The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.6, pp.251~259
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.6.251

Development of a Deep Learning-based Midterm PM2.5 Prediction Model Adapting to Trend Changes

Dong Jun Min† · Hyerim Kim†† · Sangkyun Lee†††

ABSTRACT

Fine particulate matter, especially PM2.5 with a diameter of less than 2.5 micrometers, poses significant health and economic risks. This study focuses on the Seoul region of South Korea, aiming to analyze PM2.5 data and trends from 2017 to 2022 and develop a mid-term prediction model for PM2.5 concentrations. Utilizing collected and produced air quality and weather data, reanalysis data, and numerical model prediction data, this research proposes an ensemble evaluation method capable of adapting to trend changes. The ensemble method proposed in this study demonstrated superior performance in predicting PM2.5 concentrations, outperforming existing models by an average F1 Score of approximately 42.16% in 2019, 58.92% in 2021, and 34.79% in 2022 for future 3 to 6-day predictions. The model maintains performance under changing environmental conditions, offering stable predictions and presenting a mid-term prediction model that extends beyond the capabilities of existing deep learning-based short-term PM2.5 forecasts.

Keywords : Particulate Matter, PM2.5, Time Series Forecast, Deep Learning, Ensemble

경향성 변화에 대응하는 딥러닝 기반 초미세먼지 중기 예측 모델 개발

민 동 준† · 김 혜 림†† · 이 상 근†††

요 약

초미세먼지, 특히 지름이 2.5㎛ 이하인 PM2.5는 인체 건강과 경제에 큰 피해를 주는 오염물질이다. 본 연구는 대한민국 서울 지역을 중심으로, 2017년부터 2022년까지 자료를 수집하여 PM2.5 데이터 분석 및 데이터 경향성 변화 추이를 분석하고, PM2.5 중기 예측 모델을 개발하는 것을 목표로 한다. 수집, 생산된 대기질 및 기상 데이터, 재분석 데이터, 수치모델 예측 데이터를 바탕으로, 모델을 학습하고 이를 통합한 경향성 변화에도 대응할 수 있는 앙상블 기법을 제안한다. 본 연구에서 제안하는 앙상블 기법은 PM2.5 농도 예측 성능 면에서 기존 모델 대비 미래 D+3~D+6 예측일 F1 Score 기준 평균 2019년 약 42.16%, 2021년 약 58.92%, 2022년 약 34.79% 높은 성능을 보였다. 제안한 모델은 변화하는 환경 조건에도 성능을 유지함으로써 안정적인 예측을 가능하게 하며, 기존 딥러닝 기반 PM2.5 단기 예측보다 먼 예측을 수행하는 중기 예측 모델을 제시한다.

키워드 : 초미세먼지, PM2.5, 시계열 예측, 딥러닝, 앙상블

1. 서 론

초미세먼지(PM2.5)는 지름이 2.5 마이크로미터(㎛) 이하인 입자로, 인구 집중화와 급격한 산업화로 인해 도심 지역에서 많이 배출되며, 황산염과 질산염 같은 물질이 대기 중에서 반응하여 2차 적으로 생성된다. WHO의 대기질 지침[1]에 따르면, PM2.5는 호흡기 및 심혈관 질환, 폐암으로 인한 사망률 증가와 연관이 있어 심각한 대기질 문제로 여겨진다. Bae et al. [2]과 Lee et al. [3]의 연구에 따르면, 대한민국에서 발생하는 고농도 PM2.5 사례는 대부분 장거리에서 유입된 PM2.5와 전구물질이 주요 원인이며, 지역적 대기 정체로 인한 오염물질 축적도 농도 증가에 기여하는 것으로 나타났다. Cha et al. [4]는 대한민국의 겨울철 고기압에 의한 대기 정체와 지역난방 수요 증가도 겨울철 고농도 PM2.5 사례에 유의미한 영향을 미친다고 보고했다. 즉, PM2.5 농도는 단지 지역 배출량으로만 결정되는 것이 아니라, 종관 규모의 대기 순환 패턴의 영향도 받는 것을 알 수 있다.

† 준 회 원 : 고려대학교 정보보호대학원 정보보호학과 석사과정
†† 비 회 원 : 이화여자대학교 사범대학 연구교수
††† 종신회원 : 고려대학교 정보보호대학원 부교수
Manuscript Received : January 17, 2024
First Revision : April 16, 2024
Accepted : May 18, 2024
* Corresponding Author : Sangkyun Lee(sangkyun@korea.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

대한민국 정부는 PM2.5 피해 예방을 위해 대한민국 환경부는 2015년부터 대기화학수송모델인 CMAQ(Community Multiscale Air Quality) 모델을 활용하여 대한민국 19개 권역의 대기오염물질(PM10, PM2.5, O3)의 단기예보 (예보 당일~D+2)를 하루 4번 발표하고 있다. 2020년 5월 환경부 소속 국립환경과학원에서는 인공지능 기반 단기예보 시스템을 도입해 대기오염물질 단기예보의 정확도를 기존 70%에서 79%로 향상했다고 보고했으며 [5], 2020년 6월 환경부는 기존 단기예보 기간 이후 4일의 예측일 포함 총 7일간의 PM2.5 중기예보를 제공한다고 보고했다 [6].

기존의 인공지능 기반 PM2.5 예측 연구는 인도[7], 중국[8], 대한민국[9] 등 다양한 지역의 예보를 수행하고 있으나 연구의 예측 시점은 대부분 예보 당일 1시간 후부터 최대 D+2, 48시간에 대한 예측을 수행하고 있으며 D+3~D+6일에 해당하는 중기 시점 예보 시스템에 관한 연구는 발표된 내용이 없었다. 또한 최근 연구[9]에서도 2019년 이후의 대기질 데이터는 사용하고 있지 않은데, 연구진은 COVID-19로 인한 대기질 패턴 이상으로 인해 2019년 이후 생산된 데이터는 시스템 구축에 포함하지 않았다고 밝혔다. 따라서 2020년 이후 기간의 PM2.5에 대한 실험과 기존 PM2.5 중기 예측 시스템을 성능 향상을 위한 연구가 필요하다.

본 연구에서는 국립환경과학원에서 제공하는 주간예보 중 중기예보 기간(D+3~D+6 예측일)의 15시 PM2.5에 대한 '낮음' (0~35㎍/㎥)과 '높음' (36㎍/㎥ 이상) 2단계 등급 분류를 목표로 한다. 논문의 두 가지 주요한 기여점은 다음과 같다. 첫째, 대한민국 서울 권역의 D+3~D+6일 PM2.5 중기 예측 모델을 설계하고, 이를 기존의 예측 모델과 성능 면에서 비교하였다. 둘째, 다양한 분류와 회귀 모델의 실험 결과를 제시하였으며, 기존 연구들이 제시한 예측 기간 이후의 데이터에 대한 분석과 자료를 제공하고 경향성 변화가 있는 환경에서도 기존 모델 대비 높은 성능을 나타낼 수 있는 예측 기법에 대한 방법론을 제안한다.

논문의 구성은 다음과 같다. 제2장에서는 기존의 미세먼지 예측 방법론과 최근의 딥러닝 연구를 검토한다. 제3장에서는 데이터와 본 논문이 제안하는 PM2.5 중기 예측체계의 설계를 상세히 기술하고, 제4장에서는 이 모델의 성능을 다양한 실험을 통해 검증한다. 마지막으로, 제5장에서는 연구 결과를 요약하고 향후 연구 방향에 대해 제시한다.

2. 관련 연구

2.1 대기화학 수송모델 기반 미세먼지 예측 연구

미세먼지 예측 연구는 다양한 방법론을 활용하여 진행되어 왔다. 대기화학수송모델 기반의 예측 방법은 대표적인 전통적 방법의 하나로, 이러한 모델은 대기 중의 화학 반응과 물질의 수송을 계산하여 미세먼지 농도를 예측한다. Lee et al. [10]의 연구에서는 CMAQ 모델을 기반으로 한 대한민국의 PM2.5 단기예보 시스템인 KAQPS v1(Korean Air Quality Prediction System version 1)을 개발하였다. 이외에도 Shin et al. [11]은 전지구 데이터인 FNL(Final Analysis)과 GFS(Global Forecast System) 데이터를 활용하여 9일간의 PM2.5 중기예측을 수행하였으며, 이를 통해 FNL 데이터가 중기예측에 유용한 입력변수로 작용한다는 것을 확인하였다. 한편, 딥러닝과 같은 최신의 기술을 활용한 대기화학 수송모델 개선 연구도 활발히 진행되고 있다. Sayeed et al. [12]은 WRF(Weather Research and Forecasting)와 CMAQ 모델의 예측 결과에 딥러닝 기법인 합성곱 신경망 (CNN)을 적용하여 PM10, PM2.5, NO2의 예측 성능을 개선하였다. 딥러닝 모델은 그 자체의 비선형으로 인해, CMAQ과 같은 전통적인 모델의 성능을 개선하는 데 유의미한 결과가 있음을 보였다.

2.2 머신러닝, 통계, 딥러닝 기반 미세먼지 예측 연구

최근에는 대기화학 수송모델을 개선하는 통계적 방법이 아닌, 딥러닝 시스템을 활용한 예측 모델 연구가 활발히 진행되고 있다. Singh et al. [7]은 선형 방법과 비선형 방법을 비교한 결과 비선형 방법이 대기질 데이터에서 복잡한 비선형성을 포착할 수 있음을 발견하고 인공 신경망(Artificial Neural Network, ANN)과 같은 비선형 모델이 선형 모델보다 대기 오염물질 예측에 더 적합하다고 보고했다. Azid et al. [13]과 Zhao et al. [8]은 각각 말레이시아와 중국의 대기질을 예측하기 위해 신경망 기반 단기 초미세먼지 예측 모델을 제시했다. 또한 최근 많은 미세먼지 예측 연구에서는 CNN-LSTM 구조의 모델을 사용하여 1시간에서 72시간까지의 단기 미세먼지 예측 연구를 수행하였다 [9, 14].

Chang-Hoi, H et al. [15]은 대한민국 서울 PM2.5에 대한 RNN에 CMAQ 예측 데이터를 입력한 RNN-CMAQ 모델을 통해 단기 예측 모델을 수행하였고 RNN-CMAQ 모델이 단일 CMAQ 모델보다 높은 성능을 보이는 것을 확인했다.

3. 연구 방법

본 장에서는 시계열 데이터 기반의 대한민국 및 중국 지역의 기상 및 대기질 데이터를 활용한 대한민국 서울 권역의 D+3~D+6일 중기 PM2.5 예측 시스템에 대하여 설명한다. 첫 번째 단락에서는 수집된 시계열 데이터에 대한 분석 결과를 공유하며 두 번째 단락에서는 전처리 기법에 관해 설명한다. 세 번째 단락은 중기 예측 모델 구조에 관해 설명한다. 마지막, 네 번째 단락에서는 모델 학습 및 평가와 앙상블 기법에 관해 설명한다.

3.1 데이터 범위 및 설명

본 연구는 Fig. 1과 같이 대한민국 및 중국 지역의 74개 관측 지역 데이터를 사용한다. 활용 데이터는 대기질 및 기상 데이터(OBS), 환경예측국제기구의 재분석 데이터(FNL), 그리고 수치모델 예측 데이터(WRF-CMAQ)로 구성된다. OBS 데이터는 대한민국 기상청과 중국의 PM25.in, Air Matters에서 수집

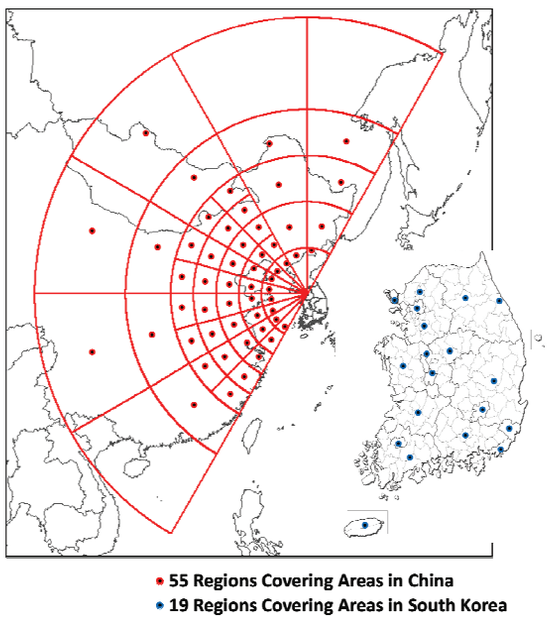

Fig. 1. Location of 74 Input Regions

하며, FNL 데이터는 미국 국립환경예측센터에서 제공한다. WRF-CMAQ 데이터는 기상청의 UM 수치예보 데이터를 기반으로 생성된다. 모든 데이터는 2017년 1월부터 2022년 12월까지, 하루 4번 (3, 9, 15, 21시) 수집되며, 총 8,764개의 데이터 포인트와 13,914개의 변수를 포함한다. 각 데이터의 영역별 변수 구성은 Table 1, 2와 같다. 표 하단의 Total은 각 전체 입력 영역의 변수의 수를 나타낸다.

1) PM2.5 데이터 경향성 분석

수집된 데이터의 경향 파악을 위해 서울 지역의 2017년 1월부터 2022년 12월까지 수집된 하루평균 PM2.5 농도 데이터에 대한 시각화 분석을 수행하였다. 본 연구에서는 Cleveland et al. [16]의 Seasonal-Trend Decomposition using LOESS (STL) 방법을 적용하여, Fig. 2에 분석 결과를 시각화하였다. 분석 기간 PM2.5 Trend는 연도별로 점진적으로 감소하는 경향을 보였으며, 특히 2021년과 2022년에는 이전 연도들과 비교해 더욱 감소하였다. Seasonal 분석에서는 연중 주기적인 변동성이 명확하게 나타났으며, 이는 계절 변화에 따른 대기질 변동성을 시사한다. 특히, 겨울과 여름에 농도가 상승하는

Table 1. Feature Information of OBS, FNL Data

	Category	hPa Lv.	Feature Name
OBS	Meta Info	-	WEEK_NO_KR, WEEK_NO_CN (2 Variables)
	Weather	surface	U, V, WS, PA, TA, TD, RH, RN (8*74=592 Variables)
	Air Pollutant	surface	PM10, PM25, O3, NO2, CO, SO2 (6*74=444 Variables)
	Total	1,038 Variables	
FNL	Wind & Temperature	{500, 700, 850, 925} hPa	U_*, V_*, WS_*, W_*, TA_*, RH_*, GEOH_*, DGH24_* (8*4*74=2,368 Variables)
	Average Deviation in Weather	{500, 700, 850, 925} hPa	ADV_U_*, ADV_V_*, ADV_WS_*, ADV_W_*, ADV_TA_*, ADV_RH_*, ADV_GEOH_* (8*4*74=2,072 Variables)
	Vorticity and Pressure Change	{500, 850} hPa	DVOR24_*, AVOR_* (2*2*74=296 Variables)
	Temperature and Humidity Indices	850 hPa	DZDT_850, EDI_850_Td (2*74=148 Variables)
	Stability Index	925 hPa	Ri_925 (74 Variables)
	Boundary Layer Height	surface	PBL_H (74 Variables)
	Total	5,032 Variables	
Target		PM2.5 of Seoul Region	

Table 2. Feature Information of WRF, CMAQ Data

	Category	hPa Lv.	Feature Name
WRF	Surface Weather	surface	T2, PSFC, U10, V10, WS, PBLH, GEOH, RH (8*74=592 Variables)
	Upper Air Weather	{500,700,850,925} hpa	U_*, V_*, W_*, WS_*, TA_*, RH_*, GEOH_* (7*4*74=2,072 Variables)
	Geopotential Height Change	{700,850,925} hpa	DGH24_* (1*3*74=222 Variables)
	Vorticity	{500,850} hpa	AVOR_*, DVOR24_* (2*2*74=296 Variables)
	Temperature and Humidity Indices	850	DZDT_850, EDI_850_Td (2*74=148 Variables)
	Stability Index	925	RI_925 (74 Variables)
	Total		3,404 Variables
CMAQ	Surface Concentration	surface	PM10, PM2_5, O3, NO2, CO, SO2, PMC, SO4, NO3, NH4, OC, EC (12*74=888 Variables)
	Upper Air Concentration	{500,700,850,925} hPa	PM10_*, PM2_5_*, O3_*, NO2_*, CO_*, SO2_*, PMC_*, SO4_*, NO3_*, NH4_*, OC_*, EC_* (12*4*74=3,552 Variables)
	Total		4,440 Variables

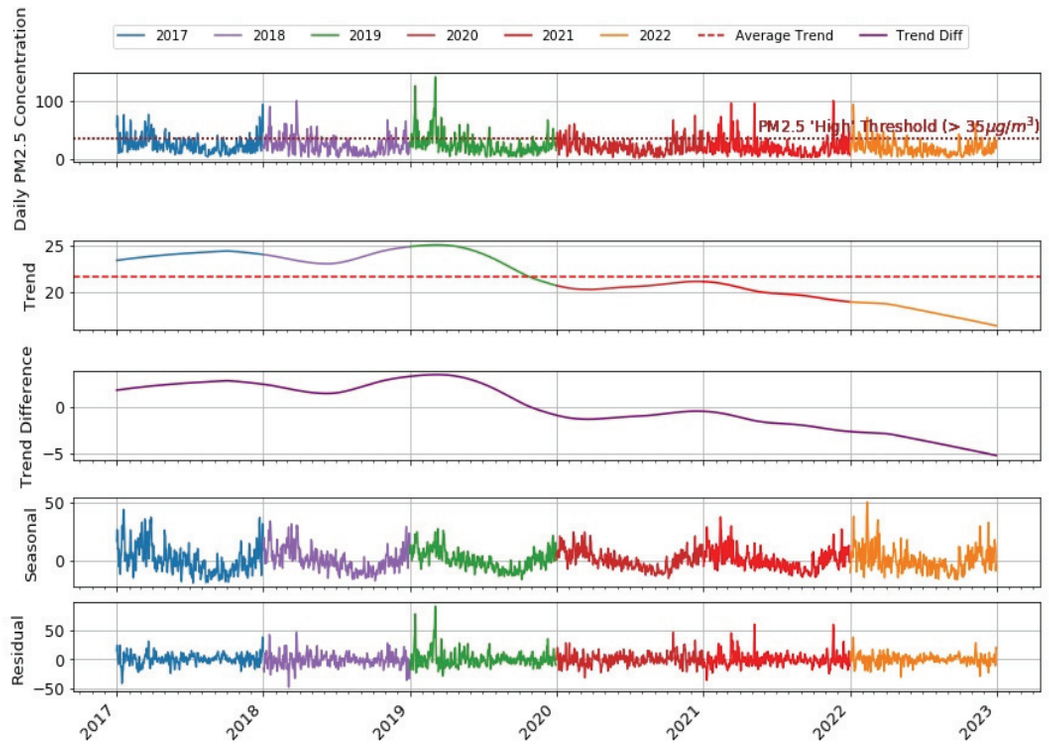

Fig. 2. Daily Seoul Region PM2.5 Data Analysis with STL method Result

Table 3. Seoul PM2.5 Data Statistics

Year	Normal Date	Bad Date	Mean	Std	Min	Max
2017	285	80	24.66	14.11	3.99	93.56
2018	289	76	22.81	15.33	2.72	99.48
2019	290	75	24.67	17.96	3.34	139.96
2020	293	73	**21.02**	**12.26**	1.86	74.33
2021	312	**53**	**19.73**	14.68	2.25	99.57
2022	313	**52**	**18.28**	**11.85**	2.09	93.27

경향은 대기 정체, 난방 및 에어컨 사용 증가와 같은 계절적 활동의 영향으로 해석된다. Residual 분석을 통해 확인된 특정 연도의 간헐적인 높은 피크는 추후 연구에서 외부 정보와의 대조를 통해 더 자세히 조사할 필요가 있다.

대한민국의 PM2.5 중기예보는 '낮음'과 '높음'의 두 단계로 분류된다. Table 3을 통해 대한민국 서울 권역의 수집된 기간에 대한 PM2.5 하루 평균 데이터 연간 통계 정보와 2단계 등급 분류 결과를 제시한다. 2017년부터 2022년까지 '높음' 등급의 사례일 수는 STL 분석과 마찬가지로 감소 추세를 보였으며, 특히 2021년과 2022년에는 이전 연도들에 비해 뚜렷한 감소세를 보였다.

데이터 시각화 및 연도별 통계 정보 분석 결과, 전체 기간 '낮음'과 '높음' 클래스 간 평균 비율은 약 0.81 대 0.19로, 클래스 간 불균형이 존재하는 것을 확인할 수 있다. 클래스 불균형 문제가 존재하는 데이터 세트를 학습하게 될 경우, 예측 모델은 다수 클래스에 편향되도록 학습될 수 있어 소수 클래스에 대한 예측 성능을 크게 떨어뜨릴 수 있다. 따라서 학습 단계에서 클래스 불균형을 완화할 방안이 필요하다.

또한 2021년과 2022년에서는 2017~2020년 대비 급격한 Trend와 '높음' 등급 일수 변화, 그리고 연도별 통곗값 변화를 통해 PM2.5 데이터에 경향성 변화가 있음을 확인하였다. 본 연구에서는 실험을 통해 경향성 변화에 따른 예측 성능 비교와 경향성 변화에도 안정적인 예측 성능을 도출하는 기법을 제시하고자 한다.

3.2 데이터 전처리

1) 데이터 표준화

데이터 전처리 단계에서는 중기예측에 사용될 OBS, FNL, WRF-CMAQ 데이터를 데이터 표준화 과정을 통해 인자 간의 척도(scale) 불일치를 해결하고, 이를 통해 모델의 학습 효율성과 정확도를 향상하고자 하였다. 데이터 표준화는 각 변수의 값들을 평균이 0이고, 표준편차가 1이 되도록 변환하는 작

업을 포함한다. 이 과정은 다음의 수식을 통해 수행되었다:

$$z = \frac{(x-\mu)}{\sigma} \quad (1)$$

Equation (1)에서 x는 변수의 입력값, μ는 해당 변수의 평균값, σ는 해당 변수의 표준편차를 나타낸다. 이렇게 변환된 z 값은 표준화된 x 변수의 값으로, 모든 변수가 같은 척도를 갖게 되어, 모델의 학습 과정에서 특정 변수가 과도하게 영향을 미치는 것을 방지한다.

2) 주성분 분석

본 연구에서 직면한 주요 과제 중 하나는 입력 변수의 크기에 비해 사용 가능한 타임 포인트의 부족이다. 이러한 데이터 구성은 '차원의 저주(Curse of dimensionality)'를 유발할 수 있으며 모든 입력 변수를 사용할 경우, 모델 학습 속도의 저하 및 예측 모델의 성능이 저하될 수 있다 [17]. 중요하지 않은 변수의 포함은 모델의 과적합(overfitting) 가능성을 증가시키며 정확도를 저하할 위험이 있다.

이러한 문제를 해결하기 위하여 본 연구에서는 주성분 분석(Principal component analysis; PCA)을 적용하여 입력 변수의 차원을 축소하였다. PCA는 고차원 데이터의 변동성을 저차원에서 효과적으로 설명하는 방법으로, 이전 연구[18]에서는 입력 데이터에 PCA를 적용한 것과 적용하지 않은 데이터 비교로 적은 수의 데이터 포인트에서 PCA를 통해 미세먼지 예보성능을 올릴 수 있음을 보였다. PCA의 기본 원리는 데이터 내에서 가장 큰 변동성을 설명할 수 있는 변수들의 선형 조합을 통해 차원을 축소하는 것이다. p개의 변수를 포함하는 벡터에서 최대 p개의 주성분을 추출할 수 있다. 각 주성분은 원본 데이터 벡터 $X(n\times p)$와 이의 공분산 행렬 $C(p\times p)$를 이용하여 계산되며 PCA에 적용하여 차원을 줄이면서 중요한 정보를 보존하는 방법으로 활용했다.

3) 타임 윈도우 쉬프팅

타임 윈도우 쉬프팅이란 예측시점 기준 고정된 시간 단위의 입력 타임 포인트 길이를 예측 모델이 학습 가능한 형태로 변환하는 전처리 과정이다. 입력되는 과거 데이터(lag)의 기간은 예측 시점(D-day)부터 과거 6일(D-6)까지 총 7일이다. 또한 미래 예측일(horizon)은 예측 시점을 기준으로 미래 D+3부터 D+6까지 미래 4일의 15시 PM2.5 1포인트를 예측 대상으로 한다. 각 데이터는 하루 4개 포인트씩 계산되며 lag는 4~28의 길이를 갖는다.

4) 클래스 불균형 완화

Table 3에서 확인한 데이터 클래스 불균형 문제를 완화하기 위한 기법으로는 리샘플링(Resampling) 기법과 데이터 증강 기법 등이 알려져 있다. 리샘플링 기법에는 불균형 데이터 세트에서 소수 클래스 샘플을 다수 클래스 샘플 비율에 맞춰 인위적으로 증가시켜 클래스 간의 균형을 맞추는 오버샘플링(Oversampling) 기법과 다수 클래스의 샘플을 소수 클래스 샘플 비율에 맞춰 줄임으로써 데이터 불균형을 해소하는 언더샘플링(Undersampling) 기법이 있다. 본 연구에서는 언더샘플링 기법은 샘플 부족으로 사용하지 않았으며, 데이터 증강 기법은 Goodfellow, Ian et al. [19]의 적대적 생성 신경망을 시계열 데이터에 적용한 연구인 J. Yoon et al. [20]의 증강 기법을 사용하여 소수 클래스 샘플 수를 증가해보고자 했으나 소수 클래스 샘플들의 분포들과 유사한 샘플 들을 생성하지 못했고 그 계산 비용 역시 큰 관계로 본 연구에서는 사용하지 못했다.

본 연구에서 사용하는 클래스 불균형 완화 기법은 오버샘플링 기법 중 랜덤 오버샘플링 기법을 적용해 학습 데이터 세트의 '높음' 클래스 데이터 샘플 수를 증가시켜 클래스 간 비율을 맞췄다.

분류 모델의 경우 국립환경과학원의 2단계 중기예보 기준에 맞춰서 예측 대상인 서울 지역의 PM2.5 값을 0 (낮음), 1 (높음) 등급으로 변환하여 라벨로 사용했으며, '높음' 클래스의 비율을 '낮음' 클래스 비율에 맞춰 해당하는 학습데이터를 증가시켜 학습하였다. 회귀 모델의 경우 표기 된 PM2.5 수치를 동일하게 라벨링 정보로 활용했으며, PM2.5 수치를 2단계 등급 기준에 맞춰 '높음' 클래스에 해당하는 PM2.5 수치 라벨과 학습데이터를 '낮음' 클래스 비율에 맞춰 오버샘플링하여 모델을 학습하였다. 이후 회귀 모델의 PM2.5 예측값과 실제 수치 값을 2단계 등급 분류 구간을 임곗값으로 사용하여 '낮음', '높음'으로 변환하여 평가하였다.

오버샘플링을 적용함으로써 예상되는 문제점으로는 소수 클래스를 지나치게 증가시켜 반복적으로 학습하는 과정에서 과적합이 발생하여 실제 평가 단계에서 성능이 매우 낮게 나오는 상황이 발생할 가능성이 있다. 따라서 오버샘플링 여부에 따른 학습된 예측 모델들의 평가데이터 세트에 대한 예측 결과 확인이 필요하다.

3.3 모델 구조

본 PM2.5 중기 예측 연구에 사용하는 모델 구조는 GoogleNet [21]의 Inception 모듈을 활용하여, 입력 데이터에서 다양한 크기의 시간 패턴을 동시에 추출하고 학습할 수 있는 Multi Source Inception Net(*MSINet*) 구조와 *MSINet*구조에 GRU Layer를 추가한 Multi Source Inception-GRU Net(*MSIGNet*) 구조를 사용한다.

*MSINet*과 *MSIGNet* 모두 Inception 모듈을 사용하는 이유는 초미세먼지 중기예측을 위한 적합한 입력 시퀀스의 길이를 특정하기 어렵기 때문이다. Inception 모듈은 다양한 크기의 합성곱 필터를 활용하여 서로 다른 시간적 범위를 가진 특징을 효과적으로 학습할 수 있다. 각각의 합성곱 필터 크기는 특정 시간적 범위의 특성을 포착하는 데 기여한다. 예를 들어, 작은 필터는 단기 패턴을, 큰 필터는 장기 패턴을 잡아낼 수 있다. 또한 입력 데이터 세트가 개별적으로 입력되는 이유는 각 데이터 세트별 실제 수집 타이밍이 다른 관계로 각각 입력

시계열의 길이가 다르기 때문이다. OBS와 FNL데이터 세트는 개별적인 Inception 모듈을 거치며 연산을 하게 된다.

모든 모델은 PCA를 통해 계산된 OBS, FNL, WRF-CMAQ의 주성분 차원 x_{obs}, x_{fnl}, x_{wc}을 입력으로 받는다. x_{obs} 과 x_{fnl} 은 각각 별도의 Inception 모듈을 통과한다. FNL 데이터는 생산 시점이 OBS 데이터보다 늦어서 타임 포인트 수가 OBS 데이터보다 2개가 적다. 실제 수집 환경을 고려하여 x_{fnl}이 입력되는 Inception 모듈의 feature map은 OBS 주성분 차원의 Inception 모듈 feature map의 크기와 1x1 합성곱 필터를 사용해 같은 크기로 맞춘다. 각각의 Inception 모듈을 거친 feature map 들은 하나로 연결(concatenation)하여 x_{concat}을 만든다. x_{wc}은 예측 모델이 목표하는 D+3~D+6 예측일의 15시 데이터 1타임 포인트만 입력된다. x_{wc}과 x_{concat}은 하나로 연결되어 x_{final}을 만들고, 두 개의 완전 연결 층(Fully Connected Layer)을 거쳐 출력된다.

*MSIGNet*은 기존 PM2.5 단기 예측 연구 [7, 13]에서 제안된 CNN-LSTM 구조를 채용하여 순환 신경망의 한 종류인 GRU (Gate Recurrent Unit)[22] Layer를 추가한 구조이다. *MSIGNet*은 *MSINet*의 x_{obs}, x_{fnl}의 특성을 추출한x_{concat}을 GRU Layer를 거쳐 두 개의 완전 연결 층을 거쳐 최종 출력을 도출한다. *MSINet*모델과 *MSIGNet* 모델들은 손실 함수에 따라 회귀 또는 분류 모델로 구성할 수 있다. 모델들은 손실 함수에 따라 회귀 또는 분류 모델로 구성할 수 있다.

3.4 모델 학습 및 평가

본 연구에 사용된 예측 모델은 분류(classification) 모델과 회귀(regression) 모델로 구분할 수 있다. 분류 모델은 미래 예측일 15시의 초미세먼지 예측등급의 확률을 나타내며 회귀 모델은 미래 예측일 15시의 초미세먼지 농도 예측값을 나타낸다. 모든 모델의 최적화 알고리즘으로는 SGD (Stochastic Gradient Descent)를 사용했으며, 목적 함수로는 분류 모델은 이진 교차 엔트로피(Binary Cross Entropy; BCE) 오차함수를 회귀 모델의 경우 평균제곱오차(Mean Squared Error; MSE) 함수를 사용했다. 모델 구조와 분류, 회귀에 따라 총 4가지 모델로 구분 지을 수 있다. 각 모델명은 *MSI*-cls, *MSI*-reg, *MSIG*-cls, *MSIG*-reg로 명명한다.

학습된 모델 성능 평가는 탐지율 (Probability of Detection; POD), 오탐율 (False Alarm Ratio; FAR) 그리고 이 둘의 조화평균으로 구해지는 F1 Score 지표를 사용하였다. POD는 Equation (2)을 따르며 실제 '높음'인 예측 결과를 얼마나 잘 감지하는지를 나타내며 100(%)에 가까울수록 성능이 좋음을 나타내며 FAR는 Equation (3)을 따르며 모델이 '높음'인 샘플을 '낮음'으로 잘못 예측한 비율을 나타내며 수치가 낮을수록 성능이 높다고 볼 수 있다. F1 Score는 Equation (4)와 같이 POD와 FAR의 조화평균이며 1에 가까울수록 높은 성능을 보인다.

모든 4가지 모델들은 학습데이터에 대하여 3-fold 교차검증(Cross Validation; CV)으로 2개의 학습데이터 fold와 1개의 검증데이터 fold로 성능 평가 및 하이퍼파라미터 튜닝을 하였다.

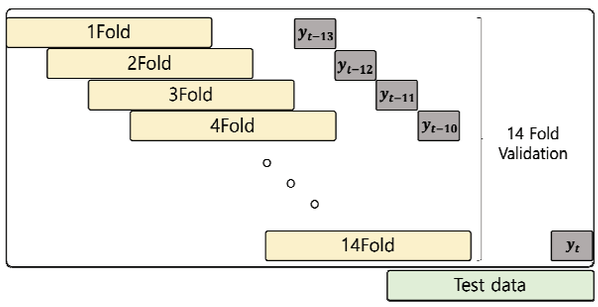

Fig. 3. Forward Chaining based Ensemble Method

$$POD = \frac{True\,Positive}{True\,Positive + False\,Negaitive} \tag{2}$$

$$FAR = \frac{False\,Positive}{False\,Positive + True\,Positive} \tag{3}$$

$$F1\,Score = 2 \times \frac{POD \times (1 - FAR)}{POD + (1 - FAR)} \tag{4}$$

3.5 Forward Chaining based Ensemble

Forward Chaining based Ensemble(*FCE*) 기법은 *MSINet*과 *MSIGNet*을 포함하는 중기 예측 모델로 구성되어 있으며, Fig. 3과 같이 Validation과 Test 두 단계로 이루어져 있다. Validation 단계는 평가데이터에 대해 예측 시점 기준 가장 최근 PM2.5 라벨을 생성 가능한 시점부터 14일 전 과거 데이터까지 검증 데이터로 사용한다. 과거 14일 데이터를 검증 데이터로 활용하며 예측 시점 방향으로 하루 단위 윈도우 쉬프팅(Window Shifting)을 수행하며 Fold를 생성한다. 4가지 모델들에 대한 14 Fold에 대한 F1 Score 성능을 평가하여 모델 평가 단계에서 앙상블 기법에 포함할 상위 3개의 모델을 선정한다. Test 단계에서는 예측 시점 데이터를 입력으로 활용하며 선정된 상위 3개 모델의 결과를 majority voting 방식으로 앙상블 결과를 도출한다. *FCE* 기법은 예측 시점에 가까운 데이터에 대해 예측 성능이 높은 모델들을 선정하고, 해당 모델들의 예측 결과를 앙상블을 통해 도출하므로, 안정적인 예측 결과를 도출할 수 있을 것으로 예상된다.

4. 실험 결과

4.1 실험 환경 구축

본 논문에서 사용한 미세먼지 중기 예측 모델의 훈련 환경은 Table 4에 요약되어 있다. PM2.5 데이터 분석, 모델의 학습 및 평가는 Python 3에서 구현 하였으며, 분류와 회귀 모델은 'PyTorch' 라이브러리를 사용하여 개발 후 그래픽카드 RTX2080Ti 환경에서 훈련하였다.

학습에 사용된 하이퍼파라미터로는 과거 입력 데이터의 길이(Lag), 오버샘플링 여부, 모델 계층 수 등이 있었다. 교차검증을 통해 검증 데이터 세트 F1 Score 성능이 가장 높은 모델을 2021년 평가데이터 세트에서 평가하였다. 주성분 분석 결과 모델 학습에 사용된 차원은 64, 128, 256, 512차원 중 기존 데이터 분산의 90% 이상을 설명하는 가장 작은 차원을 선택했다. 이에 대하여 OBS 데이터에 대해서는 256차원(92.38%)을, FNL 데이터는 512차원(93.12%), WRF-CMAQ 데이터는 512차원(91.19%)을 사용한다.

또한 오버샘플링 여부에 따른 2021년 평가데이터 예측 결과 확인 결과 모든 예측일에서 오버샘플링을 적용한 모델의 성능이 높아 본 실험에서 학습 데이터 세트에 클래스 불균형을 완화하기 위한 전처리 기법으로 오버샘플링 기법이 유효한 것으로 확인되었다.

본 연구에서 사용되는 학습데이터 기간은 2017년 1월~2020년 12월까지이며 평가데이터 기간은 2021년 1월~12월과 2022년 1월~12월이다.

4.2 분류 및 회귀 모델 성능 평가

모델 구조로 제시된 *MSINet*과 *MSIGNet* 구조의 분류 및 회귀 모델의 성능 평가를 진행하였다. 성능 평가 지표는 POD, FAR, F1 Score 지표를 사용하며, 이들은 각각 D+3~D+6 예측일에 대해 계산되었다. 평가 결과는 예측일별 성능 지표들을 모두 포함하여 Fig. 4에 상세히 정리하였다. 실험 결과 *MSI*-reg과 *MSIG*-reg와 같은 회귀 모델은 다른 분류 모델들보다 대부분의 예측일에서 상대적으로 낮은 FAR를 보여, 회귀 모델이 '높음'일에 잘못된 예측을 하는 경우가 상대적으로 적은 것을 알 수 있다. 반면 *MSI*-cls와 *MSIG*-cls과 같은 분류 모델들은 다른 회귀 모델들보다 모든 예측일에서 상대적으로 높은 POD, FAR를 보였다. 모델 구조별 성능을 비교한 결과 *MSINet*은 회귀 모델 성능이 높았지만, *MSIGNet*는 분류 모델 성능이 높은 결과를 보여 최적의 단일 중기 예측 모델을 결정하기엔 다소 어려움이 존재한다.

4.3 모델 성능 비교 평가

본 연구에서 최적의 단일 중기 예측 모델을 도출하는데 존재하는 어려움과 2021년, 2022년의 경향성 변화에 더 안정적인 결과를 도출하기 위한 기법으로 제시된 *FCE* 기법과 기존 연구의 모델 성능을 비교한 결과를 Fig. 5, 6에 정리하였다.

Table 4. Experiment Environment Specification

experiment tool	version
OS	Ubuntu 18.04
GPU	RTX 2080ti
Python	3.7
PyTorch	1.7.2
sklearn	1.0.2

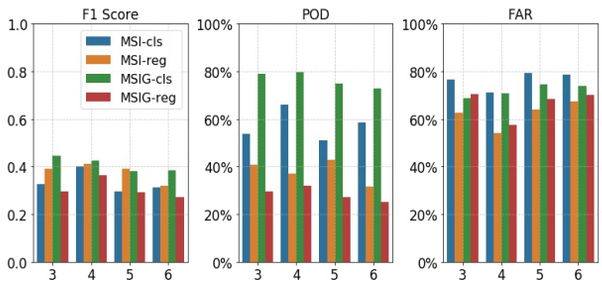

Fig. 4. Result of MSINet, MSIGNet Regression & Classification Models on 2021 Test Set

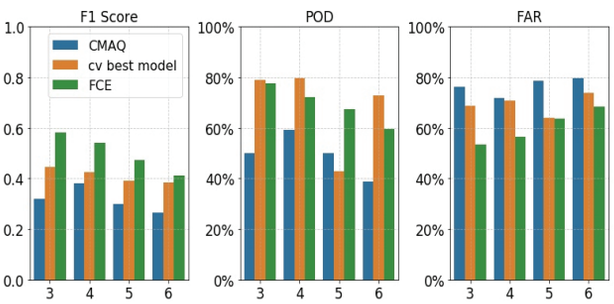

Fig. 5. Compare Result on 2021 Test Set

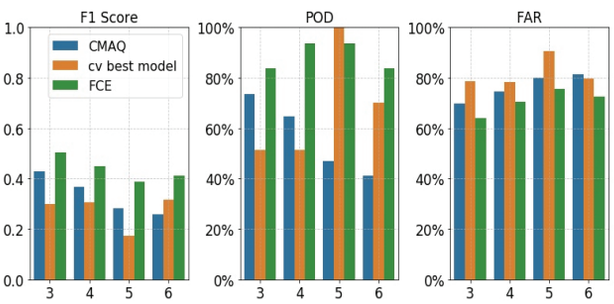

Fig. 6. Compare Result on 2022 Test Set

비교 분석에는 기존 중기 예측에 활용되는 CMAQ 모델과 4.3절에서 예측일별 가장 높은 F1 Score를 보인 모델을 'cv best model'로 선정하여 비교하였다. 실험은 2021년과 2022년의 평가데이터를 바탕으로 진행되었으며, 성능 평가 지표로는 4.2 절에서 언급된 POD, FAR, F1 Score를 사용하였다.

Fig. 5 결과에 따르면 2021년 평가데이터를 바탕으로 한 결과에서, 제안된 *FCE* 기법은 F1 Score 기준 모든 예측일에서 기존 CMAQ 모델 및 cv best model보다 높은 성능을 보였다. 다만, 제안한 기법의 POD는 일부 예측일에서 cv best model보다는 낮았지만, FAR는 모든 예측일에서 모든 모델 중 가장 낮았다. CMAQ 모델 성능 대비 전체 예측일 평균 약 58.92% 성능이 개선된 것을 확인할 수 있었으며 cv best model 성능 대비 약 21.13% 성능이 개선된 것을 확인할 수 있었다.

Fig. 6 결과에 따르면 2022년 평가데이터 결과에서도 *FCE* 기법은 F1 Score 기준 다른 비교 모델 대비 모든 예측일에서 높은 성능을 보였다. 2022년 데이터 기준 제안 기법 성능은

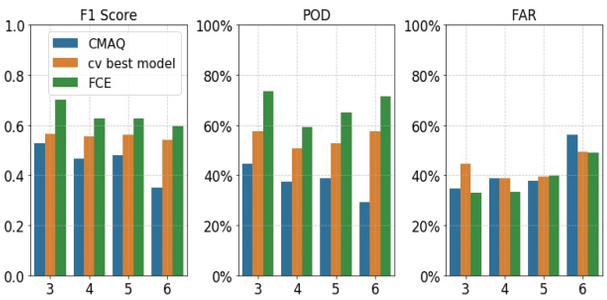

Fig. 7. Compare Result on 2019 Test Set

CMAQ 모델 성능 대비 D+3~D+6 예측일 평균 약 34.79% 성능이 개선된 것을 확인할 수 있었으며 'cv best model' 성능 대비 약 67.41% 성능이 개선된 것을 확인할 수 있었다.

2021년 대비 2022년의 실험 결과를 비교하면, *FCE* 기법은 Fig. 6의 2022년 평가데이터에서 Fig. 5의 2021년 평가데이터 결과 대비 POD 성능의 향상이 있었지만, FAR 증가로 F1 Score 성능의 하락이 관찰되었다. 이와 대조적으로, CMAQ 모델은 2021년 평가데이터 대비 2022년 평가 결과 성능이 소폭 향상되었고, cv best model은 2022년 평가 결과 성능이 하락하였다.

4.4 Ablation Study

앞서 경향성 변화가 존재하는 2021년과 2022년 평가데이터 외에 경향성 변화가 없는 기간에 대해서도 본 연구진들이 제안하는 *FCE* 기법의 성능이 유효한지 판단하기 위해 학습 및 평가 기간을 변경하여 실험을 진행했다. 실험에 사용된 데이터 학습 기간은 2017년 1월~2018년 12월까지 학습하였고, 2019년에 대하여 모델별 성능 평가를 진행했다. Fig. 7은 2019년 평가데이터 세트에 대한 성능 비교 결과이다.

2019년 평가데이터 기반 실험 결과, 각 예측일 별 제안하는 기법과 비교 모델 모두 F1 Score 성능이 2021년, 2022년에 비해 상승하고 낮은 FAR를 보이는 것을 확인할 수 있었다. 제안하는 *FCE* 기법의 경우 F1 Score 및 POD 성능은 비교 모델 가운데 가장 높은 성능을 보였다. FAR 성능 또한 다른 모델 대비 D+5 예측일을 제외한 모든 예측일에서 가장 낮았다. 2019년 평가데이터 기준 제안 모델 F1 Score 기준 성능은 CMAQ 모델 성능 대비 D+3~D+6 예측일 평균 약 42.16% 성능이 개선된 것을 확인할 수 있었으며 cv best model 성능 대비 약 26.27% 성능 개선을 확인할 수 있었다.

실험 결과, 본 연구에서 제안한 *FCE* 기법은 2019년, 2021년, 2022년 모든 평가데이터 기간에서 다른 모델들과 비교하여 가장 높은 F1 Score와 가장 낮은 FAR를 보였다. 2021년 데이터에 대한 평가에서는 다른 모델들과 비교하여 모든 지표에서 더 높은 성능을 보였으며, 2022년 데이터 세트에서 F1 Score가 소폭 감소하지만 높은 성능을 유지했다. 또한 데이터의 경향성 변화가 없는 2019년 평가데이터에서도 본 연구에서 제안한 기법이 가장 우수한 성능을 보여 안정적인 성능을 확인할 수 있었다. 그러나 예측일이 멀어질수록 성능이 감소하는 경향이 관찰되어 향후 연구에서는 앙상블 기법을 개선하여 POD를 향상하고 FAR를 감소시켜 전반적인 성능을 더욱 개선할 필요가 있다.

5. 결 론

본 연구에서 제안한 *FCE* 기법은 학습데이터와 평가데이터 간 경향성 변화가 발생하는 상황에서도 안정적인 성능을 보였으며, 경향성 변화가 없는 기간에서도 가장 안정적인 성능을 보였다. 모든 예측일에서 본 연구에서 제안하는 기법은 다른 비교 모델과 전통적인 방법들을 넘어서는 높은 F1 Score와 가장 낮은 FAR을 기록하였고, 기존 CMAQ 모델 대비 2019년은 D+3~D+6 예측일 평균 약 42.16%, 2021년은 D+3~D+6 예측일 평균 약 58.92%, 2022년은 D+3~D+6 예측일 평균 약 34.79% 성능 개선을 보여 제안한 기법의 안정성과 효율성을 입증하였다. 하지만 2022년에는 제안한 모델의 FAR이 증가하는 경향을 보였다. 이는 *FCE* 기법의 POD 향상 및 FAR을 줄이기 위한 추가적인 연구가 필요함을 시사한다.

종합적으로, 본 연구에서 제안된 모델은 인공지능 기반의 뛰어난 예측 능력을 통해 대한민국 PM2.5 중기 예측 분야에 기존 예측체계 성능을 초과하는 성능을 보였다. 또한 다양한 예측 모델의 앙상블을 통해 정확하고 안정적인 예측이 가능하게 되었으며, 대한민국 중기 대기질 모니터링과 예보에 크게 기여할 것으로 기대된다.

References

[1] WHO Health Organization, "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide" [Internet], https://www.who.int/publications/i/item/9789240034228

[2] M. Bae, Y.-H. Kang, E. Kim, S. Kim, and S. Kim, "A multifaceted approach to explain short- and long-term PM2.5 concentration changes in Northeast Asia in the month of January during 2016-2021," *Science of The Total Environment*, Vol.880, pp.163309, 2023.

[3] S. Lee, et al., "Analysis of long-range transboundary transport (LRTT) effect on Korean aerosol pollution during the KORUS-AQ campaign," Atmospheric Environment, Vol.204, pp.53-67, 2019.

[4] Y. Cha, C.-K. Song, K. Jeon, and S.-M. Yi, "Factors affecting recent PM2.5 concentrations in China and South Korea from 2016 to 2020," *Science of The Total Environment*, Vol.881, pp.163524, 2023.

[5] Ministry of Environment, Artificial intelligence prediction improves ozone forecast accuracy [Internet], https://www.me.go.kr/home/web/board/read.do?menuId=10525&boardMasterId=1&boardCategoryId=39&boardId=1372380

[6] Ministry of Environment, "Expansion of Weekly Forecast Regions and Strengthening Response to the 2nd Fine Dust Seasonal Management," Air Quality Integrated Forecast Center [Internet], www.me.go.kr
[7] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla, "Linear and nonlinear modeling approaches for urban air quality prediction," *Science of The Total Environment*, Vol.426, pp.244-255, 2012.
[8] J. Zhao, F. Deng, Y. Cai, and J. Chen, "Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction," *Chemosphere*, Vol.220, pp.486-492, 2019.
[9] H. S. Kim, K. M. Han, J. Yu, J. Kim, K. Kim, and H. Kim, "Development of a CNN+LSTM Hybrid Neural Network for Daily PM2.5 Prediction," *Atmosphere*, Vol.13, No.12, Art. no.12, 2022.
[10] K. Lee et al., "Development of Korean Air Quality Prediction System version 1 (KAQPS v1) with focuses on practical issues," *Geoscientific Model Development*, Vol. 13, No.3, pp.1055-1073, 2020.
[11] U. Shin et al., "Predictability of PM2.5 in Seoul based on atmospheric blocking forecasts using the NCEP global forecast system," *Atmospheric Environment*, Vol.246, pp.118141, 2021.
[12] A. Sayeed, Y. Lops, Y. Choi, J. Jung, and A. K. Salman, "Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks," *Atmospheric Environment*, Vol.253, pp.118376, 2021.
[13] A. Azid et al., "Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: a Case Study in Malaysia," *Water Air Soil Pollut*, Vol.225, No.8, pp.2063, 2014.
[14] C.-J. Huang and P.-H. Kuo, "A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities," *Sensors*, Vol.18, No.7, Art. No.7, 2018.
[15] H. Chang-Hoi et al., "Development of a PM2.5 prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, Republic of Korea," *Atmospheric Environment*, Vol.245, pp.118021, 2021.
[16] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition," *Journal of Official Statistics*, Vol.6, No.1, pp.3-73, 1990.
[17] R. Bellman, "Dynamic programming," *Science*, Vol.153, No.3731, pp.34-37, 1966.
[18] S. W. Choi and B. H. S. Kim, "Applying PCA to Deep Learning Forecasting Models for Predicting PM2.5," *Sustainability*, Vol.13, No.7, Art. No.7, 2021.
[19] I. J. Goodfellow, et al., "Generative Adversarial Networks," *Neural Information Processing Systems*, 27, 2014.
[20] J. Yoon et al., "Time-series generative adversarial networks," *Neural Information Processing Systems*, 32, 2019.
[21] C. Szegedy et al., "Going deeper with convolutions," *Computer Vision and Pattern Recognition*, pp.1-9, 2015.
[22] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *Empirical Methods in Natural Language Processing*, pp.1724-1734, 2014.

민 동 준

https://orcid.org/0000-0000-3720-5950
e-mail : mindongjun@korea.ac.kr
2018년 한양대학교 교통물류공학과(학사)
2020년 ~ 현 재 고려대학교 정보보호대학원 정보보호학과 석사과정
관심분야 : 시계열 예측, 이상 탐지, 연합 학습

김 혜 림

https://orcid.org/0000-0002-0704-751X
e-mail : hyerimkim@ewha.ac.kr
2006년 연세대학교 대기과학과(학사)
2010년 연세대학교 대기과학과(석사)
2017년 연세대학교 대기과학과(박사)
2012년 ~ 2014년 극지연구소 연구원
2017년 ~ 2022년 한양대학교(ERICA) 해양대기과학연구소 PostDoc.
2017년 ~ 2022년 한양대학교(ERICA) 해양대기과학연구소 연구조교수
2022년 ~ 2023년 고려대학교 정보보호연구원 연구교수
2023년 ~ 현 재 이화여자대학교 사범대학 연구교수
관심분야 : 기후 변화와 태평양 기후 변동성, 기후모델링, 해양 대기 상호작용, 미래기후 예측성 향상을 위한 인공지능 활용

이 상 근

https://orcid.org/0000-0001-8415-6368
e-mail : sangkyun@korea.ac.kr
2003년 서울대학교 컴퓨터공학(학사)
2005년 서울대학교 전기컴퓨터공학(석사)
2011년 미국 Wisconsin-Madison 대학 컴퓨터과학(석 · 박사)
2011년 ~ 2017년 독일 TU Dortmund 대학 PostDoc Fellow, Project Leader
2017년 ~ 2020년 한양대학교(ERICA) 소프트웨어학과 조교수
2020년 ~ 2021년 고려대학교 정보보호대학원 조교수
2022년 ~ 현 재 고려대학교 정보보호대학원 부교수
관심분야 : 신뢰 가능한 인공지능, 인공지능 모델 복제 공격/방어, 설명 가능한 인공지능, 언어모델 보안, 인공지능 기반 코드 취약점 탐지