Journal of Next-generation Convergence Technology Association ISSN 2508-8270 EISSN : 2765-4133
Vol.7, No.9, pp. 1-000, 2023 https://doi.org/10.33097/JNCTA.2023.07.09.000

서울시 대기 중 미세먼지 농도 예측을 위한 3D 심층 합성곱 신경망 앙상블 예측

이준민[1], 김경태[2], 최재영[3*]
[1]한국외국어대학교 컴퓨터공학과 석사과정, [2]한국외국어대학교 컴퓨터공학과 박사과정,
[3]한국외국어대학교 컴퓨터공학과 부교수

Ensemble Prediction of 3D Deep Convolutional Neural Networks for the Prediction of PM2.5 Level for the Seoul, Republic of Korea

Joon Min Lee[1], Kyeong Tae Kim[2], Jae Young Choi[3*]
[1]Master Student, Division of Computer Engineering, Hankuk University of Foreign Studies
[2]Ph.D Student, Division of Computer Engineering, Hankuk University of Foreign Studies
[3]Associate Professor, Division of Computer Engineering, Hankuk University of Foreign Studies

요 약 대기 중 미세먼지 (Particular Matter)의 높은 농도에 노출되는 것은 인간 건강에 심각한 영향을 미친다. 많은 연구가 미세먼지가 특히 높은 농도에서 인간에게 매우 유독하다는 것을 보여주었다. 이러한 이유로 여러 국가가 미세먼지 농도를 규제하기 위해 상당한 노력을 기울이고 있다. 미세먼지 피해를 줄이기 위해 PM 농도 기반의 조기 경보 시스템이 필요하다. 본 논문에서 대한민국 서울의 PM2.5 농도를 예측하기 위해 3D CNN을 활용한 새로운 앙상블 예측 접근법을 제안한다. 이 방법은 매 6시간 최대 2일 동안 관측된 데이터에 대한 시공간 데이터의 특징을 추출하고 결합하기 위해 공간 정보뿐만 아니라 시간 정보도 활용할 수 있는 3D CNN의 앙상블 예측을 사용한다. 결합한 특징은 공유 계층을 통해 의미 있는 특징 정보를 추출하고 개별 계층을 통해 시간별로 예측된 농도를 반환한다. 제안된 방법은 서울 메트로폴리탄 지역의 PM2.5 및 기상 데이터를 사용하여 2015년 1월 1일부터 2021년 2월 28일까지 예측을 수행하였다. 제안된 방법은 PM 2.5 관측 (실제 값), CMAQ 예측 및 ConvLSTM에 대해 종합적으로 모델을 평가하였다. 예측 성능 측면에서, 제안된 방법은 널리 사용되는 예측 모델과 비교하여 7.06%의 향상된 예측 정확도, 4.28%의 향상된 고농도 탐지율, 오경보율을 11.77%만큼 개선하였다.

주제어 : PM2.5, 대기질, 3D 합성곱 신경망(3D-CNN), 앙상블 모델, 인공지능, 딥러닝, 예측

Abstract Exposure to high concentrations of airborne particulate matter (Particular Matter) has serious effects on human health. Much research work showed that particulate matter is very toxic to humans, especially in high concentrations. For this reason, many countries make considerable efforts to regulate PM concentrations. In order to implement PM measures to reduce damage, an early warning system based on PM concentration level is essentially required. In this paper, we propose a novel 3D CNN ensemble prediction approach for forecasting PM2.5 concentrations in Seoul, Republic of Korea, which is observed every 6 hours

* This research was supported by Hankuk University of Foreign Studies Research Fund.
This research was supported by Korea Institute of Marine Science & Technology Promotion(KIMST) funded by the Ministry of Oceans and Fisheries(20220180).
This work was also supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1092322).

* Corresponding Author : Jae Young Choi (jychoi@hufs.ac.kr)

Received August 18, 2023 Revised September 13 2023
Accepted September 14 2023 Published September 28 2023

for up to two days. The proposed method uses an ensemble of 3D deep CNNs that can utilize not only spatial information but also temporal information to extract and combine features for spatiotemporal data corresponding to each input. The combined features aim at extracting meaningful feature information through the shared layer and return the time-series predicted concentration for each period through the individual layer. The proposed method performs prediciton using PM2.5 and meteorological data of Seoul metropolitan area, ranging from 2015-01-01 to 2021-02-28. The proposed method was comprehensively evaluated for PM2.5 observation (Ground-truth), CMAQ forecast, and ConvLSTM. In terms of prediction performance, the proposed method showed improvement in performances with an increase of 7.06% in accuracy, an increase of 4.28% in probability of detection, and a decrease of 11.77% in false alram rate, compared to the widely used forecast model.

Key Words : PM2.5, Air quality, 3D Convolutional Neural Network(3D-CNN), Ensemble model, Artificial intelligence (AI), Deep Learning, Prediction

Ⅰ. 서론

한국의 경제적 빠른 성장과 자동차 및 산업의 증가로 인해 대기 오염 문제는 매우 심각한 문제가 되고 있다. 최근에는 미세먼지(Particular Matter)의 높은 수준이 글로벌 문제로 부상하였다. PM10과 PM2.5는[1] 각각 10㎛와 2.5㎛보다 작은 미세 입자로, 많은 역학 연구들이 PM이 높은 농도에서 인간에게 매우 유독하다는 것을 보여주었다[2]. 고농도의 단기간 미세먼지 노출은 사망률을 증가시키는 요인이다[3]. 장기간 노출은 일시적인 심혈관 영향, 호흡기 질환 및 폐암을 유발할 수 있다[4-6]. 이러한 문제들을 고려하여 세계 보건 기구(WHO)는 PM2.5를 1급 발암물질로 고려하고 여러 국가에서 인간이 미세먼지 노출되는 수치를 평가하기 위해 PM10과 PM2.5 모니터링을 개선해야 한다고 발표했다[7].



Fig. 1. Description of the counties with the PM2.5 exposures in 2022.

Fig. 1에서는 한국의 PM2.5 노출에 대한 EPI(Environmental Performance Index), 환경성과지수가 180개 국가 중 45위임을 보여주며[8] 한국이 PM2.5에 대한 노출이 심하다는 것을 의미한다.

대부분 국가는 공중보건과 환경에 해로운 오염물질에 대한 국가 대기질 기준을 가지고 있다. 한국에서는 PM2.5 농도를 다음과 같이 네 가지 수준으로 분류한다. 낮음 (PM2.5 ≤ 15), 보통 (16 ≤ PM2.5 ≤ 35), 높음 (36 ≤ PM2.5 ≤ 75) 및 매우 높음 (76 ≤ PM2.5)[9]. 한국에서 '매우 높음' 수준의 PM2.5가 발생하거나 지속될 경우, 정부는 즉시 배출량을 줄이고 오염으로부터 국민을 보호하기 위해 "고농도 미세먼지 비상저감 조치"를 시행한다[10]. 이 조치는 대기 오염의 원인인 불법 배출, 차량 배기가스, 불법 소각을 규탄하고 건설 현장의 운영 상태를 점검한다. 또한 이 조치는 방출시설의 운영 시간을 변경하고 운영률을 조절하며, 건설 현장에서 시공 시간의 변경 또는 조정과 같은 완화 조치를 시행한다. 또한 주변 지역의 물 청정을 확대하여 분산 먼지의 생성을 억제한다. 이러한 조치가 취해지기 전에 미세먼지를 예측하고 대비하면 건강을 유지하고 삶의 질을 향상시킬 수 있다. 따라서 미세먼지를 예측하는데 상당한 연구들이 수행되고 있다[11-13].

한국 환경부 하에 있는 국립환경과학원(NIER)은 2013년 8월 13일 미국 환경보호국(US EPA)의 CMAQ Community Multiscale Air Quality) 모델을 사용하여

대기질 예측을 시작했다[14]. 이러한 작업은 상당한 인력 자원이 필요하여 단기간 내에 수행하기 어렵다. 또한 비선형 패턴을 인식하지 못하며, 시간 경과에 따른 대기 오염 예측이 어렵다. 최근 하드웨어와 빅데이터 관리 기술의 빠른 발전으로, 기계 학습 방법(특히 딥러닝)이 대기질 예측에 매우 인기를 끌고 있다. 대부분 시계열 예측에 사용되며, 신뢰성이 입증되었다. 전통적인 기계 학습 형태인 인공 신경망(ANN)은 비선형 매핑 문제를 잘 처리할 수 있으며, 특정 모델을 사전에 지정할 필요가 없으므로 미세먼지 농도 예측에 대해서도 다른 기본적인 기계 학습 방법보다 더 나은 결과를 제공한다[15-17]. ANN(Artificial Neural Network) 기반 모델의 단점은 메모리 셀이 부족하여 시계열 데이터 입력 데이터 간의 연결을 찾기 어렵다는 것이다[18].

시계열 데이터에 대해 독특한 재귀 구조를 사용하는 순환 신경망 모델(RNN)은 시계열 예측에 적합하며, 순차적인 데이터 패턴을 활용하는 지도 학습 모델이다. 그러나 RNN(Recurrent Neural Network)은 시계열 데이터를 처리하는 동안 장기간 의존성을 처리하는 데 어려움이 있고, 기울기 소실(Gradient Vanishing) 문제에 취약하다[19]. LSTM(Long Short-Term Memory)은 시계열 예측을 수행하기 위해 RNN보다 더 나은 모델로 RNN과 비교하여 LSTM은 장기간의 시계열 데이터를 처리하여 예측할 수 있다. 그러나 ANN, RNN 및 LSTM은 일반적으로 1차원 시계열 데이터 처리를 위해 설계되었으며, 시간에 따라 변하는 공간 데이터를 처리하기에 적합하지 않다. 2D 또는 3D 이미지와 같은 공간 데이터를 처리하기 위해 합성곱 신경망(CNN)이 사용되며 이미지 처리 관련 딥러닝 분야에서 주로 사용된다. CNN(Convolutional Neural Network)은 공간적 특징 정보를 추출하여 데이터의 패턴과 지역적인 공간 의존성을 포착하여 실질적인 패턴 인식, 분류 등 연구 성과를 보여주었다. 그러나 전형적인 CNN 구조를 사용하여 시계열 이미지 데이터와 시간적 정보를 처리하는 것은 어렵다. 이러한 문제를 해결하기 위해 CNN-RNN 및 CNN-LSTM과 같은 결합 모델이 제안되었으며, 이들은 미세먼지 예측에 널리 사용되고 있다[13,20]. 그러나 이러한 결합 모델은 모델들이 순차적으로 작동하기 때문에 시공간적 특징을 동시에 분석하는 데 어려움이 있다.

한 모델로 시계열 이미지 데이터를 처리하는 방법으로는 3D-CNN과 ConvLSTM이 있다. 이 두 모델은 모두 시계열을 예측할 수 있지만, 학습 방법 측면에서 차이가 있다. 3D-CNN은 기존 비디오 데이터에서 움직임과 같은 미세한 움직임을 포착하기 위해 설계되었다. 기본 구조로 CNN을 가지며, 특정 위치의 인접한 픽셀들이 서로 높은 상관관계를 가지고 있고, 그렇지 않은 경우는 낮아지는 특성을 기반으로 설계되었다. 이러한 특성 때문에 시계열 데이터에서 공간 정보를 가진 정보를 찾는데 유용하다. ConvLSTM은 LSTM 모델 내에서 CNN을 적용한 모델로 합성곱이 포함되어 기존 LSTM에 비해 가중치 수를 크게 줄였다. LSTM에서는 1차원 데이터만 다룰 수 있는 구조였지만, 이를 확장하여 2차원 시계열 데이터를 입력으로 사용할 수 있게 개선되었다. 하지만 ConvLSTM은 불확실성으로 인해 모델 자체에서 많은 가정들이 있기 때문에, 다양한 데이터가 평균적으로 혼합되어 흐릿한 이미지나 특성 정보가 생성되는 단점이 있습니다. 이를 처리하기 위해 대안적인 아키텍처나 정규화 기법을 탐구해서 사용하는 연구들이 진행되고 있다[21].

미세먼지 예측 분야에서, 3D CNN을 사용하여 PM2.5를 예측하는 연구는 드물다. 본 논문에서는 한국 서울의 PM2.5 농도를 예측하기 위해 3D CNN을 활용한 새로운 앙상블 예측 방법을 제안한다. 이 방법은 최대 2일 동안 매 6시간 관측되는 데이터를 사용한다. 제안된 방법은 Multi-output Ensemble 3D Convolutional Neural Network (ME3CNN)이라고 명명되었으며, 여러 개의 CNN을 사용하여 새로운 앙상블 예측 방법을 제안한다. 이 방법은 공간 정보뿐만 아니라 시간 정보도 동시에 활용할 수 있는 여러 개의 3D CNN을 사용하여 각각의 시공간 데이터로부터 특징을 추출하고 결합할 수 있다. 3D CNN 앙상블을 사용하여 추출된 특징은 1차원 시계열 데이터와 통합되며, 공유 레이어(Shared Layer)를 통해 의미 있는 특징 정보를 추출한다. 이후 개별 레이어를 디자인하여 각 시간대에 대한 다중 출력 예측이 가능하도록 예측 성능을 극대화한다. 본 방법은 2015년 1월 1일부터 2021년 2월 28일까지 서울의 PM2.5 및 기상 데이터를 사용하여 예측을 수행하였다. 또한 PM2.5 관측값(Ground-truth), CMAQ 예측 및 ConvLSTM과 종합적

으로 평가되었다. 예측 성능 측면에서, ConvLSTM과 비교하여 제안된 방법은 ACC를 7.06% 향상, POD를 11.77% 향상, FAR를 10.82% 감소시켰다.

II. 데이터셋

2.1 관측 데이터

북동부 지역에서 관측된 대기질 및 기상 변수들을 사용하였으며, 해당 변수들은 다음과 같다. 시간별 대기질 데이터는 PM10, PM2.5, O3, NO2, CO, SO2 농도와 풍속 및 풍향, 상대습도, 기압, 대기 및 이슬점 온도, 복사 및 누적 강수량으로 구성된다. 한국의 대기질 데이터는 AirKorea 웹사이트(www.airkorea.or.kr)로부터, 중국의 대기질 데이터는 PM25.in 웹사이트(pm25.in, 2022년 6월 이후 운영 중지)로부터 얻었다. 중국과 한국의 기상 데이터는 한국 기상청 API 허브 서비스 웹사이트(https://apihub.kma.go.kr/)로부터 얻었다. 대기질 및 기상 관측소의 수는 연도별로 증가하고 있다. 2023년 기준, 한국에는 523개의 대기질 관측소와 103개의 기상 관측소가 있다. 관측된 대기질 및 기상 데이터는 크리깅(Kriging) 보간기법을 사용하여 모델링 지역(중국과 대한민국을 포함)에서의 2차원 분포 지도로 변환하였다. 이러한 지도들은 중국과 대한민국 사이의 PM2.5 이동에 대한 공간적 관계를 정의하는 데 사용된다. 크리깅 데이터에 대한 설명은 2.3절에서 자세하게 다룬다.

2.2 CMAQ, WRF 예측 데이터

한국 대기질 예보 시스템(Korean Air Quality Forecasting System, KAQFS)은 지표면, 850hPa 및 925hPa에서의 PM2.5 농도와 기상 변수에 대한 예측 데이터를 구축하는 데 사용된다. 이 시스템은 WRF(Weather Research and Forecast) 모델과 동북아시아 상공에서의 화학 수송 시뮬레이션을 위해 CMAQ 모델로 구성되어 있다. 기상 모델로는 WRF 모델의 버전 3.4.1을 사용하였으며[22-23], WRF의 수직 레이어 수는 35개다. WRF의 이니셜라이징(initializing)는 National Center for Environmental Prediction (NCEP) Final Analyses (FNL) 데이터를 사용하였다. 화학 수송 모델로는 미국 환경보호청 모델-3 CMAQv4.7.1 (Community Multiscale Air Quality version 4.7.1)을 사용하여 PM2.5 농도를 예측하였다[24-25]. CMAQ는 12개의 수직 레이어로 구성되어 있으며, 이는 35개의 WRF 레이어에서 보간되었습니다. 첫 번째 표면 레이어의 높이는 36m로, 이는 WRF 모델의 첫 번째 레이어와 동일하다.

2.3 PM2.5 농도와 기상 변수의 공간 분포

제안된 모델에서는 넓은 범위의 기상 이동, 변화의 영향을 반영하기 위해 중국과 한국의 대기질 및 기상 관측소에서 얻은 PM2.5 농도와 기상 변수를 크리깅(Kriging) 방법을 사용하여 이차원 분포 데이터로 변환하였습니다. 크리깅 보간법(Kriging Interpolation)[26-27]은 지정된 관측점에서의 지점 데이터를 결합하여 이차원 필드 데이터를 생성하는 지리 통계적 기법이다. 이 방법은 보간을 위해 주변 측정 값들을 선형적으로 결합하고 이를 사용하여 값을 추정한다. IDW 보간법(역가중 거리 보간법, Inverse Distance Weighting interpolation)[28]과는 달리, 크리깅 보간법은 주변에 있는 많은 측정 값들을 통계적으로 분석하여 값을 추정한다. 즉, 값 추정 시 크리깅 보간법은 측정값으로부터의 거리뿐만 아니라 각 이웃 측정값 사이의 상관관계 강도를 반영한다. Fig. 2는 크리깅 보간법의 예시를 보여준다.

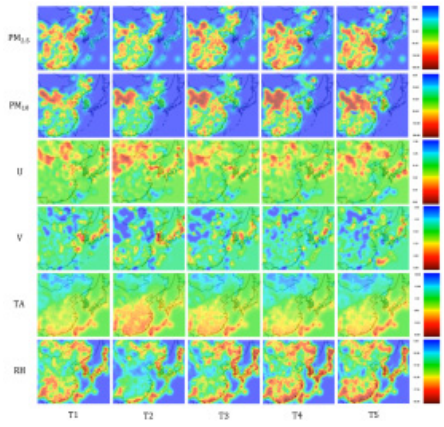


Fig. 2. Observed data samples for T_1 to T_5 of 2020-01-01 using kriging interpolation

III. 미세먼지 농도 예측 방법 제안

Fig. 3은 PM2.5 예측을 위한 제안된 앙상블 3D 심층 컨볼루션 신경망(DCNN)의 전체적인 구조를 보여준다. Fig. 3에서 보여지는 것처럼, 제안 방법은 두 가지 주요 단계로 구성된다. 첫 번째 단계에서, 각 유형의 공간적 및 시간적 데이터는 각각의 3D-CNN에 입력으로 제공되어 해당 특징을 추출한다. 추출된 모든 특징들은 하나의 통합된 특징으로 결합 된다. 두 번째 단계에서, 통합된 특징들은 시간적 데이터와 융합되고, 그 후 완전 연결층(FC-Layer)을 통해 PM2.5 농도를 예측한다. 완전 연결층은 총 두 개의 레이어로 구성되어 있으며, 첫 번째 레이어는 각 입력 데이터 사이에서 의미 있는 특징을 추출하고, 두 번째 레이어는 각 시간에 대한 PM2.5 농도를 예측한다. 우리는 평균 제곱 오차 (Mean Squared Error) 손실 함수를 사용하여 예측된 PM2.5 농도와 관측된 PM2.5 농도(실제 값)를 비교하여 학습을 진행했다. 제안된 앙상블 3D-DCNN 예측 모델은 end-to-end 방식으로 훈련된다.

3.1 3D-DCNN 기반 시공간데이터의 특징 추출

우리는 입력 데이터의 공간적 및 시간적 특징 정보를 활용할 수 있는 VGG16 기반의 3D 심층 합성곱 신경망을 설계했다. 기존의 VGG16[29]은 2D 컨볼루션 레이어로 구성되어 있지만, 본 연구에서는 해당 모델을 기반으로 2D 컨볼루션 레이어에서 3D 컨볼루션 레이어로 수정하여 모델을 개선하였다. Table 1은 베이스라인 2D VGG16 모델을 수정하여 생성된 3D 심층 합성곱 신경망 구조를 보여준다. 3D-DCNN의 입력 데이터에 포함되어 있는 입력 인자들을 $\left\{X_k^i, y_k^{(T_{n+1}, T_{n+2}, \cdots, T_N)}\right\}_{k=1}^{m}$ 이라고 정의한다. 여기서 m은 입력 데이터의 총 개수, y는 각 시간대마다 관측된 PM2.5 농도, X^i는 각 입력 데이터를 뜻하며, $i \in \left\{PM_{2.5}^o, U^o, V^o, TA^o, RH^o, PM_{2.5}^f\right\}$는 각 입력 인자를 뜻한다. 여기서 윗 첨자 o는 관측 데이터를, f는 예측 데이터를 나타낸다. 관측된 대기질 데이터 $i = PM_{2.5}^o$는, $X^i = \left\{x^{T_1}, x^{T_2}, x^{T_3}, \ldots, x^{T_n}\right\}$, 이며 구성은

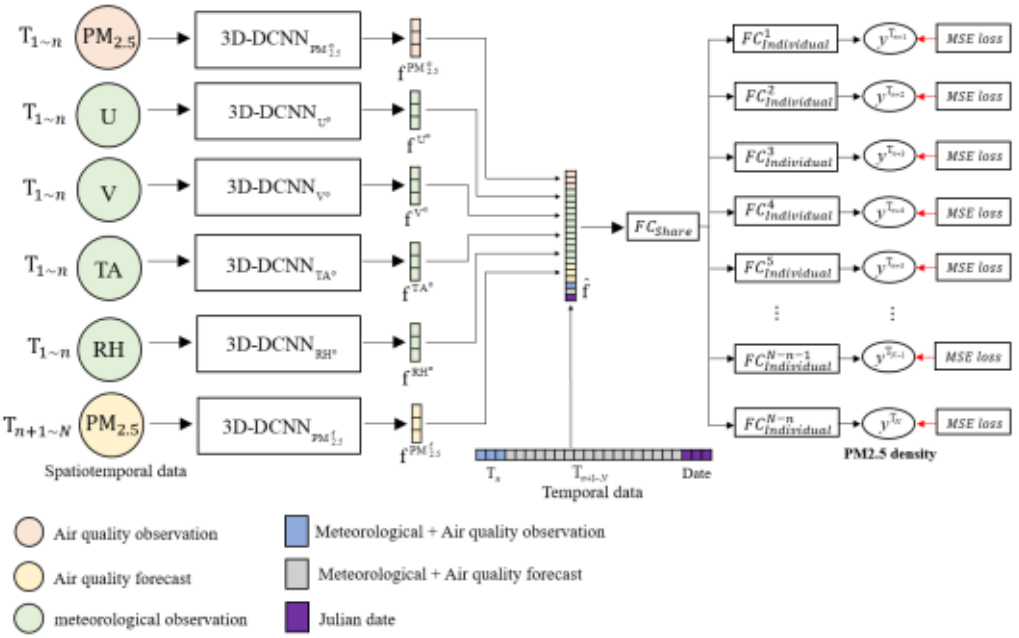


Fig. 3. Overall framework of the proposed approach, named Multi-output Ensemble 3D Convolutional Neural Network (ME3CNN), for $PM_{2.5}$ predictions.

$X^i \in R^{W \times H \times T_n}$로 설명할 수 있다. 기상 관측 데이터 $i = \{U^o, V^o, TA^o, RH^o\}$또한 관측된 대기질 데이터와 같은 구성이다. 대기질 예측 데이터 $i = PM_{2.5}^f$는, $X^i = \{x^{T_{n+1}}, x^{T_{n+2}}, x^{T_{n+3}}, ..., x^{T_N}\}$, $X^i \in R^{W \times H \times T_{N-n}}$로 구성한다. 각각 데이터는 3D-DCNN에 입력되어, 이후에는 식 (1)을 통해 특징이 추출된다.

$$f^i = \theta^i_{3D-DCNN}\left(X^i_k\right)^m_{k=1}, \quad i \in \left\{PM^o_{2.5}, U^o, V^o, TA^o, RH^o, PM^f_{2.5}\right\} \tag{1}$$

여기서 $\theta^i_{3D-DCNN}$는 i번째 3D-DCNN 모델로 특징 정보를 추출하는 것을 의미하며, i번째 3D-DCNN 모델을 사용한 특징 정보 추출 결과물은 $f^i \in R^{1 \times 1 \times 1 \times D}$와 같이 표현할 수 있다.

Table 1. Architecture of proposed 3D deep convolutional neural network.

No.	Layer Type	Description
1	Input	W x H x T_n
2	3D Conv-64	3x3x3x64;p=1,st=1
3	ReLU	
4	3D Conv-64	3x3x64x64;p=1,st=1
5	ReLU	
6	3D maxpool	2x2x2; st=2x2x2
7	3D Conv-128	3x3x3x128;p=1,st=1
8	ReLU	
9	3D Conv-128	3x3x128x128;p=1,st=1
10	ReLU	
11	3D maxpool	2x2x2; st=2x2x2
12	2D Conv-256	1x3x3x256;p=1,st=1
13	ReLU	
14	2D Conv-256	1x3x3x256;p=1,st=1
15	ReLU	
16	2D Conv-256	1x3x3x256;p=1,st=1
17	ReLU	
18	2D maxpool	1x2x2; st=1x2x2
19	2D Conv-512	1x3x3x512;p=1,st=1
20	ReLU	
21	2D Conv-512	1x3x3x512;p=1,st=1
22	ReLU	
23	2D Conv-512	1x3x3x512;p=1,st=1
24	ReLU	
25	2D maxpool	1x2x2; st=1x2x2
26	Reshape	1x3x4x512→1x1x1x6144
st: stride, p: padding		

3.2 PM2.5 농도 예측을 위한 특징 및 데이터 퓨전(Fusion)

3D-DCNN을 통해 추출된 특징은 시간적 데이터와 결합하여 식 (2)와 같이 구성된다.

$$\hat{f} = \Theta\left\{\begin{matrix} f^{PM^O_{25}}, f^{U^O}, f^{V^O}, f^{TA^O}, f^{RH^O}, \\ f^{PM^f_{25}}, f^{temp} \end{matrix}\right\} \tag{2}$$

Θ는 특성들을 연결(Concatenation)을 나타낸다. $\{f^{U^O}, f^{V^O}, f^{TA^O}, f^{RH^O}\}$는 기상 관측 특성, $f^{PM^O_{25}}$는 대기질 관측 특성, $f^{PM^f_{25}}$는 대기질 예측 특성, 그리고 시계열 데이터 f^{temp}를 통합하여 최종적으로 결합된 $\hat{f}$를 의미한다. 시계열 데이터의 형식은 $f^{temp} \in R^{1 \times 186}$이며, 시계열 데이터에 대한 자세한 내용은 4.2절에서 설명한다. 결합된 특성들은 공유된 완전 연결 레이어(shared fully-connected layer)와 개별 완전 연결 레이어(individual fully-connected layer)에 입력되어 각 시간대의 PM2.5 농도를 예측한다. 공유된 완전 연결 레이어는 결합된 특성들에 대한 의미 있는 정보를 추출하고 동시에 차원을 줄입니다. 그리고 개별 완전 연결 레이어는 독립적으로 설계되어 각 시간대의 PM2.5 농도를 정확하게 예측한다. 이는 식 (3)을 확인

$$\hat{f}_{fc} = \Theta_{Shared_FC}(\hat{f})$$
$$\left[\hat{y}^{T_{n+1}}, \hat{y}^{T_{n+2}}, ..., \hat{y}^{T_N}\right] = \Theta^{[1,2,...,N-n]}_{Individual_FC}\left(\hat{f}_{fc}\right) \tag{3}$$

공유된 완전 연결 레이어 Θ_{Shared_FC}는 특징 결합 $\hat{f}_{fc}$을 출력하며, $\Theta^{[1,2,...,N=n]}_{Individual_FC}$는 하나의 개별 완전 연결 레이어로, 총$N-n$개의 개별 완전 레이어를 구성한다. 이 경우 N은 미세먼지 농도를 예측할 시간이며, n은 현재까지 관측된 시간을 나타낸다. $\hat{y}$은 제안하는 앙상블 모델로 예측한 미세먼지 농도 값을 뜻하며, 모델은 총N번의 예측 결과를 출력한다. Table 2는 공유 완전 연결 레이어와 개별 완전 연결 레이어의 구조를 나타낸다. 제안 모델은 관측된 PM2.5 농도 값과 예측한 PM2.5 농도 값의 오차를 계산하며 학습한다. 이때 손실 함수는 평균 제곱 오차(Mean Squared Error)를 사용했다.

$$L_{MSE} = \frac{1}{m}\sum_k^m \left(y_k^{[T_{n+1}, T_{n+1}, ..., T_N]} - \hat{y}_k^{[T_{n+1}, T_{n+1}, ..., T_N]}\right)^2 \tag{4}$$

y은 관측된 PM2.5 농도(실제값)이며, $\hat{y}$는 제안 모델을 사용하여 예측된 PM2.5 농도 값이며, MSE 손실 함수를 사용해 관측값과 예측값 사이의 평균 제곱 차이를 측정하며, 예측된 값이 실제 값에 얼마나 근접하는지를 나타내고, MSE 값이 낮을수록 모델의 예측 성능이 좋다는 의미다.

Table 2. Architecture of share and individual fully-connected layer

No.	Layer Type	Description
1	Input	1x1x37,050
2	FC-Share	1x1x4000
3	ReLU	
4-1-1	FC-Individal_1	1x1x1000
4-1-2	ReLU	
4-1-3	FC-output_1	1x1x1
4-2-1	FC-Individal_2	1x1x1000
4-2-2	ReLU	
4-2-3	FC-output_2	1x1x1
4-3-1	FC-Individal_3	1x1x1000
4-3-2	ReLU	
4-3-3	FC-output_3	1x1x1
4-4-1	FC-Individal_4	1x1x1000
4-4-2	ReLU	
4-4-3	FC-output_4	1x1x1
4-5-1	FC-Individal_5	1x1x1000
4-5-2	ReLU	
4-5-3	FC-output_5	1x1x1
4-6-1	FC-Individal_6	1x1x1000
4-6-2	ReLU	
4-6-3	FC-output_6	1x1x1
4-7-1	FC-Individal_7	1x1x1000
4-7-2	ReLU	
4-7-3	FC-output_7	1x1x1
4-8-1	FC-Individal_8	1x1x1000
4-8-2	ReLU	
4-8-3	FC-output_8	1x1x1
4-9-1	FC-Individal_9	1x1x1000
4-9-2	ReLU	
4-9-3	FC-output_9	1x1x1
4-10-1	FC-Individal_10	1x1x1000
4-10-2	ReLU	
4-10-3	FC-output_10	1x1x1

Ⅳ. 실험 환경

실험에 사용된 데이터셋은 6시간마다 측정된 PM2.5와 기상 데이터로 구성되며, 2015년 1월 1일부터 2021년 2월 28일까지 6년 2개월에 걸쳐 수집되었다. 전체 데이터셋은 학습 데이터(2015년 1월 1일부터 2018년 12월 31일까지), 검증 데이터(2019년 1월 1일부터 2019년 12월 31일까지), 그리고 테스트 데이터(2020년 1월 1일부터 2021년 2월 28일까지)로 나누었다. 학습 데이터, 검증 데이터 및 테스트 데이터의 개수는 각각 1144, 180, 299개로 데이터가 측정되지 않거나 오류로 인해 발생한 날짜는 제외되었다. 측정 간격은 $T1 \cdots T5$이며, 예측 간격은 $T6 \cdots T15$이며, 각 간격은 6시간이며, 구현 환경은 CPU: i9-10900F, GPU: NVIDIA GeForce RTX 2080ti, RAM 128GB에서 실험하였다.

4.1 데이터 정규화

각 입력 요소의 값 분포가 다르기 때문에, 데이터 정규화를 통해 특정 범위로 동일하게 설정하는 것이 필요하다. 실제로, PM2.5의 최대값은 84이고 최소값은 2이며, TA의 최대값은 306이고 최소값은 259이다. 이처럼 입력 데이터의 값이 크고, 입력 데이터 간의 차이가 큰 경우 신뢰할 수 있는 학습이 진행되기 어렵다. 따라서, 입력 데이터 사이즈의 불균형을 해결하기 위해 미니멈-맥스 정규화(Min-Max normalization)를 사용하여 입력 데이터를 [0, 1] 범위로 정규화하였다.

$$\hat{x}=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{5}$$

여기서 x는 원래 데이터를 뜻하며, $x_{\min}$은 입력 데이터 요소의 최소값, $x_{\max}$는 입력 데이터 요소의 최대값이다. 식(5)와 같이 정규화를 진행하면 $\hat{x}$로, 입력 데이터가 모두 0과 1사이의 값을 가지며, 입력 데이터들 간의 크기 차이를 해소하여 모델의 학습을 원활하게 할 수 있다.

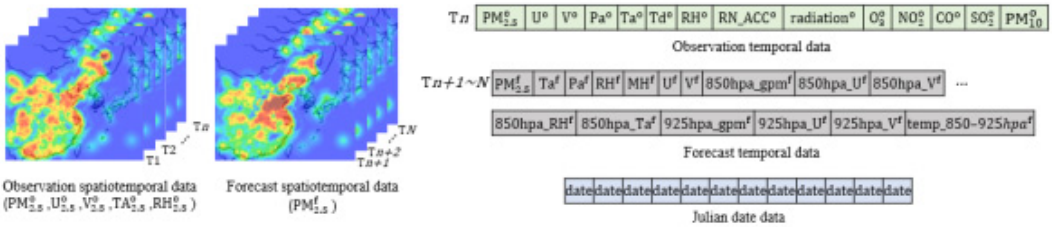
4.2 입력 데이터 구성

제안하는 모델의 입력 데이터 유형은 다음과 같다. 1) 관측 시공간 데이터, $W\times H\times T_n$ 차원의 데이터로,

$PM^{o}_{2.5}$, U^{o}, V^{o}, TA^{o}, RH^{o} 각 데이터가 크리깅을 통해 이미지로 생성되고, 시간 구간($T_1 \cdots T_n$)마다 순차적으로 쌓여있다. 2) 예측 시공간 데이터, $W \times H \times T_{N-n}$ 차원의 데이터로, $PM^{f_{25}}$의 데이터가 크리깅을 통해 이미지로 생성되고 시간 구간($T_{n+1} \cdots T_N$)마다 순차적으로 쌓여 있다. 본 연구에서는 Width(w)와 Height(h)가 각각 (146, 122)로 설정되었다. 3) 관측 시계열 데이터는 $[1 \times 14]$ 차원으로, $PM^{o}_{2.5}$, U^{o}, V^{o}, Pa^{o}, ta^{o}, td^{o}, RH^{o}, RN_ACC^{o}, $radiation^{o}$, O^{o}_{3}, NO^{o}_{2}, CO^{o}, SO^{o}_{2}, PM^{o}_{10} 로 구성된다. 4) 예측 시계열 데이터의 구성은 $PM^{f}_{2.5}$, Ta^{f}, Pa^{f} RH^{f}, MH^{f}, U^{f}, V^{f}, $850hpa_gpm^{f}$, $850hpa_U^{f}$, $850hpa_V^{f}$, $850hpa_RH^{f}$, $850hpa_Ta^{f}$, $925hpa_gpm^{f}$, $925hpa_U^{f}$, $925hpa_V^{f}$, $temp_850-925hpa^{f}$ 이며 $1 \times [16 \times (N-n)]$ 차원이며, ($T_{n+1} \cdots T_N$)의 데이터를 순차적으로 연결한 것으로 구성 된다. 본 연구에서는 총 시간대 길이(N)을 15로 설정하였으며, 관측 길이는(n) 5로 설정하였다. 따라서 예측해야할 시간대의 길이는 10이다. 5) 율리우스(Julian) 데이터(Fig. 4 참조), 1월부터 12월까지 해당 월의 요소를 나타내는 12개의 요소를 사용하며, 관련 월과 인접 월의 값을 멤버십 함수(membership function)[30]를 사용하여 계산한다. 데이터의 크기는 $[1 \times 12]$이다.

4.3 모델 평가 방법

모델의 성능을 평가하기 위해 평균 제곱근 오차, RMSE(Root Mean Square)$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$, 평균 절대 오차, MAE(Mean Absolute Error)$= \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$, 평균 편향 오차, MBE(Mean Bias Error)$= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$, 결정 계수, R^2(Coefficient of Determination)$= \left(\frac{Cov(\hat{y}, y)}{\sigma_y \sigma_{\hat{y}}}\right)$, Nash-Sutcliffe 효율성 계수 NSE(Nash-Sutcliffe model efficiency coefficient)$= 1 - \frac{\sum_{i=1}^{N}(\hat{y} - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$를 사용하였다. $\hat{y}$는 모델의 예측값, y는 관측값(실제 값)을 나타낸다. N는 전체 샘플의 수를, i는 특정 시간의 특정 샘플을, Cov는 공분산(Convariance)를, σ는 표준 편차를, $\bar{y}$는 시간 평균 관측값을 의미한다. RMSE는 예측한 값과 실제 값 간의 차이를 측정하는 데 자주 사용되는 지표다. MAE는 쌍으로 구성된 관측치 간의 오차를 나타내는 지표며, MBE는 주로 모델의 평균 편향을 추정하



Type	Input Variables
Observation	$PM_{2.5}$, U, V, Pa, Ta, Td, RH, *RN_ACC*, radiation, O_3, NO_2, CO, SO_2, PM_{10}
Forecast by CMAQ and WRF	$PM_{2.5}$, Ta, Pa, RH, MH, U, V, 850hpa_gpm, 850hpa_U, 850hpa_V, 850hpa_RH, 850hpa_Ta, 925hpa_gpm, 925hpa_U, 925hpa_V, temp_850-925hpa
Julian date	Membership function of Julian date to reflect the monthly change characteristics of PM2.5 concentration (Lee et al., 2022)
SO_2 (Sulfur dioxide), NO_2 (Nitrogen dioxide), O_3 (Ozone), CO (Carbon monoxide), PM_{10} and $PM_{2.5}$ (Particulate matter with aerodynamic diameters less than 10 μm and 2.5 μm, respectively), V and U (horizontal velocities), RN_ACC (accumulative precipitation), RH (relative humidity), Td (dew point temperature), Pa (Pressure), Rad (Solar radiation), Ta (temperature), MH (Mixing height), gpm (geopotential height), and Temp_850-925 hPa (Potential temperature difference between 850 and 925 hPa). 850 hPa and 925 hPa preceding the variables indicate the altitudes at which the variable was extracted from the three-dimensional WRF prediction.	

Fig. 4. Configuration information of Observation and Forecast spatiotemporal data, Observation and Forecast temporal data, and Julian date data.

고 모델 편향을 수정해야 하는지 여부를 결정하는 데 사용된다. RMSE, MAE, MBE의 값은 모두 $[-\infty \cdots \infty]$ 입니다. R^2는 통계 모델의 적합도를 측정하는 지표다[31]. 이 $[-0 \cdots 1]$ 값이다. NSE는 모델 시뮬레이션에서 결과 변수를 얼마나 잘 예측할 수 있는지를 측정하는 데 사용되는 지표이다[32]. 이 값은 $[-\infty \cdots 1]$다.

일반적으로 모델의 성능은 정확도, ACC(Accuracy), 탐지 확률, POD(Probability Of Detection), 거짓 경보율, FAR(False Alarm Rate)[33]과 같은 지표로 표현된다. Table 3을 참조하여 각 분포를 식에 대입하여 지표를 추출한다. $\mathrm{ACC}=\frac{A1+B2+C3+D4}{N}\times 100(\%)$, $\mathrm{POD}=\frac{IV}{III+IV}\times 100(\%)$, $\mathrm{FAR},=\frac{II}{II+IV}\times 100(\%)$. ACC는 측정(관측 또는 측정)된 값이 실제 값과 얼마나 근접하며, 정확한 수치를 예측하는지, POD는 모델이 특별한 상황(극고농도, 극저농도 등)에서도 잘 예측하는 능력을 나타낸다. FAR은 실제로 발생하지 않은 사건에 대해서 예측하는 것을 뜻한다. 이상적인 수치는 ACC와 POD는 100%의 값을 가지고, FAR은 0의 값에 가까워야 한다.

Table 3. Air pollution level expression method for evaluating the forecasting performance of the forecasting model by National Institute of Environmental Research (NIER)

Observation \ Predict	Good	Moderate	Bad	Very Bad
Good	A1	B1	C1	D1
Moderate	A2	B2	C2	D2
Bad	A3	B3	C3	D3
Very Bad	A4	B4	C4	D4

I = [A1, A2, A3, A4], II = [B1, B2, B3, B4]
III = [C1, C2, C3, C4], IV = [D1, D2, D3, D4]

V. 실험 결과

비교 실험에서는 PM2.5 농도의 예측 성능을 CMAQ, ConvLSTM, 하나의 시간대만 예측 가능한 앙상블 3D CNN 모델 (SE3CNN) 및 전체 시간대를 예측 가능한 앙상블 3D CNN 모델 (ME3CNN)과 비교했다. ConvLSTM과 SE3CNN은 제안된 ME3CNN의 입력 데이터와 동일하다. ConvLSTM은 일반적으로 Encoder와 Decoder로 구성되지만, 이 실험에서는 Encoder 부분만 사용되고, Encoder 이후에 MLP (Multi-Layer Perceptron)가 연결되어 미세먼지 농도를 예측하였다. SE3CNN은 제안된 ME3CNN의 구조와 거의 유사하다. 그러나 모델은 단일 간격을 예측하는 데에만 사용되며, 시공간 및 시간 예측 데이터 중 예측에 사용되는 기간의 데이터만 입력으로 사용한다. 예를 들어, 시공간 관측 데이터를 처리하는 모델은 이전 모델과 동일한 3D CNN을 사용하고, 예측 데이터를 처리하는 모델은 2D CNN으로 사용된다. 또한, 식 (2)에서 특징을 결합할 때, 예측할 기간의 데이터만 결합한다.

5.1 PM2.5 농도 예측 평가

서울의 현장 측정 데이터를 기반으로 6시간 평균 PM2.5 농도 예측이 평가하였다. Table 4에서 CMAQ는 관측값과 비교하여 RMSE, MAE 그리고 MBE 수치에서 매우 큰 차이가 있으며, R^2는 대부분 0에 가까운 값을 보였다. NSE는 T6을 제외하고 모두 0보다 작은 값을 가졌다. NSE가 0보다 작을 때는 예측 오차가 관측값의 분산보다 크다는 것을 의미한다. SE3CNN과 ME3CNN은 모델 평가 수치를 통해 분석하였을 때, CMAQ보다 훨씬 더 우수한 성능을 보이며, ConvLSTM보다도 더 나은 성능을 나타냈다. ME3CNN은 T6을 제외하고 RMSE, MAE, R^2에서 SE3CNN보다 더 나은 수치를 보였다. 이는 3D CNN이 공간 및 시계열 데이터에 대해 충분한 공간적 및 시간적 정보를 얻을 수 있으며, 하나의 통합된 예측 모델로 전체 기간의 예측 간격을 예측하는 것이 하나의 예측 모델을 설계하여 각 예측 기간을 예측하는 것보다 더 나은 이유라고 할 수 있다. 그러나 ME3CNN의 예측 성능이 T6에서 SE3CNN보다 낮은 이유는 T6을 예측할 때 T7~T15의 예측 데이터를 사용하지 않기 때문이다. SE3CNN은 ME3CNN과 달리 여러 개의 PM2.5 농도를 예측할 필요가 없으므로 한 개의 PM2.5 농도를 예측하는 데 집중할 수 있다. 그러나 ME3CNN이 T7~T15에서 SE3CNN보다 높은 이유는 이전 미세먼지 농도가 미래

Table. 4. Evaluation statistics for $PM_{2.5}$concentrations predictions in the Seoul. Mean absolute error (MAE), mean bias error (MBE), the coefficient of determination (R^2), and Nash-Sutcliffe efficiency (NSE) are presented.

	Model	RMSE	MAE	MBE	R2	NSE		Model	RMSE	MAE	MBE	R2	NSE
T6	CMAQ	13.07	9.64	-7.03	0.008	0.19	T11	CMAQ	18.33	13.71	-11.08	0.072	-0.44
	ConvLSTM	10.77	8.51	-3.05	0.518	0.45		ConvLSTM	14.33	10.46	-4.39	0.328	0.12
	SE3CNN	7.38	5.62	-2.27	0.771	0.74		SE3CNN	13.55	9.81	-3.56	0.360	0.21
	ME3CNN	8.42	6.11	-2.43	0.722	0.67		ME3CNN	12.27	9.20	-5.27	0.546	0.35
T7	CMAQ	16.32	12.66	-10.12	0.005	-0.09	T12	CMAQ	20.93	16.77	-13.16	0.125	-0.94
	ConvLSTM	12.40	8.69	-2.89	0.466	0.37		ConvLSTM	12.31	9.35	-3.86	0.446	0.33
	SE3CNN	12.23	8.89	-5.83	0.594	0.39		SE3CNN	12.62	9.55	-4.01	0.440	0.30
	ME3CNN	10.25	7.37	-3.41	0.643	0.57		ME3CNN	11.46	8.96	-5.42	0.595	0.42
T8	CMAQ	17.92	14.76	-12.10	0.021	-0.37	T13	CMAQ	20.22	15.79	-12.61	0.258	-0.69
	ConvLSTM	11.81	8.69	-2.89	0.477	0.41		ConvLSTM	13.99	10.59	-3.76	0.319	0.19
	SE3CNN	10.63	7.90	-3.93	0.614	0.52		SE3CNN	13.66	9.75	-2.07	0.345	0.23
	ME3CNN	9.57	7.25	-3.38	0.671	0.61		ME3CNN	11.11	8.71	-4.44	0.599	0.49
T9	CMAQ	19.00	15.38	-12.77	0.069	-0.44	T14	CMAQ	18.74	13.33	-9.40	0.341	-0.72
	ConvLSTM	13.16	9.83	-2.54	0.396	0.31		ConvLSTM	13.94	11.05	-5.60	0.344	0.05
	SE3CNN	11.20	8.26	-4.15	0.599	0.50		SE3CNN	13.41	10.14	-4.24	0.368	0.12
	ME3CNN	9.59	7.40	-2.86	0.674	0.63		ME3CNN	11.04	8.48	-4.06	0.540	0.40
T10	CMAQ	17.54	12.42	-9.35	0.082	-0.43	T15	CMAQ	21.75	15.89	-13.08	0.336	-1.22
	ConvLSTM	12.86	10.11	-4.66	0.410	0.23		ConvLSTM	14.56	10.95	-5.59	0.319	0.00
	SE3CNN	11.23	8.72	-3.60	0.538	0.41		SE3CNN	15.77	11.40	-6.34	0.336	-0.17
	ME3CNN	10.59	8.11	-3.84	0.595	0.48		ME3CNN	13.30	10.06	-5.68	0.447	0.17

미세먼지 농도를 예측하는 데 도움이 되기 때문이다. 또한, 하나의 통합된 모델이 특징 정보를 공유한다는 사실이 ME3CNN이 높은 성능을 가지는 이유다.

5.2 PM2.5 농도 단계 평가

예측을 위해 PM2.5 농도를 [낮음, 보통, 높음, 매우 높음]과 같은 범주형으로 변환하여 일반 대중이 직관적으로 이해할 수 있도록 하였다. 이 연구에서는 현장 측정 및 CMAQ, CovLSTM, SE3CNN, ME3CNN의 예측값을 4단계로 변환하여 범주형 PM2.5 수준 예측 성능 평가를 제시하였다. Table 5에서 관찰된 PM2.5 농도 수준은 분석 기간의 상당 부분에서 ME3CNN에 의해 정확하게 예측되었다. 더 직관적인 확인을 위해 Table 6을 제시한다. Table 6은 T6에서 T15까지의 예측된 PM2.5 농도를 범주형으로 변환한 경우의 ACC, POD, FAR을 보여준다. 이 연구에서 ME3CNN은 CMAQ 및 ConvLSTM보다 뛰어난 성능을 보였으며, T6을 제외한 T7에서 T15까지의 SE3CNN보다 더 좋은 성능을 나타냈다. 여기서 주목해야 할 부분은 POD와 FAR이다. 분명히 CMAQ의 POD는 매우 높지만 동시에 FAR도 매우 높다. 대부분의 모델은 POD가 높을수록 FAR도 높아지는 trade-off 문제가 있다. 우수한 모델을 만들기 위해서는 높은 ACC, POD와 낮은 FAR을 가져야 한다. ME3CNN은 T6을 제외한 모든 예측 간격에서 다른 모델들과 비교하여 상대적으로 높은 ACC, POD와 낮은 FAR을 보여준다. 더 직관적으로 확인하기 위해 ME3CNN이 ConvLSTM과 비교되었을 때 T6에서 T15까지의 ACC와 POD FAR을 평균하여, ACC가 7.06% 향상되었고, POD가 11.77% 향상되었으며, FAR가 10.82% 감소했다. 또한, ME3CNN이 SE3CNN과 비교되었을 때, ACC가 2.78% 향상되었고, POD가 3.62% 향상되었으며, FAR가 5.92% 감소했다.

Table 5. A contingency table displaying the number of days observed and forecasted for PM2.5 concentration level of low(PM2.5 $\leq 15\mu g/m^3$), moderate($16\mu g/m^3 \leq$ PM2.5 $\leq 35\mu g/m^3$), high($36\mu g/m^3 \leq$ PM2.5 $\leq 75\mu g/m^3$), very high(76 $\mu g/m^3 \leq$ PM2.5).

Forecast (CAMQ / ConvLSTM / SE3CNN / ME3CNN)						
Time	Level	Low	Moderate	High	Very high	Total
T6	Low	34/23/40/43	42/55/37/35	2/0/1/0	0/0/0/0	78/78/78/78
	Moderate	2/6/1/3	70/79/94/89	37/25/15/18	1/0/0/0	110/110/110/110
	High	0/0/0/0	4/11/7/6	29/28/33/34	8/2/1/1	41/41/41/41
	Very high	0/0/0/0	0/0/0/0	0/0/0/0	0/0/0/0	0/0/0/0
	Total	36/29/41/46	116/145/138/130	68/53/49/52	9/2/1/1	299/299/299/299
T7	Low	18/30/24/28	47/36/43/40	3/2/1/0	0/0/0/0	68/68/68/68
	Moderate	6/9/4/6	63/84/86/93	48/25/28/19	2/1/1/1	119/119/119/119
	High	2/1/0/0	1/13/4/8	28/24/33/30	8/1/2/1	39/39/39/39
	Very high	0/0/0/0	0/0/0/0	1/3/2/3	2/0/1/0	3/3/3/3
	Total	26/40/28/34	111/133/133/141	80/54/64/52	12/2/4/2	299/299/299
T8	Low	22/30/29/28	41/34/37/39	4/3/1/0	0/0/0/0	67/67/67/67
	Moderate	4/8/9/4	45/84/81/95	67/24/26/17	0/0/0/0	116/116/116/116
	High	2/2/1/0	4/14/9/11	29/28/34/33	9/0/0/0	44/44/44/44
	Very high	0/0/0/0	0/0/0/0	1/2/2/2	1/0/0/0	2/2/2/2
	Total	28/40/39/32	90/132/127/145	101/57/63/52	10/0/0/0	299/299/299/299
T9	Low	19/28/25/24	44/35/39/41	2/2/1/0	0/0/0/0	65/65/65/65
	Moderate	1/8/2/3	38/72/83/91	69/29/24/15	1/0/0/0	109/109/109/109
	High	0/0/0/0	4/16/6/11	34/33/46/40	14/0/0/1	52/52/52/52
	Very high	0/0/0/0	0/1/0/0	2/2/3/3	1/0/0/0	3/3/3/3
	Total	20/38/27/27	86/124/128/143	107/66/74/58	16/1/0/1	299/299/299/299
T10	Low	36/25/36/30	40/53/43/50	5/3/2/1	0/0/0/0	81/81/81/81
	Moderate	7/4/6/1	56/79/75/89	42/25/27/18	3/0/0/0	108/108/108/108
	High	1/0/0/0	4/12/5/8	23/26/33/32	12/2/2/0	40/40/40/40
	Very high	0/0/0/0	0/0/0/0	0/0/0/0	0/0/0/0	0/0/0/0
	Total	44/29/42/31	100/144/123/147	70/54/62/51	15/2/2/0	299/299/299/299
T11	Low	21/31/28/26	48/38/41/44	3/3/3/2	0/0/0/0	72/72/72/72
	Moderate	7/7/11/5	55/87/83/88	54/24/24/24	3/1/1/2	119/119/119/119
	High	1/0/0/0	3/13/15/8	23/20/19/26	8/2/1/1	35/35/35/35
	Very high	0/0/0/0	0/0/0/0	1/3/3/3	2/0/0/0	3/3/3/3
	Total	29/38/39/31	106/138/139/140	81/50/49/55	13/3/2/3	299/299/299/299
T12	Low	24/28/31/29	42/42/41/43	6/2/0/0	0/0/0/0	72/72/72/72
	Moderate	6/9/9/4	39/76/75/86	67/29/30/24	2/0/0/0	114/114/114/114
	High	2/2/1/0	2/7/8/8	26/32/32/32	11/0/0/1	41/41/41/41
	Very high	0/0/0/0	0/0/1/0	2/2/1/2	0/0/0/0	2/2/2/2
	Total	32/39/41/33	83/125/125/137	101/65/63/58	13/0/0/1	299/299/299/299
T13	Low	23/21/35/26	39/41/32/42	8/8/2/2	0/0/1/0	70/70/70/70
	Moderate	4/5/7/4	44/79/80/80	59/24/21/24	1/0/0/0	108/108/108/108
	High	1/1/0/0	6/12/15/5	2736/32/43	15/0/2/1	49/49/49/49
	Very high	0/0/0/0	1/1/1/0	1/1/1/2	0/0/0/0	2/2/2/2
	Total	28/27/42/30	90/133/128/127	95/69/56/71	16/0/3/1	299/299/299/299

Table 6. Categorical statistics for assessing daily PM2.5 level predictions in Seoul during January 2020 to February 2021 include accuracy, probability of detection (POD), and false alarm rate (FAR).

	Model	ACC↑	POD↑	FAR↓		Model	ACC↑	POD↑	FAR↓
T6	CMAQ	61.1(140/229)	87.8(36/41)	52.6(40/76)	T11	CMAQ	47.6(109/229)	89.5(34/38)	61.4(54/88)
	ConvLSTM	59.8(137/229)	68.3(28/41)	41.7(20/48)		ConvLSTM	62.9(144/229)	63.2(24/38)	47.8(22/46)
	SE3CNN	76(174/229)	82.9(34/41)	29.2(14/48)		SE3CNN	58.5(134/229)	55.3(21/38)	54.3(25/46)
	ME3CNN	73.4(168/229)	80.5(33/41)	29.8(14/47)		ME3CNN	62.9(144/229)	78.9(30/38)	44.4(24/54)
T7	CMAQ	49.3(113/229)	90.5(38/42)	56.8(50/88)	T12	CMAQ	42.4(97/229)	88.4(38/43)	65.1(71/109)
	ConvLSTM	63.3(145/229)	61.9(26/42)	48(24/50)		ConvLSTM	60.3(138/229)	74.4(32/43)	46.7(28/60)
	SE3CNN	65.9(151/229)	88.1(37/42)	41.3(26/63)		SE3CNN	61.6(141/229)	72.1(31/43)	43.6(24/55)
	ME3CNN	70.7(162/229)	81(34/42)	34.6(18/52)		ME3CNN	65.1(149/229)	81.4(35/43)	38.6(22/57)
T8	CMAQ	45(103/229)	84.8(39/46)	62.5(65/104)	T13	CMAQ	42.4(97/229)	84.3(43/51)	60.6(66/109)
	ConvLSTM	62(142/229)	63(29/46)	47.3(26/55)		ConvLSTM	59.8(137/229)	66.7(34/51)	45.2(28/62)
	SE3CNN	65.5(150/229)	73.9(34/46)	39.3(22/56)		SE3CNN	65.1(149/229)	64.7(33/51)	40(22/55)
	ME3CNN	69(158/229)	73.9(34/46)	30.6(15/49)		ME3CNN	65.9(151/229)	84.3(43/51)	35.8(24/67)
T9	CMAQ	42.4(97/229)	90.9(50/55)	58(69/119)	T14	CMAQ	53.3(122/229)	81.6(31/38)	56.9(41/72)
	ConvLSTM	59.4(136/229)	65.5(36/55)	42.9(27/63)		ConvLSTM	55(126/229)	65.8(25/38)	52.8(28/53)
	SE3CNN	68.1(156/229)	87.3(48/55)	30.4(21/69)		SE3CNN	60.3(138/229)	63.2(24/38)	48.9(23/47)
	ME3CNN	69.4(159/229)	80(44/55)	24.1(14/58)		ME3CNN	64.6(148/229)	73.7(28/38)	40.4(19/47)
T10	CMAQ	53.3(122/229)	85(34/40)	58(47/81)	T15	CMAQ	44.1(101/229)	80(28/35)	69.2(63/91)
	ConvLSTM	59(135/229)	65(26/40)	44.7(21/47)		ConvLSTM	56.8(130/229)	68.6(24/35)	61.9(39/63)
	SE3CNN	64.2(147/229)	85(34/40)	43.3(26/60)		SE3CNN	55.9(128/229)	71.4(25/35)	59.7(37/62)
	ME3CNN	66.8(153/229)	75(30/40)	38.8(19/49)		ME3CNN	61.1(140/229)	71.4(25/35)	53.7(29/54)

Ⅵ. 결론

이 논문에서는 대한민국 서울 지역에서 2일 동안 매 6시간 간격으로 관측되는 PM2.5 농도를 예측하기 위해 3D CNN을 사용한 새로운 앙상블 예측 접근 방식을 제안하였다. 제안된 모델은 2015년 1월부터 2019년 12월까지 매일 PM2.5 농도와 네 가지 범주의 예측 데이터를 사용하여 5년간의 데이터를 훈련한 후, 2020년 1월부터 2021년 2월까지의 데이터를 사용하여 성능을 종합적으로 평가했다. 평가 지표를 기반으로, 제안된 방법의 평균 MAE는 8.165이며, CMAQ 및 ConvLSTM 모델과 비교했을 때 각각 5.031과 1.658로, 제안된 모델이 다른 모델에 비해 높다. 미세먼지 농도 예측 성능은 제안하는 모델과 CMAQ를 비교했을 때, 평균 ACC가 18.8% 증가, 평균 POD가 8.27% 증가, 평균 FAR가 23.03% 감소되어 전체적인 성능이 향상되었다. 제안 모델이 ConvLSTM과 비교하여 평균 ACC가 7.06% 증가, 평균 POD가 4.28% 증가, 평균 FAR가 11.77% 감소하여 성능이 향상되었다. 예측 실험 결과, ME3CNN은 CMAQ와 ConvLSTM보다 더 나은 성능을 보여주었다. 따라서 제안하는 PM 농도 예측 시스템은 1차적으로 대기질 예보관에게 유용하고 보완적인 정보를 제공한다. 또한 이러한 정보를 바탕으로 건강, 교통, 환경 보호, 에너지 효율성, 농업 등 실생활의 다양한 분야에서 도움을 기대할 수 있다.

REFERENCES

[1] United States Environmental Protection Agency. (2023, July 11). *What is PM, and how does it get into the air?* Particulate Matter(PM) Basics. https://www.epa.gov/pm-pollution

[2] F. Lu, D. Xu, Y. Cheng, S. Dong, C. Guo, X. Jiang & X. Zheng. (2015). Systematic review

and meta-analysis of the adverse health effects of ambient PM25 and PM10 pollution in the Chinese population. *Environmental Research, 136*, 196-204.

[3] N. A. H. Janssen, P. Fischer, M. Marra, C. Ameling & F. R. Cassee. (2013). Short-term effects of PM2. 5, PM10 and PM2. 5 - 10 on daily mortality in the Netherlands. *Science of the Total Environment, 463*, 20-26.

[4] M. L. Scapellato, C. Canova, A. De Simone, M. Carrieri, P. Maestrelli, L. Simonato & G. B. Bartolucci. (2009). Personal PM10 exposure in asthmatic adults in Padova, Italy: Seasonal variability and factors affecting individual concentrations of particulate matter. *International Journal of Hygiene and Environmental Health, 212*(6), 626-636.

[5] S. Wu, F. Deng, Y. Hao, X. Wang, C. Zheng, H. Lv ... & X. Guo. (2014). Fine particulate matter, temperature, and lung function in healthy adults: Findings from the HVNR study. *Chemosphere, 108*, 168-174.

[6] M. C. Turner, D. Krewski, C. A. Pope III, Y. Chen, S. M. Gapstur & M. J. Thun. (2011). Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers. *American Journal of Respiratory and Critical Care Medicine, 184*(12), 1374-1381.

[7] Y. S. Lin, Y. H. Chang & Y. S. Chang. (2016, December). Constructing PM2.5 map based on mobile PM2.5 sensor and cloud platform. In *2016 IEEE International Conference on Computer and Information Technology(CIT)* (pp. 702-707). IEEE.

[8] Environmental Performance Index. (2023, July 21). *PM2.5 exposure*. RESULTS. https://epi.yale.edu/epi-results/2022/component/pmd

[9] Air Korea. (2023a, July 20). *Introduction to the Comprehensive air-quality index(CAI)*. What's CAI. https://www.airkorea.or.kr/eng/khaiInfo?pMENU_NO=166

[10] Air Korea. (2023b, July 20). *Emergency measures to reduce fine dust*. Fine Dust situation. https://www.airkorea.or.kr/portal/web/contents/contentView/?pMENU_NO=148

[11] H. Chang-Hoi, I. Park, H. R. Oh, H. J. Gim, S. K. Hur, J. Kim & D. R. Choi. (2021). Development of a PM2.5 prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, Republic of Korea. *Atmospheric Environment, 245*, 118021.

[12] T. C. Bui, V. D. Le & S. K. Cha. (2018). *A deep learning approach for forecasting air pollution in South Korea using LSTM*. arXiv:1804.07891.

[13] C. J. Huang & P. H. Kuo. (2018). A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities. *Sensors, 18*(7), 2220.

[14] L. S. Chang, A. Cho, H. Park, K. Nam, D. Kim, J. H. Hong & C. K. Song. (2016). Human-model hybrid Korean air quality forecasting system. *Journal of the Air & Waste Management Association, 66*(9), 896-911.

[15] K. Chakraborty, K. Mehrotra, C. K. Mohan & S. Ranka. (1992). Forecasting the behavior of multivariate time series using neural networks. *Neural Networks, 5*(6), 961-970.

[16] A. K. Paschalidou, S. Karakitsios, S. Kleanthous & P. A. Kassomenos. (2011). Forecasting hourly PM 10 concentration in Cyprus through artificial neural networks and multiple regression models: Implications to local environmental management. *Environmental Science and Pollution Research, 18*, 316-327.

[17] W. Z. Lu, W. J. Wang, H. Y. Fan, A. Y. T. Leung, Z. B. Xu, S. M. Lo & J. C. K. Wong. (2002). Prediction of pollutant levels in

causeway bay area of Hong Kong using an improved neural network model. *Journal of Environmental Engineering, 128*(12), 1146-1157.

[18] G. Yang, H. Lee & G. Lee. (2020). A hybrid deep learning model to forecast particulate matter concentration levels in Seoul, South Korea. *Atmosphere, 11*(4), 348.

[19] J. Schmidhuber & S. Hochreiter. (1997). Long short-term memory. *Neural Comput, 9*(8), 1735-1780.

[20] H. C. Chen, K. T. Putra & J. Chun-WeiLin. (2021). *A novel prediction approach for exploring PM2. 5 spatiotemporal propagation based on convolutional recursive neural networks*. arXiv:2101.06213.

[21] T. Yasuno, A. Ishii & M. Amakata. (2021). Rain-code fusion: Code-to-code ConvLSTM forecasting spatiotemporal precipitation. In *Pattern recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII* (pp. 20-34). Springer International Publishing.

[22] W. C. Skamarock & J. B. Klemp. (2008). A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics, 227*(7), 3465-3485.

[23] R. Borge, V. Alexandrov, J. J. Del Vas, J. Lumbreras & E. Rodríguez. (2008). A comprehensive sensitivity analysis of the WRF model for air quality applications over the Iberian Peninsula. *Atmospheric Environment, 42*(37), 8560-8574.

[24] D. W. Byun, J. K. Ching, J. Novak & J. Young. (1998). Development and implementation of the EPA's models-3 initial operating version: Community multi-scale air quality (CMAQ) model. *Air Pollution Modeling and Its Application XII*, 357-368.

[25] D. Byun & K. L. Schere. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Applied Mechanics Reviews, 59*(2), 51-77.

[26] H. Mahmoudabadi & G. Briggs. (2016, December). Directional kriging implementation for gridded data interpolation and comparative study with common methods. In *AGU Fall Meeting Abstracts* (Vol. 2016, pp. G21B-1006).

[27] J. Meng. (2021). Raster data projection transformation based-on Kriging interpolation approximate grid algorithm. *Alexandria Engineering Journal, 60*(2), 2013-2019.

[28] P. M. Bartier & C. P. Keller. (1996). Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers & Geosciences, 22*(7), 795-799.

[29] K. Simonyan & A. Zisserman. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv:1409.1556.

[30] J. B. Lee, J. B. Lee, Y. S. Koo, H. Y. Kwon, M. H. Choi, H. J. Park & D. G. Lee. (2022). Development of a deep neural network for predicting 6 h average PM 2.5 concentrations up to 2 subsequent days using various training data. *Geoscientific Model Development, 15*(9), 3797-3813.

[31] J. L. Devore. (1995). *Probability and statistics for engineering and the sciences* (Vol. 5). Belmont: Duxbury Press.

[32] J. E. Nash & J. V. Sutcliffe. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology, 10*(3), 282-290.

[33] D. Kang, R. Mathur, K. Schere, S. Yu & B. Eder. (2007). New categorical metrics for air quality model evaluation. *Journal of Applied Meteorology and Climatology, 46*(4), 549-555.

저 자 약 력

이준민(Lee, Joon Min) [회원]



- 2019년 6월 : 북경제2외국어대학교 번역학과 학사
- 2021년 2월~현재 : 한국외국어대학교 컴퓨터공학과 석사과정

E-Mail : joonminlee@hufs.ac.kr

<관심분야>
딥러닝, 패턴인식, 영상처리, 예측

김경태(Kim, Kyeong Tae) [회원]



- 2016년 6월 : 중원대학교 학사
- 2019년 3월 : 한국외국어대학교 컴퓨터공학과 석사
- 2019년 2월~현재 : 한국외국어대학교 컴퓨터공학과 박사과정

E-Mail : rudxo9251@gmail.com

<관심분야>
머신러닝, 패턴인식, 영상처리

최재영(Choi, Jae Young) [회원]



- 2011년 8월 : KAIST 전기및전자공학과 박사
- 2008년 11월~2012년 5월 : 토론토대학 연구원
- 2012년 7월~2013년 6월 : 펜실베니아대학 연구원
- 2013년 9월~2014년 8월 : 삼성전자 책임연구원
- 2014년 9월~2016년 8월 : 중원대학 의료공학과 조교수
- 2023년 3월~2024년 2월 : 브리티시컬럼비아대학 컴퓨터공학과 방문교수
- 2016년 9월~현재 : 한국외국어대학 컴퓨터공학부 부교수

E-Mail : jychoi@hufs.ac.kr

<관심분야>
딥러닝, 머신러닝, 패턴인식, 영상처리