



Research Paper

Evaluation of Machine Learning Application on the Prediction of Particulate Matter Concentrations in Small/Medium-Sized City

Yeji Lee^{1,2} · Iseul Na^{1,2} · Younggyu Son^{1,2†}

¹Department of Environmental Engineering, Kumoh National Institute of Technology, Republic of Korea

²Department of Energy Engineering Convergence, Kumoh National Institute of Technology, Republic of Korea

(Received August 5, 2024; Revised August 28, 2024; Accepted October 2, 2024)

Objectives : In this study, Machine Learning (ML) algorithms were evaluated to predict the concentration of particulate matter (PM₁₀ and PM_{2.5}) using air quality and meteorological data in small/medium-sized city.

Methods : ML models, including Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Random Forest (RF), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGB), were used to predict PM₁₀ and PM_{2.5} concentrations. Five air quality variables, including NO₂, SO₂, CO, PM₁₀ and PM_{2.5}, and seven meteorological variables, including temperature, humidity, vapor pressure, wind speed, precipitation, local atmospheric pressure, and sea-level atmosphere pressure, were collected from three air quality monitoring stations and one meteorological observatory from 2017 to 2022. A total of 52,583 sets of data were used for ML. The prediction accuracies of the applied ML models were evaluated using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of determination (R²).

Results and Discussion : Higher ML performance was obtained when using the data including PM₁₀ and PM_{2.5} compared to the data excluding these variables. Among five different ML models, the XGB model showed the highest accuracy in predicting PM₁₀ and PM_{2.5} one hour in the future. However, poorer performance was obtained as the predicted period increased from one hour to 72 hours.

Conclusion : The application of ML algorithms for the short-term prediction of PM₁₀ and PM_{2.5} was successful in this study. However, more input variables and deep learning algorithms are needed for long-term prediction of PM₁₀ and PM_{2.5}.

Keywords : Machine learning, Air pollution, PM₁₀, PM_{2.5}, XGB model

The Korean text of this paper can be translated into multiple languages on the website of <http://jksee.or.kr> through Google Translator.

† Corresponding author

E-mail: yson@kumoh.ac.kr

Tel: 054-478-7637 Fax: 054-478-7859

© 2024, Korean Society of Environmental Engineers



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

연구논문

중소도시 미세먼지 농도 예측을 위한 머신러닝 기법 적용 가능성 평가

이예지^{1,2} · 나이슬^{1,2} · 손영규^{1,2*}

¹국립금오공과대학교 환경공학과

²국립금오공과대학교 에너지공학융합전공

목적: 본 연구는 중소도시의 대기환경 및 기상 데이터를 이용하여 미세먼지(PM₁₀ 및 PM_{2.5}) 농도를 예측하는 머신러닝 모델의 성능을 평가하였다.

방법: Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Random Forest (RF), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGB)의 5개 머신러닝 모델을 이용하여 PM₁₀과 PM_{2.5} 농도를 예측하였다. 2017년부터 2022년까지 5개의 대기환경 데이터(NO₂, SO₂, CO, PM₁₀, PM_{2.5})와 7개의 기상 데이터(기온, 습도, 증기압, 풍속, 강수량, 현지기압, 해면기압)를 3개의 대기질 측정소와 1개의 기상대에서 수집하였고, 이를 머신러닝 학습, 검증, 테스트 자료로 사용하였다. 머신러닝 예측 성능은 Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), Coefficient of determination(R²)의 3가지 성능 평가 지표를 이용하여 정확도를 평가하였다.

결과 및 토의: 미세먼지(PM₁₀ 및 PM_{2.5}) 농도 예측에서 입력 변수에 예측하고자 하는 미세먼지 정보가 포함된 데이터셋을 학습한 모델이 그렇지 않은 모델보다 더 높은 예측 성능을 보였다. XGB 모델이 다른 머신러닝 모델들보다 대부분의 경우에 가장 우수한 성능을 나타내었다. 그러나 미세먼지 농도를 예측하는 시간이 길어질수록 모든 머신러닝의 예측 정확도가 크게 감소하였다.

결론: 본 연구에서 머신러닝 기법을 이용한 1시간 후의 미세먼지(PM₁₀ 및 PM_{2.5}) 농도 예측은 충분히 적용 가능한 수준의 결과로 확인되었다. 그러나 장기간의 미세먼지를 예측하기 위해서는 딥러닝 모델을 적용하거나 미세먼지 농도에 영향을 미치는 추가적인 영향 인자 데이터를 확보하여 적용하는 것이 필요할 것이다.

주제어: 머신러닝, 대기오염, PM₁₀, PM_{2.5}, XGB 모델

1. 서론

미세먼지(Particulate Matter, PM)는 공기 중에 부유하는 흡입 가능한 고체 화합물로¹⁾, 직경에 따라 10 μm 이하의 PM₁₀과 2.5 μm 이하의 PM_{2.5}으로 구분한다. 미세먼지는 다양한 자연적, 인위적 발생원에 따라 생성되는데, PM₁₀은 사업장 및 자동차 배출가스와 황사 등으로 인하여 발생하며, PM_{2.5}은 황산화물(SO_x), 질소산화물(NO_x), 휘발성유기화합물(VOCs)과 같은 전구물질의 대기화학적 상호작용으로 인해 2차 생산된다.²⁾ WHO는 2019년 대기 오염으로 인해 전세계적으로 연간 420만 명이 사망하는 것으로 추정하였으며, 이는 미세먼지에 대한 노출로 인해 발생하였다고 발표하였다.³⁾ 미세먼지의 장시간 노출은 천식, 만성 폐쇄성 폐질환, 폐암 등 호흡기 질환은 물론, 골다공증, 당뇨병, 심혈관 질환, 면역 질환, 신경 질환, 정신 질환, 저체중아 출산, 영아 사망률과 관련이 있으며, 이에 대한 연구가 다수 보고된 바 있다.⁴⁻⁷⁾ 이러한 미세먼지의 위험성 저감과

대기질 개선을 위하여 미세먼지 농도 모니터링 및 예측에 대한 연구가 활발하게 보고되고 있다.

최근 수질, 대기, 토양 등 다양한 환경 분야에 IoT 기술이 활용되어 많은 데이터가 실시간으로 축적되고 있다. 축적된 데이터의 양적 및 질적 수준이 높아지고, 기존의 분석 기법으로 처리가 어려워짐에 따라 보다 효과적인 분석 기법의 도입이 요구되었다. 이러한 요구를 충족시킬 수 있는 방안 중 하나로 머신러닝 기법이 제안되고 있다.⁸⁾ 머신러닝이란 인공지능의 한 분야로서, 컴퓨터가 명시적으로 프로그래밍되지 않고도 주어진 데이터를 반복적으로 학습하여 목표한 결과를 찾아내는 기술을 말한다.⁹⁾ 환경 분야 중 머신러닝 기법이 가장 활발하게 적용되고 있는 분야는 상대적으로 많은 데이터가 축적되어 있는 대기환경분야이다.⁸⁾ Kang 등¹⁰⁾은 머신러닝 모델을 적용하여 동아시아의 NO₂, O₃ 농도를 추정하였으며, Du 등¹¹⁾과 Hosseinpour 등¹²⁾은 미국의 O₃ 농도를 예측하였고, Araki 등¹³⁾과 Chi 등¹⁴⁾은 각각 오사카와 중국 전역의

NO₂ 농도를 예측하였다. 또한, Cole 등¹⁵⁾은 머신러닝 모델을 통하여 중국의 지역 봉쇄가 대기 오염 수준에 미치는 영향을 확인하였으며, Xiao 등¹⁶⁾은 딥러닝 모델을 이용하여 중국의 발전소에 의한 대기 오염 모델링을 수행하였다.

머신러닝은 특히 상대적으로 관심이 높은 미세먼지 분석 연구에도 많이 사용되고 있는데, 주로 Multiple Linear Regression(MLR), Decision Tree Regression(DTR), Random Forest(RF), Extreme Gradient Boosting(XGB), Light Gradient Boosting Machine(LGB), Support Vector Regression(SVR) 등의 머신러닝 모델^{10,11,17,18)}과 Long Short Term Memory(LSTM), Convolutional Neural Network(CNN), Deep Neural Network(DNN) 등의 딥러닝 모델^{19,21)}을 사용하고 있다. 위성 데이터와

기상 데이터를 같이 이용하거나²²⁻²⁴⁾, 대기환경 데이터와 기상 데이터를 이용하여 미세먼지 농도를 예측한 결과들이 보고되고 있다.^{19,20,25,26)}

머신러닝을 활용한 미세먼지 농도 예측 연구에서 입력 변수를 구분하여 살펴보면, 예측하고자 하는 미세먼지 항목을 학습시키지 않고 나머지 대기환경 데이터 및 기상 데이터를 이용하여 미세먼지 농도를 예측하는 방법과^{19,21,23,25,27)}, 예측하고자 하는 미세먼지 항목을 학습에 포함하는 방법으로 구분할 수 있다.^{20,22,28,29)} 기존 연구에서 두 가지 방법을 동시에 적용하여 비교한 연구는 거의 보고되지 않고 있어 어떤 방법이 더 적합한지에 대한 연구가 필요할 수 있다. 또한 머신러닝을 활용하여 미래의 미세먼지의 농도를 예측하는 연구는

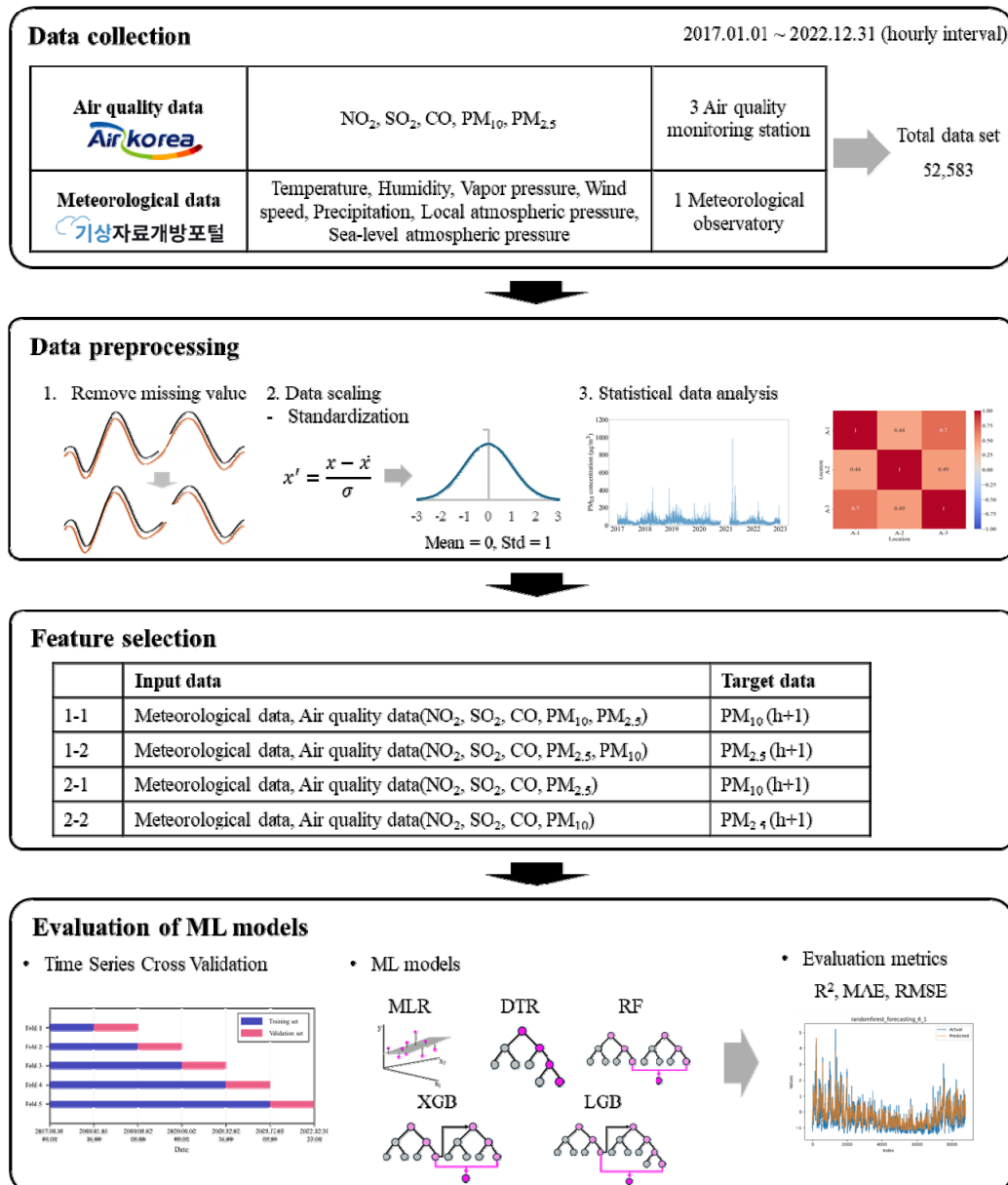


Fig. 1. Brief flowchart of machine learning for particulate matter prediction.

국의 연구에서 다수 보고되어 있으나 국내에서는 현재까지 거의 보고되지 않고 있다.^{20,21,29-31)}

본 연구에서는 미세먼지 예측에 자주 사용되는 머신러닝 모델을 활용하여 위에서 언급한 바와 같이 미세먼지의 포함 여부를 구분하여, 1시간 후의 미세먼지 농도 예측을 수행한 후 두 방법의 예측 성능을 비교하였다. 또한, 학습된 머신러닝 모델이 높은 신뢰성으로 1시간 보다 더 먼 미래의 미세먼지 농도를 예측할 수 있는지 여부를 확인하였다. 본 연구에서 사용된 머신러닝 기법은 MLR, DTR, RF, XGB, LGB이며, 학습된 모델의 미세먼지 농도 예측 성능을 비교평가하기 위해 Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), Coefficient of determination(R^2)을 활용하였다.

2. 실험방법

대기환경 데이터와 기상 데이터를 이용하여 미세먼지 농도를 예측하는 머신러닝 연구방법을 Fig. 1에 나타내었다. 본 연구에서 머신러닝 연구는 데이터 수집, 데이터 전처리, 입력/목표 변수 구분 및 학습 데이터셋 분할, 학습된 머신러닝 모델의 성능 평가 등을 포함한다.

2.1. 데이터 수집

미세먼지 농도 예측을 위한 머신러닝 입력 데이터 수집을 위하여 대기환경 데이터는 한국환경공단 에어코리아(<https://www.airkorea.or.kr>)에서 기상데이터는 기상자료개방포털(<https://data.kma.go.kr>)에서 다운로드 받아 이용하였다. 데이터 수집 대상 지역은 대한민국 경상북도 A시로 선정하였으며, 데이터 수집 기간은 2017년 1월 1일부터 2022년 12월 31일까지로 하였다. A시는 615 km²의 면적을 가지며, 국가산업단지를 보유한 도농복합도시로, 북쪽 지역은 주로 농업이 발달되어 있으며, 남서쪽 지역에는 상업 및 주택단지가

위치하고 있고, 남동쪽 지역에는 국가산업단지가 있다. 이러한 지역적 특성을 고려하여 Fig. 2에 나타낸 바와 같이 북쪽 지역에는 대기질 측정소를 설치/운영하고 있지 않으며, 총 5개의 대기질 측정소(A-1 - A-5) 모두 남쪽 지역에 설치/운영되고 있다. 본 연구에서는 3개 측정소(A-1, A-2, A-3)를 대상 측정소로 하였는데, 이는 이들 측정소가 가장 먼저 설치/운영되어 상대적으로 많은 5년 이상의 데이터를 확보할 수 있기 때문이다. 결과적으로 대기질 측정소 A-1, A-2, A-3는 각각 공업지역, 주거지역, 상업지역을 대표하는 것으로 판단하였다. 또한, A시의 기상 데이터는 한 곳에서만 측정/수집되었기 때문에 대상 측정소 3곳(A-1, A-2, A-3)의 기상 데이터는 모두 동일한 기상 데이터를 적용하였는데, 대상 지역의 기상 데이터는 대기질 측정소 지점에 따라 크게 변화하지는 않을 것으로 판단하였다. 다만, 보다 의미 있는 결과 도출을 위해서는 대기질 측정소에서의 기상 데이터 확인이 필요할 것으로 판단된다. 대기환경 데이터는 NO₂, SO₂, CO, PM₁₀, PM_{2.5}를, 기상 데이터는 기온, 풍속, 강수량, 현지기압, 해면기압, 습도, 증기압을 수집하여 머신러닝에 적용하였다. 대기환경 데이터가 1시간 간격으로 제공되어, 기상 데이터 역시 1시간 간격의 데이터를 이용하였다.

2.2. 데이터 전처리

수집한 데이터셋은 6년간의 총 52,583개이며, 일부 데이터셋은 항목별로 전체 데이터셋의 0.01%에서 9.10% 수준의 결측값이 존재하였다(강수량의 경우, 결측된 데이터는 비가 오지 않는 것으로 고려하여 0으로 대체하였으며, 결측값에서 제외하였음). 결측값은 최소 1시간에서 최대 5개월 동안 측정되지 않았다. 선행 연구에서는 linear interpolation, k-nearest neighbors 등의 방법^{27,31,32)}으로 결측값을 처리하였으나, 긴 기간 동안 결측값을 단순히 예측하여 처리할 경우, 계절적/시간적인 오차가 발생할 수 있어 본 연구에서는 결측값이

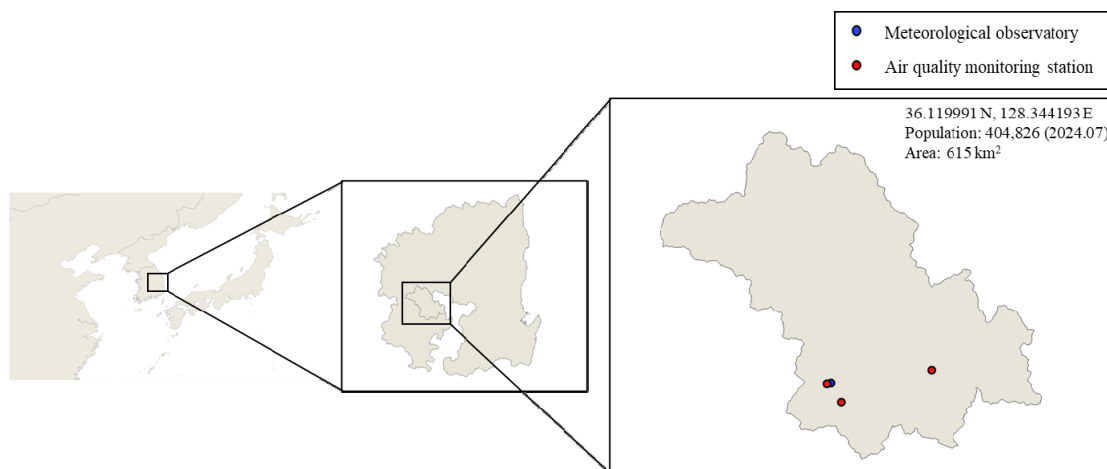


Fig. 2. Locations of the meteorological observatory and air quality monitoring stations in City A.

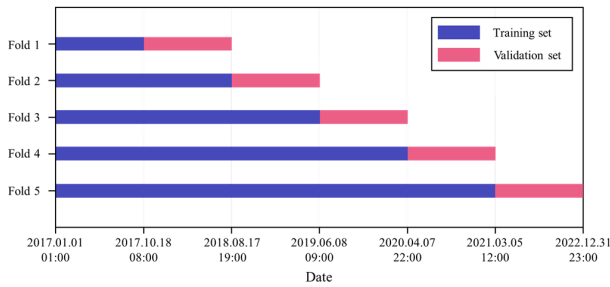


Fig. 3. Time Series Cross Validation using 5-year datasets for station A-1 (K = 6).

포함된 기간에 대한 데이터를 모두 제거하였다. 결측값 처리 후 측정소 별로 A-1: 43,829개, A-2: 42,968개, A-3: 46,142개 데이터셋을 이용하였다.

대기환경 및 기상 데이터의 항목별 스케일의 차이가 있으며, 이러한 차이가 머신러닝 모델의 학습에 영향을 미칠 수 있기 때문에 정규화 과정이 필요하다. 수집된 데이터의 항목별로 스케일의 차이를 조정하기 위하여 모든 항목의 데이터에 대해 평균값이 0이며, 표준편차가 1이 되게 하는 정규화(Standardization)를 수행한 후 이를 모델 학습에 적용하였다.^{33,34)}

머신러닝 모델의 학습 및 학습결과 평가를 위해 대상 데이터를 학습 데이터와 테스트 데이터로 분할하는 것이 필요하다. 총 6년간의 전체 데이터 중 5년간의 데이터(A-1: 36,451개, A-2: 35,675개, A-3: 38,916개 데이터셋, 2017년~2021년)를 학습 데이터로 지정하였고, 나머지 1년 간의 데이터를 테스트 데이터(A-1: 7,378개, A-2: 7,293개, A-3: 7,226개 데이터셋, 2022년)로 하였다. 또한, 다양한 조건의 데이터를 학습시켜 모델의 일반화 성능을 확인하기 위해 **Fig. 3**과 같이 시계열 교차 검증(Time Series Cross Validation)을 사용하였다(K = 6). 이는 데이터셋을 시기별로 K개로 구분하고, 구분된 데이터셋을 다시 학습 데이터(과거 데이터)와 검증 데이터(미래 데이터)로 구분하여 (K-1)번 반복 학습하는 것을 의미한다. 반복 학습이 수행된 모델의 성능 평가는 각 학습에 대한 성능 평가 지표의 평균값을 이용하였다.³⁵⁻³⁸⁾

2.3. 모델 적용 및 모델의 예측 성능 평가

미세먼지 농도를 예측하기 위하여 기존의 머신러닝을 활용한 미세먼지 예측 연구를 포함하는 환경공학 분야의 머신러닝 연구에서 자주 사용되는 5개의 머신러닝 모델을 적용하였다.^{22,28,30,32,39,40)} 비교적 간단한 알고리즘을 갖고 있는 머신러닝 모델인 MLR, DTR과 최근 자주 사용되는 트리 기반의 앙상블 모델인 RF, XGB, LGB를 본 연구에서 사용하였다. MLR 모델은 독립 변수의 영향력을 쉽게 이해할 수 있어 해석이 용이하나 선형 관계를 가정하므로 비선형 패턴을 모델링하기에 어렵다.⁴¹⁾ DTR 모델은 비선형 관계를 잘 포착하며

직관적이고 시각적으로 이해하기 쉽지만, 과적합에 취약하며 데이터에 민감하여 작은 변화에도 크게 영향을 받는다.⁴²⁾ RF 모델은 과적합에 강하며, 노이즈에 강하고 안정적인 예측 성능을 보이지만 연산 자원과 시간이 많이 필요하며, 많은 트리 생성을 통해 예측하기 때문에 해석이 어렵다.⁴³⁾ XGB 모델은 과적합 방지 및 결측값이 존재하는 경우에도 학습이 가능하나, 비교적 복잡한 구조의 모델로 튜닝이 필요하며 시간이 많이 필요하다.⁴⁴⁾ LGB 모델은 연산 자원과 시간이 비교적 적게 필요하나, 해당 모델도 튜닝이 필요하며, 과적합될 가능성이 있다.⁴⁵⁾ 결과적으로 아직까지 다양한 분야에 범용적으로 사용할 수 있는 머신러닝 모델이 제시되지 않고 있으며, 기존 연구에서는 다양한 머신러닝 모델을 복수로 선정·적용하여 그 결과를 비교하고, 적용 사례에 적합한 모델을 보고하고 있다. 본 연구에서도 기존의 이러한 연구 동향을 고려하여 5개의 머신러닝을 선정하여 적용하고 그 결과를 비교 분석하였다.

적용 모델은 scikit-learn (ver. 1.4.1, MLR, DTR, RF 포함), XGBoost(ver. 2.0.3, XGB 포함), LightGBM(ver. 4.3.0, LGB 포함) 라이브러리를 이용하였다. 또한, 머신러닝 모델의 제어 프로그램 작성을 위해 Python(ver. 3.12.3)을 사용하였다.

머신러닝에 의해 예측된 미세먼지 농도와 실제 미세먼지 농도의 비교를 통해 학습된 머신러닝의 성능 평가를 수행하였는데, 이를 위해 Eq. 1 ~ 3에 나타난 바와 같이 관련 연구에서 많이 사용되고 있는 RMSE, MAE, R²의 세 가지 성능 평가 지표를 활용하였다.^{19,25,29)} 각 성능 평가 지표에서 y_i 는 실제값, $\hat{y}_i$ 는 예측값, $\bar{y}_i$ 는 실제값의 평균, n은 샘플의 수이다. (그림파일 참고) RMSE는 실제값과 예측값 간의 오차의 제곱 평균의 제곱근을 계산한 값이며, MAE는 실제값과 예측값 간의 절대 오차의 평균을 계산한 값이다. RMSE는 큰 오차에 더 민감한 평가 지표이며, MAE는 모든 오차에 동일한 중요도를 부여하는 지표이다.^{19,46,47)} R²는 모델이 실제 데이터를 얼마나 잘 설명하는지를 나타내는 지표로⁴⁸⁾, 해당 지표를 이용하여 모델의 성능을 다양한 측면에서 평가하고자 하였다. MAE와 RMSE는 0에 가까울수록, R²는 1에 가까울수록 모델의 예측 성능이 높다.⁴⁶⁾

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (1)$$

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

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

Table 1. Descriptive statistics of the meteorological data.

	Temp. (°C)	Wind speed (m/s)	Precipitation (mm)	LAP* (hPa)	SAP** (hPa)	Humidity (%)	Vapor pressure (hPa)
Mean	13.9	1.0	1.3	1,010.7	1,016.6	67.0	12.9
Std	10.3	0.8	3.2	8.0	8.3	24.3	9.3
CV	0.7	0.8	2.5	0.008	0.008	0.4	0.7
Min	-14.5	0.0	0.0	965.4	970.8	2.0	0.2
25%	5.4	0.4	0.0	1,004.4	1,010.1	48.0	5.1
50%	14.7	0.9	0.2	1,010.9	1,016.8	68.0	10.0
75%	22.5	1.5	1.2	1,017.0	1,023.0	90.0	20.0
Max	37.7	7.2	57.2	1,032.8	1,039.3	99.0	41.9

* LAP: Local Atmospheric Pressure

** SAP: Sea-level Atmospheric Pressure

3. 결과 및 고찰

3.1. 수집된 데이터 통계 분석

2017년부터 2022년까지 6년 동안 수집한 대기환경(SO₂, CO, NO₂, PM₁₀, PM_{2.5}) 및 기상 데이터(기온, 풍속, 강수량, 현지기압, 해면기압, 습도, 증기압)에 대한 기본 통계 분석을 수행하였고, 이를 **Table 1**과 **Table 2**에 각각 나타내었다. 대상 지역의 기상 데이터는 여름철에 기온, 강수량 등이 증가하고, 현지기압 등이 감소하며 겨울철에는 반대 현상이 일어나는 등의 전형적인 한국 기상 데이터 양상을 나타내었다. **Fig. 15**에 나타난 바와 같이, 대기환경 데이터의 경우에도 전반적으로 각 항목별 시기별 변화가 나타났다. SO₂, CO, NO₂의 경우

겨울철인 11월부터 2월까지 평균 농도가 증가하였으며, O₃는 1부터 6월까지 농도가 증가하고 7월부터 12월까지 감소하였다. PM₁₀은 7월부터 10월까지 농도가 상대적으로 낮게 확인되고 나머지 기간에는 유사한 수준이었다. PM_{2.5} 역시 7월부터 10월기간 동안 상대적으로 낮은 농도가 확인되었으나, 11월부터 3월까지의 기간에 4월부터 6월까지의 기간보다 높은 수준을 나타내었다. 이는 겨울철에는 화석 연료 사용의 증가로 NO₂, SO₂, CO, PM₁₀, PM_{2.5}의 농도가 증가하며, 봄철에는 중국에서 이동하는 미세먼지로 인해 농도가 증가하게 된다. 또한, 여름철의 잦은 강수로 인하여 PM₁₀, PM_{2.5}의 농도가 감소하게 된다.⁴⁹⁾ 결과적으로 해당지역의 대기환경 항목들은 우리나라의 전형적인 계절적 특성이 뚜렷하게 나타났으며, 대상 기간동안 지속적으로 감소하는 경향을 나타내었다.

Table 2. Descriptive statistics of the air quality data for 3 stations.

Station	Air quality data	Mean	Std*	CV**	Min	25%	50%	75%	Max
A-1	SO ₂ (ppm)	0.003	0.001	0.333	0.001	0.002	0.003	0.004	0.036
	CO (ppm)	0.4	0.2	0.5	0	0.3	0.4	0.6	1.4
	NO ₂ (ppm)	0.011	0.007	0.636	0.001	0.006	0.009	0.014	0.060
	O ₃ (ppm)	0.032	0.019	0.609	0	0.016	0.030	0.043	0.132
	PM ₁₀ (μg/m ³)	41	31	0.756	1	23	35	52	1133
	PM _{2.5} (μg/m ³)	21	15	0.714	0	10	17	27	250
A-2	SO ₂ (ppm)	0.003	0.001	0.333	0.001	0.002	0.003	0.004	0.016
	CO (ppm)	0.4	0.2	0.5	0	0.3	0.4	0.5	2.3
	NO ₂ (ppm)	0.017	0.011	0.647	0.001	0.009	0.014	0.024	0.076
	O ₃ (ppm)	0.032	0.020	0.637	0.001	0.016	0.029	0.044	0.137
	PM ₁₀ (μg/m ³)	42	32	0.762	1	23	34	52	1000
	PM _{2.5} (μg/m ³)	20	14	0.700	1	10	16	26	140
A-3	SO ₂ (ppm)	0.003	0.001	0.333	0.001	0.002	0.003	0.004	0.014
	CO (ppm)	0.4	0.2	0.5	0	0.3	0.4	0.5	2.7
	NO ₂ (ppm)	0.015	0.011	0.733	0.001	0.008	0.012	0.020	0.086
	O ₃ (ppm)	0.032	0.020	0.625	0.001	0.016	0.031	0.044	0.127
	PM ₁₀ (μg/m ³)	38	28	0.737	1	22	31	46	985
	PM _{2.5} (μg/m ³)	20	15	0.750	0	9	16	26	171

* std: standard deviation

** CV: Coefficient of variation (CV = std/mean)

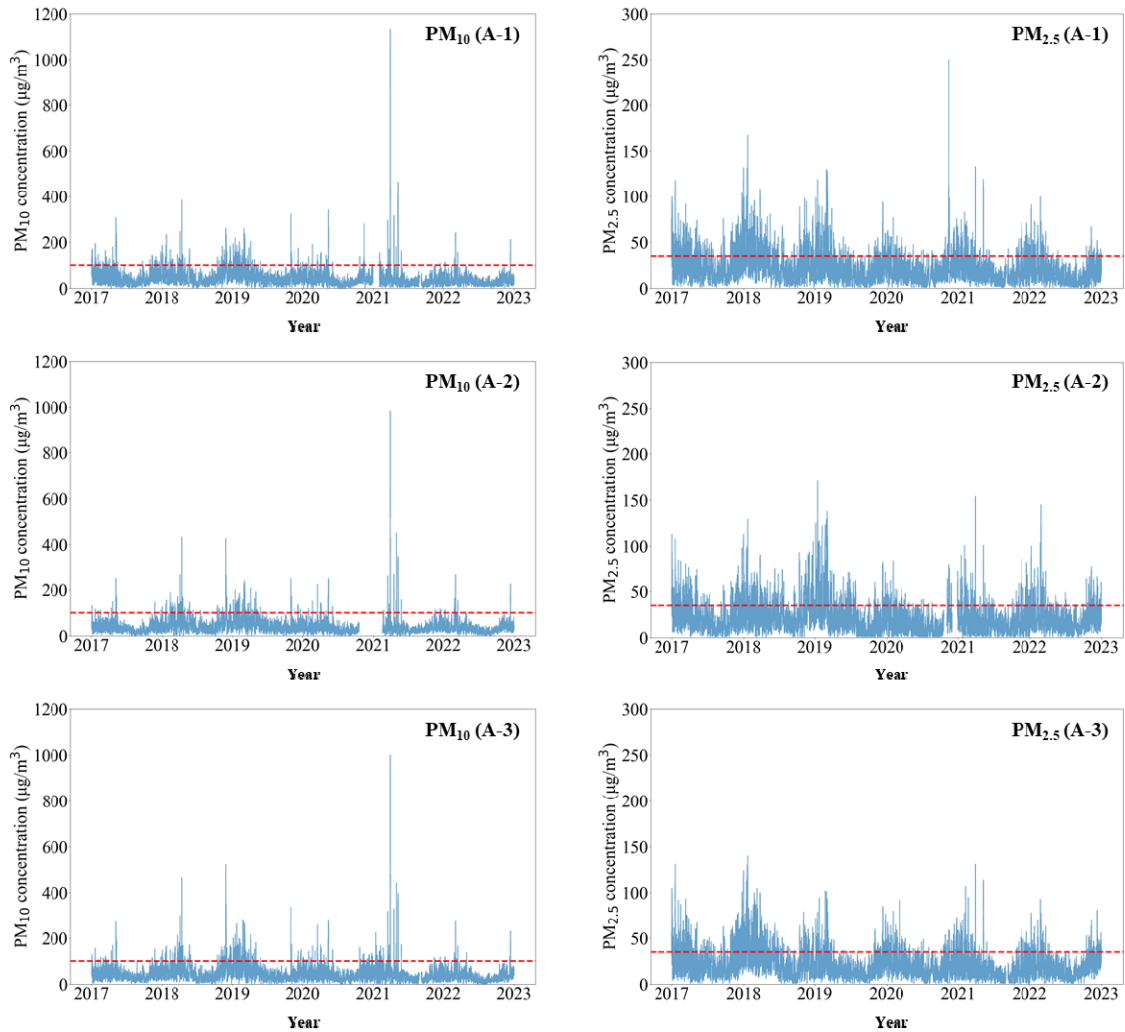


Fig. 4. Changes of hourly PM_{10} and $PM_{2.5}$ concentrations observed at 3 stations (Red line represents the national ambient air quality standard for PM_{10} and $PM_{2.5}$ (PM_{10} : $100 \mu g/m^3$, $PM_{2.5}$: $35 \mu g/m^3$)).

공업지역(A-1), 주거지역(A-2), 상업지역(A-3)으로 구분한 데이터 측정소 별로 살펴보면, 대부분의 항목에서 큰 차이는 확인되지 않았다. 그러나 NO_2 의 경우 A-1에서, PM_{10} 의 경우 A-3에서 각각 다른 측정소보다 상대적으로 낮은 수준을 나타내는 등 각 지역 특성에 따른 유의미한 차이를 논의하기 어려울 것으로 판단하였다. 이는 각 지역의 측정소가 각 지역을 대표하는 지점에 설치되지 않은 부분이 크게 작용하였을 것으로 고려된다.

각 항목별 각각의 대기환경 데이터에 대한 환경기준 만족 수준을 살펴보면, SO_2 , CO , NO_2 의 경우 환경기준을 초과한 사례가 한 건도 확인되지 않았으며, O_3 의 경우 세 개 측정소에서 각각 114~158회의 초과 사례가 확인되었는데, 이는 각 측정소 전체 측정 횟수 대비 0.26%~0.37% 수준이었다. PM_{10} 및 $PM_{2.5}$ 는 환경기준 초과 사례가 봄, 겨울철에 매우 높게 확인되었는데, 측정소별로 대기 환경 기준 초과 일수는 PM_{10} 의 경우 A-1: 61일, A-2: 39일, A-3: 64일, $PM_{2.5}$ 의 경우 A-1:

265일, A-2: 230일, A-3: 222일이었다 (PM_{10} : 24hr 평균 $100 \mu g/m^3$ 이하, $PM_{2.5}$: 24hr 평균 $35 \mu g/m^3$ 이하). Fig. 4에 세 개 측정소에 대한 PM_{10} 및 $PM_{2.5}$ 의 시계열 농도 변화를 나타내었다(다른 대기환경 데이터 농도 변화 그래프는 생략하였음). 2020년 11월 14일에는 A-1 측정소가 위치한 지역에서 화재로 인하여 높은 PM_{10} 및 $PM_{2.5}$ 농도가 관찰되었다. 2021년 3월 29일에 매우 높은 PM_{10} 및 $PM_{2.5}$ 농도가 관찰되었는데, 이는 황사로 인한 것으로 확인되었다.

2020년 이후의 PM_{10} , $PM_{2.5}$ 의 연평균 농도가 기존 수치 대비 각각 $10 \mu g/m^3$, $5 \mu g/m^3$ 정도 감소한 경향을 보였으나^{2,50)}, 여전히 대기환경 기준을 만족하지 못하는 사례는 확인되었다. 결과적으로 해당 지역은 $PM_{2.5}$ 및 PM_{10} 에 대한 모니터링 및 농도저감을 위한 관리가 지속적으로 필요할 것으로 판단되었다.

수집된 데이터를 대상으로 기상 및 대기환경 항목간 선형적 상관관계를 분석하기 위해 측정소별로 피어슨 상관 계수

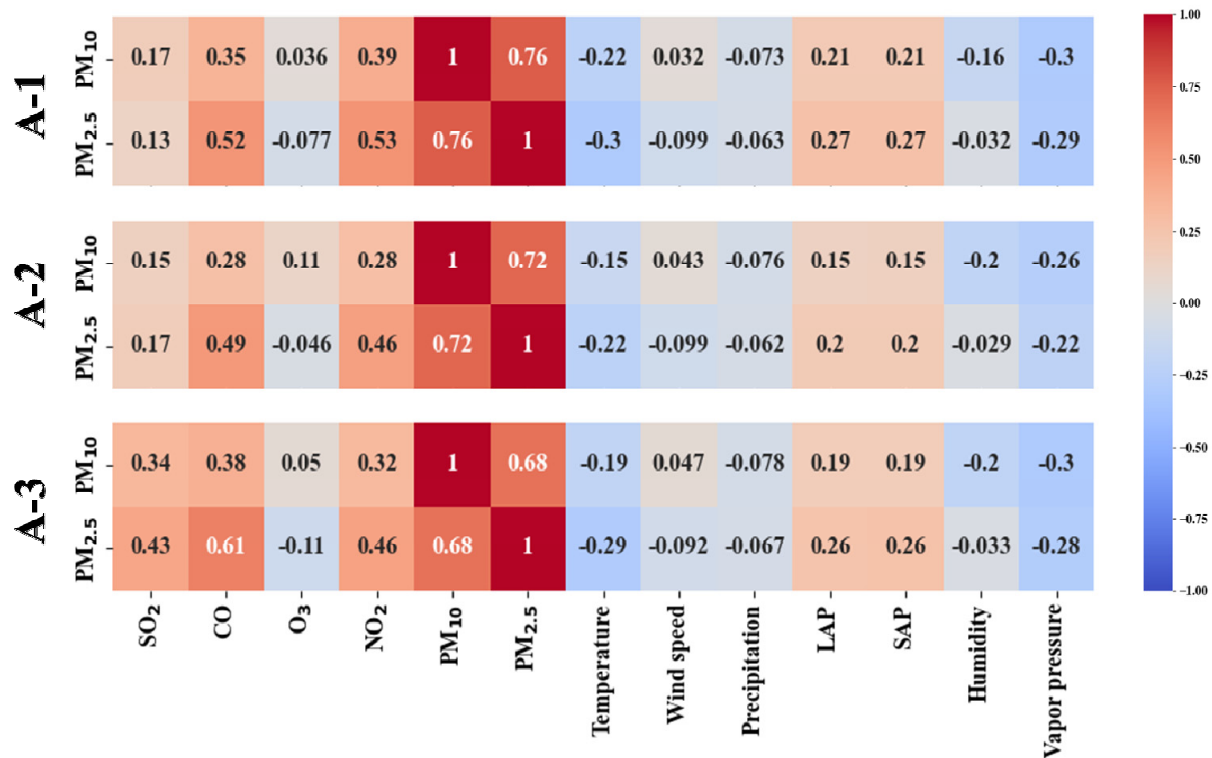


Fig. 5. Pearson correlation analysis of particulate matter (PM_{2.5} and PM₁₀) and other features for 3 stations.

를 산정하여 비교분석을 수행하였다. 세 개 측정소 모두 전반적으로 항목간의 선형적 상관성이 높지는 않은 것으로 확인되었다. 기존 연구에서는 상관 계수가 0.3~0.4 이상인 경우 상관성이 보통 이상인 것으로 보고하였다.^{21,50)} 일반적으로 기상 요인의 경우 풍속, 기압은 오염물질의 혼합에 영향을 주며, 강우는 습식 침착을 통해 미세먼지의 농도를 감소시킬 수 있는 것으로 알려져 있다.⁴⁹⁾ 그러나 본 연구에서는 이러한 내용이 뚜렷하게 확인되지 않았는데, 이는 기상데이터 측정 장소와 대기질 데이터 측정장소의 위치가 동일하지 않았기 때문일 수 있으며, 추가적인 연구가 필요할 것으로 판단되었다. 대기환경 항목간 가장 높은 피어슨 상관 계수는 서로 간의 선형적 상관성이 높을 수밖에 없는 PM_{2.5}와 PM₁₀ 사이의 상관 계수였으며, 0.68~0.76 수준을 나타내었다. PM_{2.5}와 PM₁₀ 사이 이외에 PM_{2.5} 혹은 PM₁₀와의 상관 계수가 높은 대기 환경 항목은 Fig. 5에 나타난 바와 같이 CO와 NO₂ 항목이었으며, PM_{2.5}가 좀 더 높은 상관성을 보였다(세 개 측정소에서 PM_{2.5}와 CO 및 NO₂에 대한 0.46~0.61의 상관 계수 확인). A-3 측정소에서는 PM_{2.5}와 SO₂와의 상관 계수가 0.43으로 상대적으로 높은 수준이 확인되기도 하였다. 미세먼지는 CO, NO₂, SO₂와 같은 연소 부산물에 의해 생성되며, NO₂, SO₂는 미세먼지의 주요 전구 물질이고, CO는 2차 에어로졸을 형성하는데 기여한다고 보고되고 있다.⁴⁹⁾ 본 연구에서 수집된 데이터 간의 선형적 상관성이 높지 않으나, 머신러닝 기법은 비선형 관계를 가진 데이터를 처리하고 정규분포를

따르지 않는 데이터의 분석이 가능하기 때문에⁵²⁾ 수집된 데이터를 통해 미세먼지 농도를 예측할 수 있음을 기대할 수 있다.

3.2. 머신러닝 학습 및 예측

기존 머신러닝에 의한 미세먼지 예측 연구에서 입력 변수에 미세먼지 항목 자체를 포함시켜 학습하는 경우^{19,21,23,25,27)}와 미세먼지 항목을 뺀 나머지 항목들로 학습하는 경우^{20,22,28,29)}로 구분되었다. 본 연구에서는 두 가지 방법을 적용하여 머신러닝 모델의 예측 성능을 비교 분석하였다. 기상 데이터 및 예측하고자 하는 미세먼지(PM₁₀ 또는 PM_{2.5})를 제외한 대기환경 데이터를 입력 변수로 이용하는 방법과 해당 미세먼지 농도를 입력 변수에 포함하는 방법으로 나누어 데이터셋을 구축하였다. 각 데이터셋을 5개의 머신러닝 모델(MLR, DTR, RF, XGB, LGB)을 통해 1시간 후의 PM₁₀ 및 PM_{2.5} 농도 예측에 대한 학습(Train), 검증(Validation), 테스트(Test)를 순차적으로 수행하였으며, 각 모델의 성능 평가 지표(RMSE, R²)를 미세먼지 항목 포함 여부에 따라 Fig. 6과 7에 각각 나타내었다 (검증 데이터에 대한 성능 평가 지표 및 학습/테스트 데이터에 대한 MAE는 부록 Fig. 2S~4S에 수록됨).

입력 변수에 예측하고자 하는 미세먼지 정보가 포함된 데이터셋을 학습한 모델이 그렇지 않은 모델보다 PM₁₀과 PM_{2.5} 모두에 대해 더 높은 예측 성능을 보였으며, 5개 적용 모델 중 RF, XGB, LGB 모델의 성능이 비교적 우수한 것으로 확

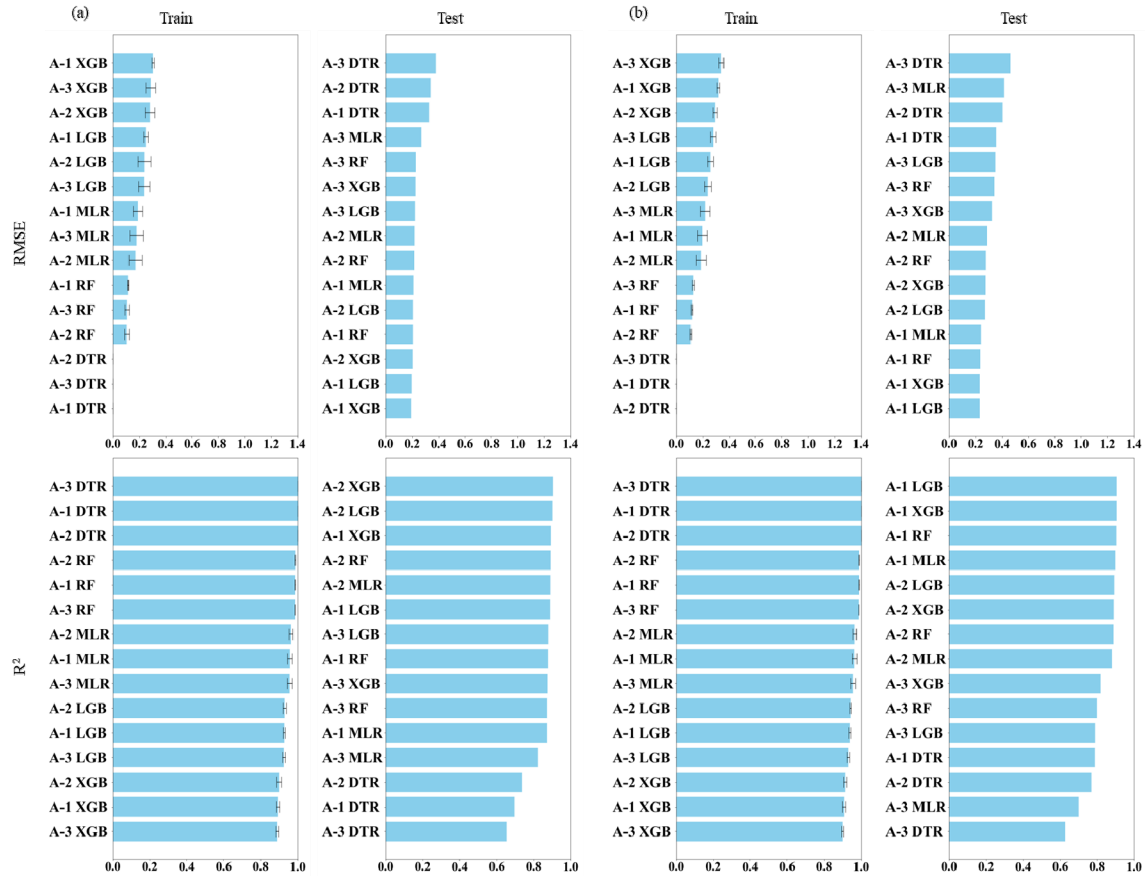


Fig. 6. Evaluation of ML models' performance using input data, including PM_{10} and $PM_{2.5}$, for predicting PM concentration one hour later at each station: (a) PM_{10} , (b) $PM_{2.5}$.

인되었다. XGB 모델이 전반적으로 가장 우수한 모델로 선정되었다(A-1: $R^2=0.895(PM_{10})$, $R^2=0.908(PM_{2.5})$ /A-2: $R^2=0.905(PM_{10})$, $R^2=0.891(PM_{2.5})$ /A-3: $R^2=0.877(PM_{10})$, $R^2=0.819(PM_{2.5})$). 본 연구에서 사용한 5가지 모델은 회귀 모델(MLR), 분류기반 모델(DTR), DTR 기반 앙상블 모델(RF, XGB, LGB)로 구분할 수 있다. 선형 회귀 모델인 MLR은 수집된 데이터 간의 선형적 상관관계가 낮기 때문에 예측 정확도가 분류 모델에 비해 낮았으며, DTR은 학습 데이터에 대한 과적합이 나타나 예측 성능이 낮게 나타났다.⁵³⁾ 결과적으로 3개의 앙상블 모델의 예측 성능이 MLR 및 DTR 모델보다 전반적으로 높게 나타났는데, 선행 연구에서도 RF, LGB 등의 앙상블 모델이 대부분 높은 예측 성능을 보였다. 모델에서 도출된 결과의 해석은 변수 중요도^{17,26,28,39)}, SHAP(Shapley Additive Explanation) method⁵⁴⁾ 등을 이용하여 변수들이 모델의 출력에 어떠한 영향을 미치는지를 통해 이루어질 수 있는데, 본 연구에서는 PM_{10} 및 $PM_{2.5}$ 가 1시간 후 PM_{10} 및 $PM_{2.5}$ 농도를 예측하는데 90% 이상의 중요도를 나타내었다. 머신러닝의 예측 성능 중 R^2 값이 음수가 나오는 경우가 일부 케이스에서 확인되었는데, 이는 머신러닝 학습이 제대로 되지 않았음을 의미한다.⁴⁶⁾

기존 연구에서 딥러닝 및 머신러닝을 이용하여 예측하고자 하는 미세먼지 데이터를 입력 변수에 포함하거나 포함하지 않은 연구 결과를 비교분석 하였다. 우선, 미세먼지 데이터를 입력 변수에 포함한 경우를 살펴보면, Wei 등²⁰⁾은 대만의 2개 지역(지룽시 및 시즈구)를 대상으로 해당 지역과 주변 지역의 PM_{10} , $PM_{2.5}$, O_3 , SO_2 , CO, NO_x , NO, NO_2 , THC, 기온, 풍속, 시간당 풍속, 풍향, 시간당 풍향, 강수량, 상대습도의 총 16개의 대기환경 및 기상 데이터를 입력 변수로 하여 1, 3, 6, 12 시간 후의 $PM_{2.5}$ 농도를 예측하였다. 사용된 모델은 DNN의 딥러닝 모델과 M5P, M5Rules, AMT(Alternating model trees), MLR의 머신러닝 모델이며, 가장 높은 성능을 나타낸 모델은 머신러닝 모델인 M5P($R^2=0.97$, 시즈구)이며, 1시간 후 예측 결과였다. Zhao 등²⁹⁾은 중국 4개 도시(베이징, 광저우, 시안, 선양)를 대상으로 $PM_{2.5}$, PM_{10} , SO_2 , NO_2 , O_3 , CO_2 , 기온, 이슬점 온도, 풍향, 풍속의 총 10개의 대기환경 데이터와 기상 데이터를 입력 변수로 하여 1, 2, 3 시간 후 $PM_{2.5}$ 농도를 예측하기 위하여 LSTM, RF-BiLSTM(RF-bidirectional LSTM)의 딥러닝 모델과 SVR, RF, DTR의 머신러닝 모델을 적용하였다. RF-BiLSTM 모델을 이용한 경우가 가장 높은 예측 성능을 얻을 수 있었으며(1시간 후 베이징

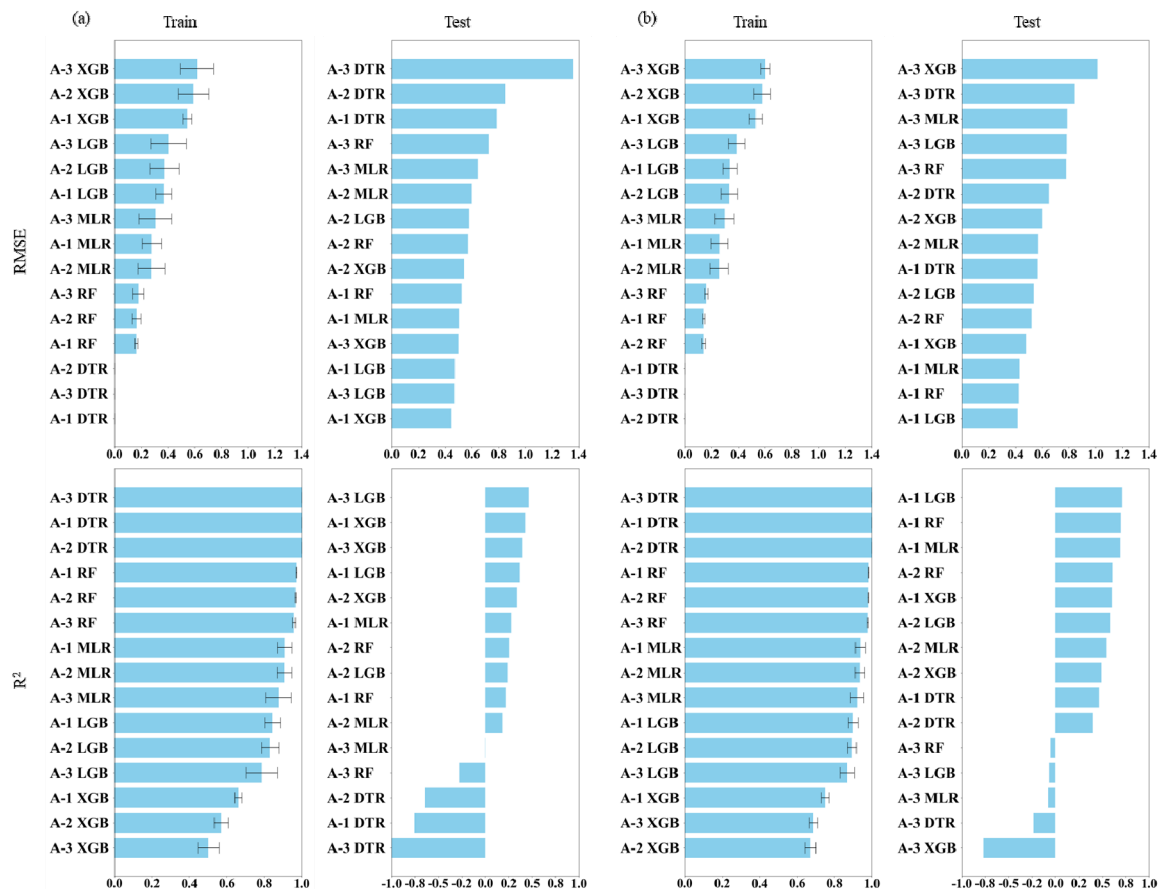


Fig. 7. Evaluation of ML models' performance using input data, excluding PM_{10} and $PM_{2.5}$, for predicting PM concentration one hour later at each station: (a) PM_{10} , (b) $PM_{2.5}$.

$PM_{2.5}$ 농도 예측 성능은 $R^2=1.0$), 머신러닝 모델 중에는 RF 모델이 가장 높은 성능을 나타내었다(1시간 후 베이징 $PM_{2.5}$ 농도 예측 성능은 $R^2=0.94$). Wood³⁰⁾는 $PM_{2.5}$ 의 값을 활용하여 생성한 변수들과 계절 성분의 13개 데이터를 입력 변수로 사용하여 딥러닝 모델인 LSTM, BiLSTM, MLP(Multi-layer perceptron),

CNN, EncDecCNN-LSTM(Encoder-Decoder-CNN-LSTM), EncDecLSTM-LSTM, GRU(Gated recurrent unit) 및 머신러닝 모델인 DT(Decision tree), AdaBoost, KNN(K-nearest neighbor), ELM(Extreme learning machine), MLR, RF, SVR, XGB 등의 15가지 모델을 이용하여 현재 $PM_{2.5}$ 농도 예측 성능을 비교하였는데, 머신러닝 모델인 MLR과 SVR 모델이 가장 높은 성능을 나타내었다. 해당 모델을 이용하여 다시 1, 3, 6, 12시간 후의 $PM_{2.5}$ 농도를 예측하였으며, 1시간 후 $PM_{2.5}$ 농도 예측 성능은 SVR 모델이 $R^2=0.83$ 으로 가장 높았다.

다음으로 미세먼지 데이터를 입력 변수에 포함하지 않는 경우를 살펴보면, Kim 등³⁹⁾은 기온, 이슬점 온도, 습도, 대기압, 상대습도, 풍향, 풍속, 가시거리, 운량, 행성경계층 높이, 강수량, 일 수, 요일, 미세먼지 관측소 위치의 기상, 날짜, 위치 데이터를 입력 변수로 하여 서울의 PM_{10} 과 $PM_{2.5}$ 농도를

예측하였다. 본 연구와 동일하게 RF, XGB, LGB의 머신러닝 모델에 적용되었으며, 가장 높은 성능을 보인 것은 LGB 모델이었다($R^2 = 0.86$). 이 연구는 1시간 간격의 데이터를 이용한 다른 연구들과는 다르게 1일 간격의 데이터를 이용하였다. Panneerselvam 등⁴⁰⁾은 LR(Linear regression), stepwise LR, Ridge, DTR, RF, SVR, GPR(Gaussian process regression)의 머신러닝 모델과 NN(Neural network)의 딥러닝 모델을 사용하여 기온, 기압, 풍속, 비습, 강수량의 5가지 기상 데이터를 입력 변수로 하여 태국 방콕의 PM_{10} 과 $PM_{2.5}$ 농도를 예측하였다. 가장 높은 성능을 보인 것은 GPR 모델로 $R^2 = 0.787$ 의 성능을 보였다. Ren 등¹⁹⁾은 딥러닝 모델인 TCN-LSTM(Temporal Convolutional network-LSTM)을 주요 모델로 적용하여 SO_2 , NO_2 , O_3 , CO, 풍속, 풍향, 상대습도, 강수량, 기압, 기온의 대기환경 및 기상 데이터를 입력 변수로 하여 중국 시안의 PM_{10} 및 $PM_{2.5}$ 농도를 예측하였으며, 해당 모델과 비교하기 위하여 딥러닝 모델인 CNN-LSTM, TCN, LSTM과 머신러닝 모델인 SVR, RF에도 동일하게 적용하여 예측을 수행하였다. TCN-LSTM 모델이 $R^2=0.88$ ($PM_{2.5}$), $R^2=0.87$ (PM_{10})로 가장 높은 예측 성능을 보였으며, 머신러닝 모델인 RF는 $R^2=0.79$ ($PM_{2.5}$), $R^2=0.75$ (PM_{10})의 예측 성능을 보였다.

Table 3. Evaluation of ML models'performance for predicting PM concentrations at station A-1 using RMSE, MAE, and R² metrics.

		PM ₁₀			PM _{2.5}		
	Model	RF	XGB	LGB	RF	XGB	LGB
1 hour	RMSE	0.206	0.192	0.197	0.236	0.233	0.233
	MAE	0.144	0.131	0.135	0.169	0.165	0.161
	R ²	0.879	0.895	0.890	0.905	0.908	0.908
6 hours	RMSE	0.485	0.450	0.443	0.527	0.515	0.505
	MAE	0.350	0.332	0.315	0.395	0.388	0.376
	R ²	0.330	0.423	0.440	0.531	0.552	0.569
12 hours	RMSE	0.567	0.538	0.523	0.610	0.594	0.577
	MAE	0.425	0.402	0.382	0.457	0.450	0.430
	R ²	0.085	0.174	0.219	0.372	0.405	0.438
24 hours	RMSE	0.669	0.627	0.615	0.738	0.670	0.684
	MAE	0.491	0.477	0.450	0.565	0.513	0.519
	R ²	-0.271	-0.116	-0.077	0.084	0.246	0.213
72 hours	RMSE	0.815	0.702	0.721	0.909	0.765	0.857
	MAE	0.625	0.552	0.551	0.695	0.603	0.651
	R ²	-0.875	-0.389	-0.468	-0.391	0.015	-0.236

이상의 입력 변수에 예측하고자 하는 미세먼지 농도가 없는 방법을 선택한 선행 연구의 예측 성능은 R²가 0.7 이상으로 미세먼지 농도를 제외하고 수행한 본 연구의 결과보다 비교적 높은 예측 성능을 보였다. 이는 미세먼지 예측 성능의 최적 조건이 지역, 연도, 계절, 위치, 농도 등 기상 및 대기환경 데이터의 특성에 따라 크게 달라질 수 있음을 의미한다.⁵⁵⁾ 또한, 미세먼지 농도가 다른 입력 변수와의 상관성이 높았을 수 있기 때문인데, 예를 들어 위의 Ren 등의 연구에서 PM₁₀과 PM_{2.5}에 대한 CO, NO₂, SO₂의 상관계수가 0.60 이상의 비교적 높은 수준을 나타내었다.¹⁹⁾

3.3. 1시간 이상의 미래 예측

1시간 이상 미래의 미세먼지 농도를 예측하였을 때의 성능 변화를 확인하기 위하여 모든 측정소에서 공통으로 성능이 가장 높은 RF, XGB, LGB 모델을 적용하여 6시간, 12시간, 24시간, 72시간 후의 미세먼지 농도 예측을 수행한 후 테스트에 대한 각 성능 평가 지표를 Table 3에 나타내었다. 모델 적용 결과, 모든 측정소에서 미세먼지 농도를 예측하는 시간이 길어질수록 머신러닝의 예측 정확도가 크게 감소하는 것으로 확인되었다. 6시간 후의 예측 정확도는 1시간 후의 예측 성능의 2배 이상 차이로 감소하였다. 또한 72시간 후 예측 결과의 대부분은 R²가 음수 값이 나타났으며, 특히 측정소 A-2의 RF 모델의 경우 12시간 후 PM₁₀ 농도 예측부터 R²가 음수로 나타난다. 이는 앞서 설명한 바와 같이 학습이 제대로 이루어지지 않은 것을 의미한다.⁴⁶⁾

예측 시간이 길어질수록 미세먼지 농도 예측 성능이 낮아지는 것은 선행 연구에서도 유사하게 나타났다. 앞서 언급한

선행 연구 중, Wei 등²⁰⁾의 연구에서 1시간 후 예측 시 가장 높은 성능을 나타낸 DNN 모델에서 예측하는 시간이 길어짐에 따라 모델의 예측 성능이 크게 낮아진 것을 확인하였다(1시간 후 PM_{2.5}: R²=0.91, 12시간 후 PM_{2.5}: R²=0.54). 또한, Zhao 등²⁹⁾의 연구에서 가장 높은 성능의 딥러닝 모델인 RF-BiLSTM의 1시간 후 예측 성능은 R²=1.00이며, 12시간 후 예측 성능은 R²=0.95으로 큰 성능의 차이가 없었으나, 머신러닝 모델인 SVM의 경우 1시간 후 예측 성능은 R²=0.97, 3시간 후 예측 성능은 R²=0.84로 성능이 감소하는 것을 확인하였다(RF-BiLSTM 모델을 제외한 나머지 모델들은 3시간 이후의 학습을 진행하지 않음). MLR과 SVR 모델을 적용하여 1, 3, 6, 12시간 후의 PM_{2.5} 농도를 예측한 Wood³⁰⁾의 연구에서도 두 개의 모델 모두 예측 시간이 길어질수록 예측 성능이 감소하는 것을 확인하였다(SVR 모델, 1시간 후: R²=0.827, 12시간 후: R²=0.396).

본 연구 결과와 선행 연구들을 통하여 미래의 미세먼지 농도를 예측하였을 때, 현재 수준의 머신러닝 기술로 학습된 데이터의 시간 간격보다 더 큰 간격으로 예측을 수행할 경우, 그 성능은 크게 감소하여 예측 결과를 이용하기 어려운 것으로 판단되었다⁵⁶⁾ (본 연구 및 대부분의 연구에서 이용 데이터는 1시간 간격 데이터를 이용함). 머신러닝 모델은 모델이 학습한 범위 내에서 값을 예측하기 때문에 미세먼지 농도가 매우 높거나 낮은 경우 해당 이벤트에 대한 데이터가 부족하여 이에 대해 모델이 잘 학습하지 못할 수 있다고 보고되었다⁵⁷⁾. 머신러닝 모델은 입력되는 데이터셋을 개별적으로 활용하기 때문에 스스로 데이터셋 간의 관계를 판단하지 못하나, 딥러닝 모델은 학습한 데이터셋 정보를 이후 학습에 사용하

여 순차적 관계를 함께 학습할 수 있어 시계열 데이터의 패턴을 잘 파악할 수 있다.⁵⁸⁾ 대부분의 연구에서는 머신러닝 모델보다 딥러닝 모델의 먼 미래 예측 성능이 보다 우수한 것으로 확인되는 사례가 있어³⁰⁾, 더 먼 미래의 예측을 위해서는 딥러닝 모델을 활용하는 연구가 필요할 것이며, 미세먼지 농도에 영향을 미치는 추가적인 영향 인자 데이터를 확보하는 것 역시 필요하다.

4. 결 론

본 연구에서는 MLR, DTR, RF, XGB, LGB의 5가지 머신러닝 모델을 활용하여 5개의 대기환경 데이터(NO_2 , SO_2 , CO , PM_{10} , $\text{PM}_{2.5}$)와 7개의 기상 데이터(기온, 습도, 증기압, 풍속, 강수량, 현지기압, 해면기압)를 통해 경상북도 1개 도시의 PM_{10} 및 $\text{PM}_{2.5}$ 농도를 예측하였다. 입력 변수에 예측하고자 하는 미세먼지 데이터를 포함하는 경우와 포함하지 않는 경우의 두가지 방법을 통하여 1시간 후의 PM_{10} 및 $\text{PM}_{2.5}$ 농도를 예측하고 각 모델에 대한 성능 평가 지표(RMSE, MAE, R^2)를 산정하였다. 입력 변수에 예측하고자 하는 미세먼지 데이터를 포함하는 방법의 성능이 그렇지 않은 방법보다 높은 정확도를 나타내었으며, 5개 적용 머신러닝 모델 중 XGB 모델이 가장 높은 성능을 보였다. 6시간 이후 예측부터 성능이 낮아지기 시작하여 그 이상의 미세먼지 농도 예측이 어려운 것을 확인하였다. 결과적으로 과거 6년 동안의 4만여개의 데이터셋을 이용하였으나 수시간 이상의 미래 미세먼지 농도 예측을 하기에는 현재 수준의 머신러닝 기법과 본 연구에서 활용한 입력 데이터 만으로는 한계가 있을 것으로 판단되었다. 보다 먼 미래의 미세먼지 농도 예측을 위해서는 다양한 딥러닝 기법의 적용 및 미세먼지와 관련된 영향 인자 데이터를 보다 많이 수집/학습하는 등의 연구가 필요할 것이다.

Acknowledgement

이 연구는 국립금오공과대학교 학술연구비로 지원되었음 (2022~2024).

References

1. J. T. Pryor, L. O. Cowley, S. E. Simonds, The physiological effects of air pollution: particulate matter, Physiology and Disease, *Frontiers in Public Health*, 10, 882569(2022).
2. National Institute of Environmental Research (NIER), Annual Report of Air Quality in Korea, 2022, 1-403(2023).
3. World Health Organization (WHO), [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health), December(2022)
4. A. Mukherjee, M. Agrawal, World air particulate matter: sources, distribution and health effects, *Environmental Chemistry Letters*, 15(2), 283-309(2017).
5. R. B. Hamanaka, G. M. Mutlu. Particulate matter air pollution: effects on the cardiovascular system. *Frontiers in Endocrinology*, 9, 680(2018).
6. J. Lee, Weerasinghe-Mudiyanselage PDE, B. Kim, S. Kang, J. S. Kim, C. Moon, Particulate matter exposure and neurodegenerative diseases: a comprehensive update on toxicity and mechanisms, *Ecotoxicology and Environmental Safety*, 115565(2023).
7. J. K. Choi, I. S. Choi, K. K. Cho, S. H. Lee, Harmfulness of particulate matter in disease progression, *Journal of Life Science*, 30(2), 191-201(2020).
8. X. Liu, D. Lu, A. Zhang, Q. Liu, G. Jiang, Data-driven machine learning in environmental pollution: gains and problems, *Environmental Science & Technology*, 56(4), 2124-2133(2022).
9. C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, *Electronic Markets*, 31(3), 685-695(2021).
10. Y. Kang, H. Choi, J. Im, S. Park, M. Shin, C. K. Song, S. Kim, Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over east asia, *Environmental Pollution*, 288, 117711(2021).
11. J. Du, F. Qiao, P. Lu, L. Yu, Forecasting ground-level ozone concentration levels using machine learning, *Resources, Conservation and Recycling*, 184, 106380(2022).
12. F. Hosseinpour, N. Kumar, T. Tran, E. Knipping, Using machine learning to improve the estimate of U.S. background ozone, *Atmospheric Environment*, 316, 120145(2024).
13. S. Araki, M. Shima, K. Yamamoto, Spatiotemporal land use random forest model for estimating metropolitan NO_2 exposure in Japan, *Science of The Total Environment*, 634, 1269-1277 (2018).
14. Y. Chi, M. Fan, C. Zhao, Y. Yang, H. Fan, X. Yang, J. Yang, J. Tao, Machine learning-based estimation of ground-level NO_2 concentrations over China, *Science of The Total Environment*, 807, 150721(2022).
15. M. A. Cole, R. J. R. Elliott, B. Liu, The impact of the wuhan covid-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach, *Environmental and Resource Economics*, 76(4), 553-580(2020).
16. D. Xiao, F. Fang, J. Zheng, C. C. Pain, I. M. Navon, Machine learning-based rapid response tools for regional air pollution modelling, *Atmospheric Environment*, 199, 463-473(2019).
17. X. Hu, J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland, Y. Liu, Estimating $\text{PM}_{2.5}$ concentrations in the conterminous united states using the random forest approach, *Environmental Science & Technology*, 51(12), 6936-6944 (2017).
18. Z. Gao, K. Do, Z. Li, Jiang X, Maji KJ, Ivey CE, et al., Predicting $\text{PM}_{2.5}$ levels and exceedance days using machine learning methods, *Atmospheric Environment*, 323, 120396 (2024).
19. Y. Ren, S. Wang, B. Xia, Deep learning coupled model based on TCN-LSTM for particulate matter concentration prediction, *Atmospheric Pollution Research*, 14(4), 101703(2023).
20. C. C. Wei, W. J. Kao, Establishing a real-time prediction system

- for fine particulate matter concentration using machine-learning models, *Atmosphere*, 14(12), 1817(2023).
21. E. Kristiani, H. Lin, J. R. Lin, Y. H. Chuang, C. Y. Huang, C. T. Yang, Short-term prediction of PM_{2.5} using LSTM deep learning methods, *Sustainability*, 14(4), 2068(2022).
22. C. Brokamp, R. Jandarov, M. Hossain, P. Ryan, Predicting daily urban fine particulate matter concentrations using a random forest model, *Environmental Science & Technology*, 52(7), 4173-4179(2018).
23. Y. Zhan, Y. Luo, X. Deng, H. Chen, M. L. Grieneisen, X. Shen, L. Zhu, M. Zhang, Spatiotemporal prediction of continuous daily PM_{2.5} concentrations across china using a spatially explicit machine learning algorithm, *Atmospheric Environment*, 155, 129-139(2017).
24. Y. Yang, Z. Wang, C. Cao, M. Xu, X. Yang, K. Wang, H. Guo, X. Gao, J. Li, Z. Shi, Estimation of PM_{2.5} concentration across china based on multi-source remote sensing data and machine learning methods, *Remote Sensing*, 16(3), 467(2024).
25. L. Mampitiya, N. Rathnayake, Y. Hoshino, U. Rathnayake, Forecasting PM₁₀ levels in sri lanka: a comparative analysis of machine learning models PM₁₀, *Journal of Hazardous Materials Advances*, 13, 100395(2024).
26. L. Lv, P. Wei, J. Li, J. Hu, Application of machine learning algorithms to improve numerical simulation prediction of PM_{2.5} and chemical components, *Atmospheric Pollution Research*, 12(11), 101211(2021).
27. A. Barthwal, D. Acharya, D. Lohani, Prediction and analysis of particulate matter (PM_{2.5} and PM₁₀) concentrations using machine learning techniques, *Journal of Ambient Intelligence and Humanized Computing*, 14(3), 1323-1338(2023).
28. B. Czernecki, M. Marosz, J. Jędruszkiewicz, Assessment of machine learning algorithms in short-term forecasting of PM₁₀ and PM_{2.5} concentrations in selected polish agglomerations, *Aerosol and Air Quality Research*, 21(7), 200586(2021).
29. J. Zhao, L. Yuan, K. Sun, H. Huang, P. Guan, C. Jia, Forecasting fine particulate matter concentrations by in-depth learning model according to random forest and bilateral long- and short-term memory neural networks, *Sustainability*, 14(15), 9430(2022).
30. D. A. Wood, Trend decomposition aids forecasts of air particulate matter (PM_{2.5}) assisted by machine and deep learning without recourse to exogenous data, *Atmospheric Pollution Research*, 13(3), 101352(2022).
31. A. Bekkar, B. Hssina, S. Douzi, K. Douzi, Air-pollution prediction in smart city, deep learning approach, *Journal of Big Data*, 8(1), 161(2021).
32. A. Pandya, R. Nanavaty, K. Pipariya, M. Shah, A comparative and systematic study of machine learning (ML) approaches for particulate matter (PM) prediction, *Archives of Computational Methods in Engineering*, 31(2), 595-614(2024).
33. W. Zhang, W. Huang, J. Tan, D. Huang, J. Ma, B. Wu, Modeling, optimization and understanding of adsorption process for pollutant removal via machine learning: recent progress and future perspectives, *Chemosphere*, 311, 137044(2023).
34. Z. Zhang, Y. Zeng, K. Yan, A hybrid deep learning technology for PM_{2.5} air quality forecasting, *Environmental Science and Pollution Research*, 28(29), 39409-39422(2021).
35. S. X. Lv, L. Peng, H. Hu, L. Wang, Effective machine learning model combination based on selective ensemble strategy for time series forecasting, *Information Sciences*, 612, 994-1023(2022).
36. P. Sadowsky, Forecasting solar stock prices using tree-based machine learning classification: how important are silver prices?, *The North American Journal of Economics and Finance*, 61, 101705(2022).
37. D. M. Zhang, X. Y. Guo, Y. M. Shen, W. D. Zhou, X. S. Chen, Data- and experience-driven neural networks for long-term settlement prediction of tunnel, *Tunnelling and Underground Space Technology*, 147, 105669(2024).
38. J. E. Park, J. H. Cho, Y. T. Kim, Prediction of PM_{2.5} dual-stage attention mechanism using surrounding area data, *Journal of Korean Institute of Intelligent Systems*, 32(5), 416-423(2022).
39. B. Y. Kim, Y. K. Lim, J. W. Cha, Short-term prediction of particulate matter (PM₁₀ and PM_{2.5}) in seoul, South Korea using tree-based machine learning algorithms, *Atmospheric Pollution Research*, 13(10), 101547(2022).
40. B. Panneerselvam, N. Ravichandran, U. C. Dumka, M. Thomas, W. Charoenlerkthawin, B. Bidorn, A novel approach for the prediction and analysis of daily concentrations of particulate matter using machine learning, *Science of The Total Environment*, 897, 166178(2023).
41. X. Su, X. Yan, C. L. Tsai, Linear regression, *WIREs Computational Statistics*, 4(3), 275-294(2012).
42. A. M. Ahmed, A. Rizaner, A. H. Ulusoy, A decision tree algorithm combined with linear regression for data classification, *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering*, 12-14(2018).
43. L. Breiman, Random forests, *Machine Learning*, 45(1), 5-32(2001).
44. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, San Francisco, pp. 785 - 794(2016).
45. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, Lightgbm: a highly efficient gradient boosting decision tree, *Advances in neural information processing systems*, 30, (2017).
46. D. Chicco, M. J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, *Peerj computer science*, 7, e623(2021).
47. T. Chai, R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature, *Geosci Model Dev*, 7(3), 1247-1250(2014).
48. A. D. Bucchianico, Coefficient of determination (R²), *Wiley StatsRef: Statistics Reference Online*, (2014).
49. S. U. Kim, K. Y. Kim, Physical and chemical mechanisms of the daily-to-seasonal variation of PM₁₀ in Korea, *Science of The Total Environment*, 712, 136429(2020).

50. W. C. Choi, K. S. Cheong, Analysis of the factors affecting fine dust concentration before and after COVID-19, *J Korean Soc Hazard Mitig*, 21(6), 395-402(2021).
51. Y. Xie, B. Zhao, L. Zhang, R. Luo, Spatiotemporal variations of PM_{2.5} and PM₁₀ concentrations between 31 chinese cities and their relationships with SO₂, NO₂, CO and O₃, *Particuology*, 20, 141-149(2015).
52. L. Li, S. Rong, R. Wang, S. Yu, Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: a review, *Chemical Engineering Journal*, 405, 126673(2021).
53. S. Buschjäger, K. Morik, Decision tree and random forest implementations for fast filtering of sensor data, *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(1), 209-222(2018).
54. M. Sharma, N. Kumar, S. Sharma, V. Jangra, S. Mehandia, S. Kumar, P. Kumar, Assessment of fine particulate matter for port city of eastern peninsular india using gradient boosting machine learning model, *Atmosphere*, 13(5), 743(2022).
55. Z. Peng, B. Zhang, D. Wang, X. Niu, J. Sun, H. Xu, J. Cao, Z. Shen, Application of machine learning in atmospheric pollution research: a state-of-art review, *Science of The Total Environment*, 910, 168588(2024).
56. B. Zhang, Y. Rong, R. Yong, D. Qin, M. Li, G. Zou, J. Pan, Deep learning for air pollutant concentration prediction: a review, *Atmospheric Environment*, 290, 119347(2022).
57. J. S. Pérez-Carrasquilla, P. A. Montoya, J. M. Sánchez, K. S. Hernández, M. Ramírez, Forecasting 24h averaged PM_{2.5} concentration in the Aburrá Valley using tree-based machine learning models, global forecasts, and satellite information, *Adv Stat Clim Meteorol Oceanogr*, 9(2), 121-135(2023).
58. M. Rahimzad, A. Moghaddam Nia, H. Zolfonoon, J. Soltani, A. Danandeh Mehr, H. H Kwon, Performance comparison of an LSTM-based deep learning model versus conventional machine learning algorithms for streamflow forecasting, *Water Resources Management*, 35(12), 4167-4187(2021).

Declaration of Competing Interest

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors and Contribution Statement

Yeji Lee

Department of Environmental Engineering, Kumoh National Institute of Technology, Department of Energy Engineering Convergence, Kumoh National Institute of Technology, M.S. Candidate, [ORCID](#) 0009-0006-6110-0277: Conceptualization, Data curation, Data analysis, Methodology, Visualization, Writing - original draft, Validation, Visualization.

Iseul Na

Department of Environmental Engineering, Kumoh National Institute of Technology, Department of Energy Engineering Convergence, Kumoh National Institute of Technology, Ph.D. Candidate, [ORCID](#) 0009-0009-1383-0242: Data curation, Data analysis, Methodology, Visualization, Writing - original draft, Validation, Visualization.

Younggyu Son

Department of Environmental Engineering, Kumoh National Institute of Technology, Department of Energy Engineering Convergence, Kumoh National Institute of Technology, Professor, [ORCID](#) 0000-0002-7379-7300: Conceptualization, Data analysis, Methodology, Writing - original draft, Funding acquisition, Project administration, Resources, Supervision, Validation.