   | 53.382       | 6.332  | 432 (N/A) + 501 (Zero) |
| ROAD_7       | 8328  | 36.027     | 0   | 61.022       | 11.231 | 432 (N/A) + 501 (Zero) |
| ROAD_8       | 8328  | 42.825     | 0   | 65.912       | 11.786 | 432 (N/A) + 501 (Zero) |

- Last observation carried forward (LOCF): The last observed non-missing value was used to fill the missing values at later points.
- Next observation carried backward (NOCB): The next non-missing observation was used to fill the missing values at earlier points.
- Interpolation: New data points were constructed within the range of a discrete set of known data.

As shown in Figure 4, the interpolation method provided the best result in estimating the missing values in the dataset. Thus, this method was used to fill in the missing values.
4.2. Training of Models

Figure 5 shows the process of data integration, model training, and testing. First, the data from three datasets were integrated into one dataset by mapping the data using the DateTime index. Here, $T$, $WS$, $WD$, $H$, $AP$, and $SD$ represent temperature, wind speed, wind direction, humidity, air pressure, and snow depth, respectively, from the meteorological dataset. $R1$ to $R8$ represent eight roads from the traffic dataset, and $PM$ indicates PM$_{2.5}$ and PM$_{10}$ from the air quality dataset. In addition, it is important to note that machine learning methods are not directly adapted for time-series modeling. Therefore, it is mandatory to use at least one variable for timekeeping. We used the following time variables for this purpose: month ($M$), day of the week ($DoW$), and hour ($H$).
4.3. Experimental Results

4.3.1. Hyperparameters of Competing Models

Most machine learning models are sensitive to hyperparameter values. Therefore, it is necessary to accurately determine hyperparameters to build an efficient model. Valid hyperparameter values depend on various factors. For example, the results of the RF and GB models change considerably based on the max_depth parameter. In addition, the accuracy of the LSTM model can be improved by carefully selecting the window and learning_rate parameters. We applied the cross-validation technique to each model, as shown in Figure 6. First, we divided the dataset into training (80%) and test (20%) data. Furthermore, the training data were divided into subsets that used a different number of folds for validation. We selected several values for each hyperparameter of each model. The cross-validation technique determined the best parameters using the training subsets and hyperparameter values.

![Cross-validation flowchart](image-url)

**Figure 6.** Cross-validation technique to find the optimal hyperparameters of competing models. Adopted from [41].

Table 2 presents the selected and candidate values of the hyperparameters of each model and their descriptions. The RF and GB models were applied using Scikit-learn [41]. As both models are tree-based ensemble methods and implemented using the same library, their hyperparameters were similar. We selected the following five essential hyperparameters for these models: the number of trees in the forest (n_estimators, where higher values increase performance but decrease speed), the maximum depth of each tree (max_depth), the number of features considered for splitting at each leaf node (max_features), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf, where a higher value helps cover outliers). We selected the following five essential hyperparameters for the LGBM model using the LightGBM Python library: the number of boosted trees (n_estimators), the maximum tree depth for base learners (max_depth), the maximum tree leaves for base learners (num_leaves), the minimum number of samples of a parent node (min_split_gain), and the minimum number of samples required at a leaf node (min_child_samples). We used the grid search function to evaluate the model for each possible combination of hyperparameters and determined the best value of each parameter. We used the window size, learning rate, and batch size
as the hyperparameters of the deep learning models. The number of hyperparameters for the deep learning models was less than that for the machine learning models because training the deep learning models required considerable time. Two hundred epochs were used for training the deep learning models. Early stopping with a patience value of 10 was used to prevent overfitting and reduce training time. The LSTM model consisted of eight layers, including LSTM, RELU, DROPOUT, and DENSE. The input features were passed through three LSTM layers with 128 and 64 units. We added dropout layers after each LSTM layer to prevent overfitting. The GRU model consisted of seven GRU, DROPOUT, and DENSE layers. We used three GRU layers with 50 units.

Table 2. Hyperparameters of competing models.

| Model | Parameter          | Description                     | Options               | Selected |
|-------|--------------------|---------------------------------|-----------------------|----------|
| RF    | n_estimators       | Number of trees in the forest   | 100, 200, 300, 500, 1000 | 500      |
|       | max_features       | Maximum number of features on each split | auto, sqrt, log2 | auto     |
|       | max_depth          | Maximum depth in each tree      | 70, 80, 90, 100       | 80       |
|       | min_samples_split  | Minimum number of samples of parent node | 3, 4, 5                | 3        |
|       | min_samples_leaf   | Minimum number of samples to be at a leaf node | 8, 10, 12            | 8        |
| GB    | n_estimators       | Number of trees in the forest   | 100, 200, 300, 500, 1000 | 100     |
|       | max_features       | Maximum number of features on each split | auto, sqrt, log2 | auto     |
|       | max_depth          | Maximum depth in each tree      | 80, 90, 100, 110      | 90       |
|       | min_samples_split  | Minimum number of samples of parent node | 2, 3, 5                | 2        |
|       | min_samples_leaf   | Minimum number of samples of parent node | 1, 8, 9, 10            | 8        |
| LGBM  | min_split_gain     | Minimum number of samples of parent node | 2, 3, 5                | 2        |
|       | min_child_samples  | Minimum number of samples of parent node | 1, 8, 9, 10            | 9        |
| GRU   | seq_length         | Number of values in a sequence  | 18, 20, 24            | 24       |
|       | batch_size         | Number of samples in each batch during training and testing | 64    | 64       |
|       | epochs             | Number of times that entire dataset is learned | 200 | 200     |
|       | patience           | Number of epochs for which the model did not improve | 10 | 10 |
|       | learning_rate      | Tuning parameter of optimization | 0.01, 0.1 | 0.01 |
|       | layers             | GRU block of deep learning model | 3, 5, 7 | 3       |
|       | units              | Neurons of GRU model | 50, 100, 120 | 50 |
| LSTM  | seq_length         | Number of values in a sequence  | 18, 20, 24            | 24       |
|       | batch_size         | Number of samples in each batch during training and testing | 64 | 64       |
|       | epochs             | Number of times that entire dataset is learned | 200 | 200     |
|       | patience           | Number of epochs for which the model did not improve | 10 | 10 |
|       | learning_rate      | Tuning parameter of optimization | 0.01, 0.1 | 0.01 |
|       | layers             | LSTM block of deep learning model | 3, 5, 7 | 5   |

The hyperparameters were tuned using cross-validation.
4.3.2. Impacts of Different Features

The first experiment compared the error rates of the models using three different feature sets: meteorological, traffic, and both combined. The main purpose of this experiment was to identify the most appropriate features for predicting air pollutant concentrations. Figure 7 shows the RMSE values of each model obtained using the three different feature sets. The error rates obtained using the meteorological features are lower than those obtained using the traffic features. Furthermore, the error rates significantly decrease when all features are used. Thus, we used a combination of meteorological and traffic features for the rest of the experiments presented in this paper.

![Graph showing RMSE values for different models and feature sets.](image)

**Figure 7.** RSME in predicting (a) PM$_{10}$ and (b) PM$_{2.5}$ with different feature sets.

4.3.3. Comparison of Competing Models

Table 3 shows the $R^2$, RMSE, and MAE of the machine learning and deep learning models for predicting the 1 h AQI. The performance of the deep learning models is generally better performance than that of the machine learning models for predicting PM$_{2.5}$ and PM$_{10}$ values. Specifically, the GRU and LSTM models show the best performance in predicting PM$_{10}$ and PM$_{2.5}$ values, respectively. The RMSE of the deep learning models is approximately 15% lower than that of the machine learning models in PM$_{10}$ prediction. Figure 8 shows the PM$_{10}$ and PM$_{2.5}$ predictions obtained using all models. The blue and orange lines represent the actual and predicted values,
respectively. The PM$_{2.5}$ values predicted by the LSTM model are 27\% more accurate than those predicted by the other models.

**Table 3.** Error rates of competing models.

| Model  | PM$_{10}$  | PM$_{2.5}$ |
|--------|------------|------------|
|        | $R^2$  | RMSE | MAE | $R^2$  | RMSE | MAE |
| RF     | 83.71\%  | 9.77  | 6.73 | 82.16\%  | 6.75  | 4.79 |
| GB     | 84.66\%  | 9.48  | 6.4  | 84.88\%  | 6.21  | 4.27 |
| LGBM   | 83.87\%  | 9.72  | 6.72 | 85.93\%  | 5.99  | 4.35 |
| GRU    | 85.62\%  | 8.34  | 5.07 | 84.01\%  | 6.62  | 4.84 |
| LSTM   | 84.81\%  | 8.72  | 5.41 | 91.16\%  | 4.96  | 3.44 |

(a) (b) (c) (d) (e) (f) (g) (h)
4.3.4. Comparison of Prediction Time

We performed an experiment to analyze the effect of time scales (1 h, 3 h, 6 h, and 12 h) on the accuracy of the machine learning and deep learning models. Figure 9 shows the results of the experiment. The LSTM model shows better RMSE values at time scales of 1 h, 3 h, and 6 h. However, the machine learning models show better RMSE values at a time scale of 12 h. The RMSE values obtained using the GB model are relatively unaffected by the time scale.

Figure 9. RMSE in predicting (a) PM\textsubscript{10} and (b) PM\textsubscript{2.5} at different time scales.
4.3.5. Influence of Wind Direction and Speed

In recent years, numerous studies have considered the influence of wind direction and speed [42-44] on air quality. Wind direction and speed are essential features used by stations to measure air quality. On the basis of wind direction and speed, air pollutants may move away from a station or settle around it. Thus, we conducted additional experiments to examine the influence of wind direction and speed on the prediction of air pollutant concentrations. For this purpose, we developed a method of assigning road weights on the basis of wind direction. We selected the air quality measurement station that was located in the middle of all eight roads. Figure 10 shows the air pollution station and surrounding roads. On the basis of the figure, we can assume that traffic on Roads 4 and 5 may increase the AQI close to the station when the wind direction is from the east. In contrast, the other roads have a weaker effect on the AQI around the station. We applied the computed road weights to the deep learning models as an additional feature.

![Figure 10. Location of the air pollution station and surrounding roads.](image)

The roads around the station were classified on the basis of the wind direction (NE, SE, SW, and NW), as shown in Table 4. According to Table 4, the road weights were set as 0 or 1. For example, if the wind direction was NE, the weights of Roads 3, 4, and 5 were 1 and those of the other roads were 0. We built and trained the GRU and LSTM models using wind speed, wind direction, road speed, and road weight to evaluate the effect of road weights. Figure 11 shows the RMSE of the GRU and LSTM models with (orange) and without (blue) road weights. For the GRU model, the RMSE values with and without road weights are similar. In contrast, for the LSTM model, the RMSE values with road weights are approximately 21% and 33% lower than those without road weights for PM$_{10}$ and PM$_{2.5}$, respectively.

| Id | Numerical Value | Categorical Value | Roads |
|----|----------------|-------------------|-------|
| 1  | 1°–90°         | NE                | 3, 4, 5 |
| 2  | 91°–180°       | SE                | 1, 4, 5 |
| 3  | 181°–270°      | SW                | 1, 2, 5, 6 |
| 4  | 271°–360°      | NW                | 1, 2, 6, 7, 8 |
5. Discussion and Conclusions

We proposed a comparative analysis of predictive models for fine PM in Daejeon, South Korea. For this purpose, we first examined the factors that can affect air quality. We collected the AQI, meteorological, and traffic data in an hourly time-series format from January 1, 2018, to December 31, 2018. We applied the machine learning models and deep learning models with 1) only meteorological features, 2) only traffic features, and 3) meteorological and traffic features. Experimental results revealed that the performance of the models with only meteorological features was better than that with only traffic features. Furthermore, the accuracy of the models increased significantly when meteorological and traffic features were used.

Furthermore, we determined a model that is most suitable to perform the prediction of air pollution concentration. We examined three types of machine learning models (RF, GB, and LGBM models) and two types of deep learning models (GRU and LSTM models). The deep learning models outperformed the machine learning models. Specifically, the LSTM and GRU models showed the best accuracy in predicting PM$_{2.5}$ and PM$_{10}$ concentrations, respectively. The accuracies of the GB and RF models were similar. We also compared the effect of time scales (1 h, 3 h, 6 h, and 12 h) on the models. The AQI predicted at a time scale of 1 h was more accurate than that predicted at the other time scales.

Finally, we have analyzed the effect of road conditions on the prediction of air pollutant concentrations. Specifically, we measured the relationship between traffic and wind direction and speed. An air pollution measurement station surrounded by eight roads was selected. We set weights for each road based on the location and wind direction. The consideration of road weights reduced the RMSE by approximately 21% and 33% for PM$_{10}$ and PM$_{2.5}$, respectively.

We conducted the experiments based on time-series data (i.e., air pollution, meteorological, and traffic), which are widely used in predicting air pollutant concentration. Considering that nowadays, most countries or cities make their environmental data open publicly, we assume that the proposed methodology can be easily applied to predict air pollutant concentration in both local and international applications.

There are several limitations of our study that should be addressed in the future. Firstly, we considered only meteorological and traffic factors for air pollution. However, air pollution is affected by several other factors, which should be further investigated. Secondly, we only considered the roads located in the city center when analyzing the effect of road conditions on the prediction of air pollutant concentration. However,
suburban roads can also help characterize the overall air pollution of the city. Finally, we used a relatively small dataset of a one-year period. In the future, we aim to improve the prediction accuracy in two manners. The first is to consider different air pollution causes, such as power plants and industrial emissions. The second is to use more data, treat outliers, and tune the models.

**Author Contributions:** Conceptualization, M.H. and S.C.; methodology, M.H. and T.C.; formal analysis, M.H. and T.C.; data curation, M.H.; writing—original draft preparation, M.H.; writing—review and editing, T.C. and A.N.; visualization, M.H. and T.C.; supervision, A.N. and S.C.; project administration, S.C.; funding acquisition, S.C. and A.N. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2021010749, Development of Advanced Prediction System for Solar Energy Production with Digital Twin and AI-based Service Platform for Preventive Maintenance of Production Facilities).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Almetwally, A.A.; Bin-Jumah, M.; Allam, A.A. Ambient air pollution and its influence on human health and welfare: An overview. *Environ. Sci. Pollut. Res.* 2020, 27, 24815–24830.
2. Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and health impacts of air pollution: A review. *Front. Public Health* 2020, 8, 14.
3. Koo, J.H.; Kim, J.; Lee, Y.G.; Park, S.S.; Lee, S.; Chong, H.; Cho, Y.; Kim, J.; Choi, K.; Lee, T. The implication of the air quality pattern in South Korea after the COVID-19 outbreak. *Sci. Rep.* 2020, 10, 22462.
4. World Health Organization. Available online: https://www.who.int/mediacentre/news/releases/2014/air-pollution/en (accessed on 10 February 2021).
5. Zhao, CX.; Wang, Y.Q.; Wang, Y.J.; Zhang, H.L.; Zhao, B.Q. Temporal and spatial distribution of PM2.5 and PM10 pollution status and the correlation of particulate matters and meteorological factors during winter and spring in Beijing. *Environ. Sci.* 2014, 35, 418–427.
6. Annual Report of Air Quality in Korea 2018, National Institute of Environmental Research: Incheon, Korea, 2019.
7. Shapiro, M.A.; Bolsen, T. Transboundary air pollution in South Korea: An analysis of media frames and public attitudes and behavior. *East Asian Community Rev.* 2018, 1, 107–126.
8. Kim, H.C.; Kim, S.; Kim, BU.; Jin, C.S.; Hong, S.; Park, R.; Son, S.W.; Bae, C.; Bae, M.A.; Song, C.K.; et al. Recent increase of surface particulate matter concentrations in the Seoul Metropolitan Area, Korea. *Sci. Rep.* 2017, 7, 4710.
9. Korean Statistical Information Service. Available online: https://kosis.kr/eng/statisticsList/statisticsListIndex.do?menuId=M_01_01 (accessed on 10 February 2021).
10. Hitchcock, G.; Conlan, B.; Branningan, C.; Kay, D.; Newman, D. *Air Quality and Road Transport—Impacts and Solutions;* RAC Foundation, London, SW1Y 5HS, UK, 2014.
11. Daejeon Metropolitan City. Available online: https://www.daejeon.go.kr/dre/index.do (accessed on 2 March 2021).
12. Kim, H.; Kim, H.; Lee, J.T. Effect of air pollutant emission reduction policies on hospital visits for asthma in Seoul, Korea: Quasi-experimental study. *Environ. Int.* 2019, 132, 104954.
13. Lee, S.; Kim, S.; Kim, H.; Seo, Y.; Ha, Y.; Kim, H.; Ha, R.; Yu, Y. Tracing of traffic-related pollution using magnetic properties of topsoils in Daejeon, Korea. *Environ. Earth Sci.* 2020, 79, 485.
14. Dasari, K.B.; Cho, H.; Jakićimović, R.; Sun, G.M.; Yim, Y.H. Chemical composition of Asian dust in Daejeon, Korea, during the spring season. *ACS Earth Space Chem.* 2020, 4, 1227–1236.
15. Jeong, Y.; Youn, Y.; Cho, S.; Kim, S.; Huh, M.; Lee, Y. Prediction of Daily PM10 Concentration for Air Korea Stations Using Artificial Intelligence with LDAPS Weather Data, MODIS AOD, and Chinese Air Quality Data. *Korean J. Remote Sens.* 2020, 36, 573–586.
16. Park, J.; Chang, S. A particulate matter concentration prediction model based on long short-term memory and an artificial neural network. *Int. J. Environ. Res. Public Health* 2021, 18, 6801.
17. Kim, S.-Y.; Song, I. National-scale exposure prediction for long-term concentrations of particulate matter and nitrogen dioxide in South Korea. *Environ. Pollut.* 2017, 226, 21–29.

18. Eum, Y.; Song, I.; Kim, H.-C.; Leem, J.-H.; Kim, S.-Y. Computation of geographic variables for air pollution prediction models in South Korea. *Environ. Health Toxicol.* 2015, 30, e2015010.

19. Jang, E.; Do, W.; Park, G.; Kim, M.; Yoo, E. Spatial and temporal variation of urban air pollutants and their concentrations in relation to meteorological conditions at four sites in Busan, South Korea. *Atmos. Pollut. Res.* 2017, 8, 89–100.

20. Lee, M.; Lin, L.; Chen, C.Y.; Tsao, Y.; Yao, T.H.; Fei, M.H.; Fang, S.H. Forecasting air quality in Taiwan by using machine learning. *Sci. Rep.* 2020, 10, 4153.

21. Chang, Z.; Guojun, S. Application of data mining to the analysis of meteorological data for air quality prediction: A case study in Shenyang. *IOP Conf. Ser. Earth Environ. Sci.* 2017, 81, 012097.

22. Choubin, B.; Abdolshahnejad, M.; Moradi, E.; Querol, X.; Mosavi, A.; Shamshirband, S.; Ghamisi, P. Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain. *Sci. Total Environ.* 2020, 701, 134474.

23. Qadeer, K.; Rehman, W.U.; Sheri, A.M.; Park, I.; Kim, H.K.; Jeon, M. A long short-term memory (LSTM) network for hourly estimation of PM2.5 concentration in two cities of South Korea. *Appl. Sci.* 2020, 10, 3984.

24. Xayasouk, T.; Lee, H.; Lee, G. Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability* 2020, 12, 2570.

25. Comert, G.; Darko, S.; Huyhn, Nathan.; Elijah, B.; Eloise, Q. Evaluating the impact of traffic volume on air quality in South Carolina. *Int. J. Transp. Sci. Technol.* 2020, 9, 29–41.

26. Adams, M.D.; Requia, W.J. How private vehicle use increases ambient air pollution concentrations at schools during the morning drop-off of children. *Atmos. Environ.* 2017, 165, 264–273.

27. Askariyeh, M.H.; Venugopal, M.; Khreis, H.; Birt, A.; Zietsman, J. Near-road traffic-related air pollution: Resuspended PM2.5 from highways and arterials. *Int. J. Environ. Res. Public Health* 2020, 17, 2851.

28. Rossi, R.; Ceccato, R.; Gastaldi, M. Effect of road traffic on air pollution. Experimental evidence from COVID-19 lockdown. *Sustainability* 2020, 12, 8984.

29. Lešnik, U.; Mongus, D.; Jesenko, D. Predictive analytics of PM10 concentration levels using detailed traffic data. *Transp. Res. D Transp. Environ.* 2019, 67, 131–141.

30. Wei, Z.; Peng, J.; Ma, X.; Qiu, S.; Wangm, S. Toward periodicity correlation of roadside PM2.5 concentration and traffic volume: A wavelet perspective. *IEEE Trans. Veh. Technol.* 2017, 66, 10439–10452.

31. Catalano, M.; Galatìoto, F.; Bell, M.; Namdeo, A.; Bergantino, A.S. Improving the prediction of air pollution peak episodes generated by urban transport networks. *Environ. Sci. Policy* 2016, 60, 69–83.

32. Askariyeh, M.H.; Zietsman, J.; Autenrieth, R. Traffic contribution to PM2.5 increment in the near-road environment. *Atmos. Environ.* 2020, 224, 117113.

33. Korea Environment Corporation. Available online: https://www.airkorea.or.kr/ (accessed on 2 March 2021).

34. Korea Meteorological Administration. Available online: https://www.kma.go.kr/eng/index.jsp (accessed on 2 March 2021).

35. Daejeon Transportation Data Warehouse. Available online: http://tportal.daejeon.go.kr/(accessed on 2 March 2021).

36. Breiman, L. Random forests. *Mach. Learn.* 2001, 45, 5–32.

37. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 2001, 29, 1189–1232.

38. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* 2017, 30, 3149–3157.

39. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* 2014, arXiv:1412.3555.

40. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* 1997, 9, 1735–1780.

41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825–2830.

42. Kim, K.H.; Lee, S.B.; Woo, D.; Bae, G.N. Influence of wind direction and speed on the transport of particle-bound PAHs in a roadway environment. *Atmos. Pollut. Res.* 2015, 6, 1024–1034.

43. Kim, Y.; Guldmann, J.M. Impact of traffic flows and wind directions on air pollution concentrations in Seoul, Korea. *Atmos. Environ.* 2011, 45, 2803–2810.

44. Guerra, S.A.; Lane, D.D.; Marots, G.A.; Carter, R.E.; Hohl, C.M.; Baldauf, R.W. Effects of wind direction on coarse and fine particulate matter concentrations in southeast Kansas. *J. Air Waste Manag. Assoc.* 2006, 56, 1525–1531.