JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

응용 기술

J. Adv. Navig. Technol. 28(5): 760-766, Oct. 2024

SHAP을 활용한 PM2.5 예측 모델 성능 및 변수 영향력 분석

PM2.5 Prediction Model Performance and Variable Impact Analysis Using SHAP

정 용 진 · 오 창 헌*
한국기술교육대학교 전기전자통신공학과

Yong-jin Jung · Chang-Heon Oh*

Department of Electrical, Electronics and Communication Engineering, Korea University of Technology and Education (KOREATECH), Chungcheongnam-do, 31253, Korea

[요 약]

본 논문에서는 SHAP을 사용하여 변수들이 예측 값에 어떠한 영향을 주었는지 분석하였다. DNN과 LSTM 알고리즘을 사용하여 $PM_{2.5}$에 대한 예측 모델을 설계하였다. 학습 및 테스트 데이터는 기상데이터와 대기오염물질데이터를 상관분석을 통해 선별하여 구성하였다. 두 예측 모델에 대해 RMSE와 AQI의 범주에 대한 정확도를 확인하였으며, SHAP을 이용하여 두 모델의 예측 값에 대해 변수들의 기여도를 확인하였다. 공통적으로 대기오염물질 데이터가 $PM_{2.5}$를 예측 하는 과정에서 기여도가 높은 것을 확인하였으며, 기상데이터 중 온도가 두 모델의 예측 과정에서 기여도가 높은 것을 확인하였다. 그리고 기여도에 따른 영향력을 확인하였을 때, 두 모델이 공통적으로 온도, 풍속, 해면기압이 값이 높을 때 예측 값을 감소시키는 영향을 주며, 값이 낮을 때 예측 값을 증가시키는 영향을 주었다. NO_2, PM_{10}, SO_2의 경우 DNN 예측 모델과는 달리 LSTM 예측 모델에서는 값이 높을 때 예측 값의 양방향으로 영향을 주는 것을 확인하였다.

[Abstract]

Machine learning and deep learning are being researched in various fields and applied in real life. Designing reliable models is crucial, and understanding the results of these models is necessary. This paper analyzes the impact of variables on prediction values using SHAP. Prediction models for $PM_{2.5}$ were designed using DNN and LSTM algorithms. The training and test data were composed by selecting weather data and air pollutant data through correlation analysis. The RMSE and accuracy for AQI categories were checked for both prediction models, with the LSTM algorithm showing slightly better performance. The contribution of variables to the prediction values of both models was confirmed using SHAP. It was found that air pollutant data had a high contribution in predicting $PM_{2.5}$, and temperature among weather data had a high contribution in the prediction process of both models. Both models showed that high values of temperature, wind speed, and sea level pressure decreased prediction values, while low values increased them. For NO_2, PM_{10}, and SO_2, the LSTM model showed a bidirectional impact on prediction values, unlike the DNN model.

Key word : SHAP, Deep learning, LSTM, Particulate matter.

http://dx.doi.org/10.12673/jant.2024.28.5.760

Received 7 October 2024; **Revised** 26 October 2024
Accepted (Publication) 29 October 2024 (31 October 2024)

***Corresponding Author; Chang-Heon Oh**

Tel: +82-41-560-1187
E-mail: choh@koreatech.ac.kr

Ⅰ. 서 론

미세먼지는 눈에 보이지 않을 정도로 작은 입자의 먼지이다. 입자의 직경 10 μm 이하의 크기를 가지는 먼지를 미세먼지(PM_{10})라 하며 직경 2.5 μm 이하의 크기를 가지는 먼지는 초미세먼지($PM_{2.5}$)라고 한다[1]. 매우 작은 입자의 먼지이기 때문에 호흡기를 통해 쉽게 노출될 수 있으며, 이에 따라 장시간 노출될 경우 호흡기 질환, 심혈관 질환, 폐 질환, 세포 노화 등 건강에 영향 좋지 않은 영향을 주는 요인으로 보고되고 있다[2]-[4]. 또한 세계보건기구(WHO; World Health Organization) 산하 국제암연구소(IARC; International Agency for Research on Cancer)는 일찍이 2013년도에 미세먼지를 1군 발암물질로 분류하였으며, 세계의 모든 사람들이 미세먼지에 대한 위험성을 인지하였다[5]. 이에 많은 대중들은 미세먼지가 심할 경우, 야외에서의 여가 및 경제 활동을 자제 등 사회의 경제 활동 저하의 원인으로 분석되고 있다[6]-[8]. 미세먼지에 대한 사회의 영향력이 큰 만큼 미세먼지 예보에 대한 관심도 높아졌으며, 더욱 정확한 예보가 요구되고 있다. 따라서 미세먼지에 대한 다양한 연구가 진행되고 있다. 이 중 머신러닝이나 딥러닝 알고리즘을 활용하여 예측 모델을 통해 미세먼지 예측 성능을 향상시키기 위한 연구들도 진행되고 있고, 예측 모델의 학습에 대해 다양한 데이터들의 분석도 진행되고 있다.

Li(2024)의 연구에서는 베이징의 대기 중 $PM_{2.5}$ 농도 변화의 주요 요인을 분석하는 연구를 진행했다. Random forest 알고리즘을 이용하여 $PM_{2.5}$ 예측 모델을 설계하였으며, 학습 데이터로 대기오염물질을 사용하였다. SHAP (shapley additive explanations)을 이용하여 예측 모델의 분석 결과 대기오염물질 중 NO_2와 SO_2이 예측 값에 대한 기여도가 높은 것을 확인하였다[9]. Gao(2022)의 연구에서는 하얼빈의 $PM_{2.5}$ 농도 예측에 대한 대기오염물질 및 기상요인의 분석을 진행하였다. Random forest 알고리즘과 대기오염물질데이터, 기상데이터, 에어로솔 광학두께(AOD; aerosol optical depth) 데이터를 사용하여 예측 모델을 설계하였다. 모델의 평가는 평균 제곱근 오차(RMSE; root mean squared error), 평균 절대 오차(MAE; mean absolute error) 사용하였으며, 13.23, 6.69의 값을 확인하였다. 변수들이 $PM_{2.5}$의 예측과정에서의 기여도을 분석하기 위해 SHAP를 사용하였으며, 평균 습도와 AOD의 기여도가 높은 것을 확인하였다[10]. Guo(2023)의 연구에서는 Catboost 알고리즘을 이용하여 실내의 $PM_{2.5}$ 농도 분류 예측 모델을 연구하였다. 학습 및 테스트에 사용하기 위한 데이터는 실내 온도, 습도, CO_2, 조도를 사용하였다. SHAP를 사용하여 예측 값에 대한 기여도를 확인한 결과 한 시간 전의 $PM_{2.5}$가 가장 기여도가 높았으며, 두 번째로 습도가 가장 높은 것을 확인하였다[11].

본 논문에서는 SHAP을 이용하여 $PM_{2.5}$ 예측 모델에 대한 변수 영향력을 분석하였다. 딥러닝 알고리즘 중 심층신경망(DNN; deep neural networks) 알고리즘과 LSTM(long short-term memory) 알고리즘을 사용하였으며, 2019년부터 5년간 천안시와 오창읍에서 측정한 기상 및 대기오염물질 데이터를 사용하였다. 상관분석을 통해 수집한 데이터 중 $PM_{2.5}$와 상관성이 있는 변수를 선정하였으며, 이 결과를 기반으로 학습 및 테스트 데이터를 구성하였다. DNN과 LSTM 알고리즘 기반의 예측 모델은 하이퍼 파라미터 탐색을 통해 최적의 파라미터를 설정하였으며, 예측 성능 평가를 위해 RMSE와 AQI(air quality index)의 범주를 활용한 정확도를 기준으로 진행하였다. 이 후 SHAP을 이용하여 예측 값에 대한 변수들의 영향력을 확인하였다.

Ⅱ. 데이터 수집 및 구성

2-1 데이터 수집 및 학습 데이터 선정

$PM_{2.5}$ 예측 모델의 학습 및 테스트 데이터는 기상청과 환경공단에서 시간 단위로 측정한 데이터를 사용하였다. 기상청의 기상자료개방포털을 통해 천안시 측정소에서 수집한 종관기상관측 데이터 중 2019년도부터 2023년도까지의 온도, 이슬점 온도, 강수량, 습도, 풍속, 풍향, 증기압, 현지기압, 해면기압, 전운량에 대한 데이터를 수집하였다. 그리고 환경공단의 Air korea를 통해 천안시의 기상데이터 측정소와 인접한 오창 측정소의 대기오염물질 데이터를 수집하였다. 대기오염물질 데이터는 $PM_{2.5}$, PM_{10}, O_3, NO_2, CO, SO_2를 포함한다.

예측 모델에 사용할 데이터로 예측 대상인 $PM_{2.5}$와 다른 데이터들 간 상관분석을 진행하였다. 상관분석은 pearson 상관 계수를 사용하였다. Pearson 상관 계수는 두 변수 사이의 선형 관계의 강도와 방향을 측정하는 방법이며, 연속형 변수와 연속형 변수 간의 관계를 확인할 때 사용한다. 표 1은 수집한 데이터와 예측 대상인 $PM_{2.5}$과의 상관분석 결과이다.

표 1. 상관분석 결과
Table 1. Correlation analysis result.

Variable		correlation analysis
Meteorological elements	Station pressure	0.24
	Sea-level pressure	0.24
	Vapor pressure	-0.30
	Temperature	-0.27
	Dew-point Temp.	-0.24
	Wind speed	-0.21
	Precipitation	-0.08
	Wind direction	-0.08
	Cloud cover	-0.04
	Humidity	0.00
Air pollutants	PM_{10}	0.80
	CO	0.63
	NO_2	0.56
	SO_2	0.27
	O_3	-0.09
	$PM_{2.5}$	1

기상데이터 중 습도는 $PM_{2.5}$와 상관성이 없는 것으로 확인되었으며, 강수량, 풍향, 전운량은 거의 상관성이 없는 것으로 확인되었다. 현지기압, 해면기압은 낮은 양의 상관관계를 보였으며, 증기압, 온도, 이슬점온도, 풍속은 낮은 음의 상관관계를 보였다. 대기오염물질 데이터 중 O_3은 거의 상관성이 없는 것으로 확인되었다. SO_2는 낮은 양의 상관관계를 보였으며, CO와 NO_2는 다소 높은 양의 상관관계를 보였다. PM_{10}은 높은 양의 상관관계를 보였다.

상관분석 결과를 활용하여 예측 모델의 학습 데이터를 현지기압, 해면기압, 증기압, 온도, 이슬점온도, 풍속, PM_{10}, CO, NO_2, SO_2로 선정하였다.

2-2 데이터 구성

예측 모델에 사용되는 데이터는 모델의 학습에 사용될 데이터와 모델의 성능을 평가하기 위한 테스트 데이터로 구성된다. 학습 데이터는 학습 과정에서 검증을 위한 검증 데이터가 필요하며, 학습데이터 중 25%로 구성하였다. 그림 1은 예측 모델의 학습 및 테스트에 사용될 dataset 구조이다.

2019년 1월부터 2022년 12월까지의 데이터는 training set으로 구성하였으며, 2023년 데이터는 test set으로 구성하였다. Training set 중 2021년 12월까지의 데이터는 train set으로 구성하였으며 2022년 데이터는 validation set으로 구성하였다. 구성한 데이터 중 학습에 필요한 데이터의 경우, min max scaler를 이용하여 전처리를 진행하였다.

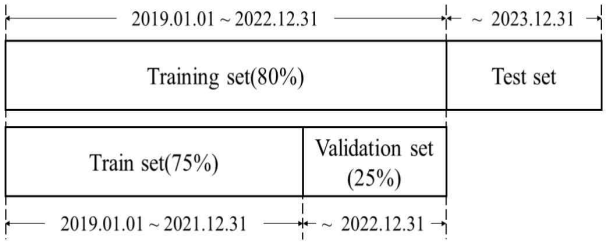


그림 1. 데이터셋 구조
Fig. 1. Structure of dataset.

표 2. 하이퍼 파라미터 탐색 결과
Table 2. Hyper parameter search result.

parameter	DNN	LSTM
units	140	80
dropout rate	0.5	0.2
batch size	200	20

Ⅲ. 예측 모델 설계

3-1 DNN 기반의 예측 모델

DNN은 기존의 인공신경망의 단점을 보완하기 위해 설계된 딥 러닝 알고리즘으로, 학습 양이 많아질수록 과적합(over fitting)에 대한 문제를 n개의 hidden layer를 사용한다. $PM_{2.5}$를 예측하기 위한 모델 중 하나로 DNN 알고리즘을 사용하여 설계하였다. 2개의 hidden layer를 사용하였으며, 뉴런과 뉴런 간의 신호 전달하는 역할의 활성화 함수는 ReLU를 각 layer에 적용하였다. 활성화 함수 ReLU는 음수의 입력에 대해 0을 출력하고 양수의 입력에 대해 양수를 출력하는 특징을 가지며, 이에 따라 기울기 소실 문제를 피하며 계산 효율이 높은 장점이 있다. 손실 함수는 예측 값과 실제 값 차이의 제곱에 대한 평균값을 이용하는 회귀 손실 함수인 mean squared error를 사용하였다. 최적화 함수는 손실함수의 값이 가장 낮은 지점을 찾기 위해 사용되며 adam을 사용하였다. layer의 뉴런에 해당하는 파라미터인 units과 과적합 방지를 위해 사용되는 dropout rate, 그리고 훈련 데이터 셋의 전체를 1회 학습할 때 작은 묶음으로 나누어 순차적으로 학습을 진행하기 위한 batch size는 모델의 예측 성능 최적화하기 위한 값을 찾기 위해 keras tuner의 random search를 이용하여 하이퍼 파라미터 탐색을 진행하여 모델을 설계하였다. 하이퍼 파라미터 탐색 결과는 표 2와 같다.

3-2 LSTM 기반의 예측 모델

인공신경망의 알고리즘 중 하나인 순환신경망(RNN; recurrent neural network)은 이전 정보를 지속적으로 참조하여 학습하는 알고리즘이다. RNN은 최근 정보일수록 예측을 위한 학습에 더 반영하도록 설계가 되어있다. 이러한 설계 방식으로 이전의 정보를 제대로 활용하지 못하는 문제가 있다. LSTM은 RNN의 문제를 해결하기 위해 메모리 셀을 추가하여 이전의 정보를 오래 기억할 수 있으며, 긴 시퀀스 데이터의 처리가 가능하다.

$PM_{2.5}$ 예측을 위해 순환신경망의 한 종류인 LSTM 알고리즘을 사용하여 설계를 진행하였다. DNN과 동일하게 2개의 hidden layer를 사용하였으며, 하루의 정보를 기억하여 예측 학습에 사용하도록 time stem의 값을 24로 설정하여 시퀀스를 구성하였다. 활성화 함수의 경우, ReLU는 0 이상의 값을 출력함에 따라 시퀀스를 통해 과거의 값들을 반복적으로 재사용하는 LSTM에 적용할 경우 출력 값이 발산할 확률이 높다. 따라서 안정적인 출력 값을 얻기 위해 −1 ~ 1사이의 값을 출력하는 tanh을 사용하였다. 손실함수와 손실함수의 최적화를 위한 함수는 DNN과 동일하게 mean squared error와 adam을 사용하였다.

그 외의 파라미터인 units, dropout rate, batch size는 DNN과 동일하게 하이퍼 파라미터 탐색을 통해 모델의 최적화를 위한 변수 값 도출하여 설계하였다.

Ⅳ. 모델 성능 평가 및 분석

4-1 성능 평가

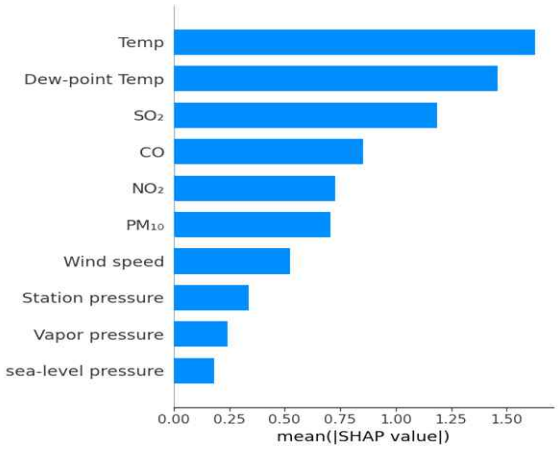
DNN과 LSTM 알고리즘을 이용하여 설계한 $PM_{2.5}$ 예측 모델의 성능 평가를 위해 test set을 이용하여 2023년의 $PM_{2.5}$ 예측을 진행하였다. 모델들의 성능을 평가하기 위해 여러 가지 지표가 사용된다. 그 중 대표적으로 정확도와 오차에 대한 지표들을 사용한다. 정확도는 실수 값의 예측 결과를 실제 값과 비교할 경우, 예측 모델의 정확한 평가가 이루어 질 수 없다. 따라서, 예측 값과 실제 값의 차이를 표현하는 방법 중 RMSE를 사용하였다. 그리고 실제 예보에 사용하고 있는 AQI를 기준으로 정확도를 확인하였다. AQI는 미세먼지를 특정 농도 값을 기준으로 '좋음', '보통', '나쁨', '매우 나쁨'의 4가지 범주로 구분하며, 해당 범주에 대해 예측 값의 결과가 실제 값의 범주를 비교하여 정확도를 확인하였다. 또한 '나쁨' 범주의 기준인 $36\mu g/m^3$ 농도 값 미만의 저농도와 $36\mu g/m^3$ 농도값 이상의 고농도에 대한 정확도를 확인하였다. 표 3은 DNN, LSTM 예측 모델의 예측 성능에 대한 평가 지표 값이다.

DNN 알고리즘을 이용한 $PM_{2.5}$ 예측 모델의 경우, 4.346의 RMSE 값을 확인하였으며, 정확도는 85.99%의 값을 확인하였다. 저농도의 경우, 98.73%의 정확도를 보였으며, 고농도의 경우, 78.95%의 정확도를 보였다. LSTM 알고리즘을 이용한 예측 모델의 경우, 3.99의 RMSE 값을 보였으며, 87.51%의 정확도를 확인하였다. 저농도의 경우, 98.85%의 정확도를 보였으며, 고농도는 80.43%의 정확도를 보였다.

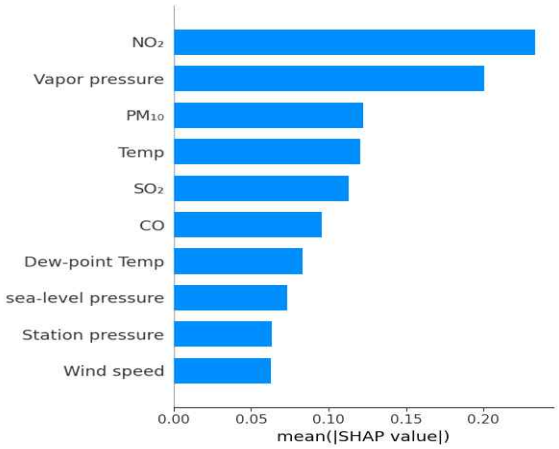
LSTM 알고리즘을 이용한 예측 모델이 DNN 알고리즘을 이용한 예측 모델보다 더 나은 성능을 보였다. 그러나 두 모델이 공통적으로 고농도의 예측에 있어 98% 수준의 저농도 정확도와 비교하였을 때 80% 수준의 정확도를 보였다. 미세먼지의 경우, 고농도에 해당하는 '나쁨' 이상의 수치에 대한 예측 정확도가 중요하다. 따라서 고농도 이상의 정확도 개선이 필요하다.

표 3. 예측 성능
Table 3. Prediction performance.

indicator	DNN	LSTM
RMSE	4.346	3.99
accuracy	85.99% (6820 / 7931)	87.51% (6920 / 7908)
low concentration accuracy (< $36\mu g/m^3$)	98.73% (6780 / 6867)	98.85% (6766 / 6845)
high concentration accuracy ($36\mu g/m^3 \leq$)	78.95% (840 / 1064)	80.43% (855 / 1063)



(a) DNN model



(b) LSTM model

그림 2. SHAP value 기반 summary plot
Fig. 2. SHAP value based summary plot.

4-2 예측 모델 분석

두 예측 모델의 결과 도출에 변수들의 영향력을 확인하기 위해 SHAP을 사용하여 분석을 진행하였다. 딥러닝 모델의 분석을 위해 DeepExplainer를 사용하였으며, 변수들이 $PM_{2.5}$의 예측에 얼마나 기여했는지에 대한 SHAP value를 측정하였다. SHAP value를 통한 기여도는 예측 결과에 긍정적인 영향과 부정적인 영향으로 구분하여 수치로 제공되며, 긍정적인 영향은 예측 값의 증가에 대한 기여로 해석할 수 있고 부정적인 영향은 예측 값의 감소에 대한 기여로 해석할 수 있다. DNN 예측 모델과 LSTM 예측 모델의 $PM_{2.5}$ 예측에 대한 SHAP value로 확인한 변수들의 기여도 순위는 그림 2와 같다.

각각의 예측 값에 영향을 준 변수들의 SHAP value를 절대값의 평균으로 순위를 정하였다. SHAP value의 절대 값으로 확인할 수 있는 것은 $PM_{2.5}$의 예측에 있어 각 변수들이 얼마 만큼의 기여를 했는지 확인하기 위함이다. DNN 예측 모델의 경우, 예측 값의 기여도가 가장 높은 변수는 온도이며, 기여도가 가장

낮은 변수는 해면기압이다. 기상데이터 중 온도와 이슬점온도가 SHAP value 1.5 이상의 높은 기여도를 보였으며, 풍속, 현지기압, 증기압, 해면기압은 0.75 미만의 낮은 기여도를 보였다. 대기오염물질 데이터의 경우, SO_2, CO, NO_2, PM_{10} 순으로 온도와 이슬점온도 다음으로 높은 기여도를 보였다. LSTM 예측 모델의 경우, NO_2가 가장 높은 기여도를 보였으며, 풍속이 가장 낮은 기여도를 보였다. 기상데이터 중 증기압이 0.2 이상의 가장 높은 기여도를 보였으며, 온도가 0.1 이상의 기여도 높은 기여도를 보였다. 그 외 이슬점온도, 해면기압, 현지기압, 풍속이 0.1이하의 기여도로 전체 데이터 중 기여도가 낮은 변수들임을 확인하였다. 대기오염물질 데이터의 경우 0.1 이상의 기여도를 보였다.

예측 모델이 예측 값을 도출하는 과정에서 각 변수들이 어떤 영향을 주었는지 확인하기 위해 그림 3과 같이 dot plot을 확인하였다. Dot plot에서 SHAP value가 음수인 영역은 예측 값을 감소시켰다는 것을 의미하며, 양수인 영역은 예측 값을 증가시켰다는 것을 의미한다. Feature value는 변수의 값이 크기를 표현한다. DNN 예측 모델에서 온도의 경우, SHAP value의 음수 영역은 붉은색 점들이 많이 분포하고 양수 영역에서는 파란색 점들이 많이 분포하는 것을 확인할 수 있다. 이는 온도가 높을 경우 예측 값을 감소시키는 영향을 주며, 온도가 낮을 경우 예측 값을 증가시키는 영향을 준다고 해석할 수 있다. 전체 데이터 중 온도, 풍속, 현지 기압이 값이 높을 경우 예측 값을 감소시키는 영향을 주었으며, 낮은 값일 경우 예측 값을 증가시키는 영향을 주었다. 나머지 변수들은 값이 낮을 경우 예측 값을 감소시키는 영향을 주며, 높은 값일 경우 예측 값을 증가시키는 영향을 주었다. LSTM 예측 모델에서 NO_2의 경우, 값이 높을 때 예측 값의 양방향으로 영향을 주며, 값이 낮을 때 예측 값을 감소시키는 영향을 주었다. 전체 변수들 중 PM_{10}과 SO_2가 이와 비슷한 영향을 주는 것을 확인하였다. 증기압의 경우, 값이 높을 때 예측 값을 증가시키는 영향을 주었으며, 값이 낮을 때 예측 값을 감소시키는 영향을 주었다. 이와 비슷한 영향력을 보여주는 변수들은 CO, 이슬점온도, 현지기압이 있다. 그 외의 온도, 해면기압, 풍속은 값이 높을 때 예측 값을 감소시키는 영향을 주었으며, 값이 낮을 때 예측 값을 증가시키는 영향을 주었다.

DNN 예측 모델과 LSTM 예측 모델의 결과를 비교하였을 때, 공통적으로 대기오염물질 데이터가 $PM_{2.5}$를 예측 하는 과정에서 기여도가 높은 것을 확인하였으며, 기상 데이터 중 온도가 두 모델의 예측 과정에서 기여도가 높은 것을 확인하였다. 변수들의 영향력에서는 온도, 풍속, 해면기압이 값이 높을 때 예측 값을 감소시키는 영향을 주며, 값이 낮을 때 예측 값을 증가시키는 영향을 주었다. 이와 반대의 영향력을 준 변수는 증기압, CO, 이슬점온도, 현지기압이다. NO_2, PM_{10}, SO_2의 경우 DNN 예측 모델과는 달리 LSTM 예측 모델에서는 값이 높을 때 예측 값의 양방향으로 영향을 주는 것을 확인하였으며, 이는 두 알고리즘의 학습 방식에 대한 차이로 해석할 수 있다.

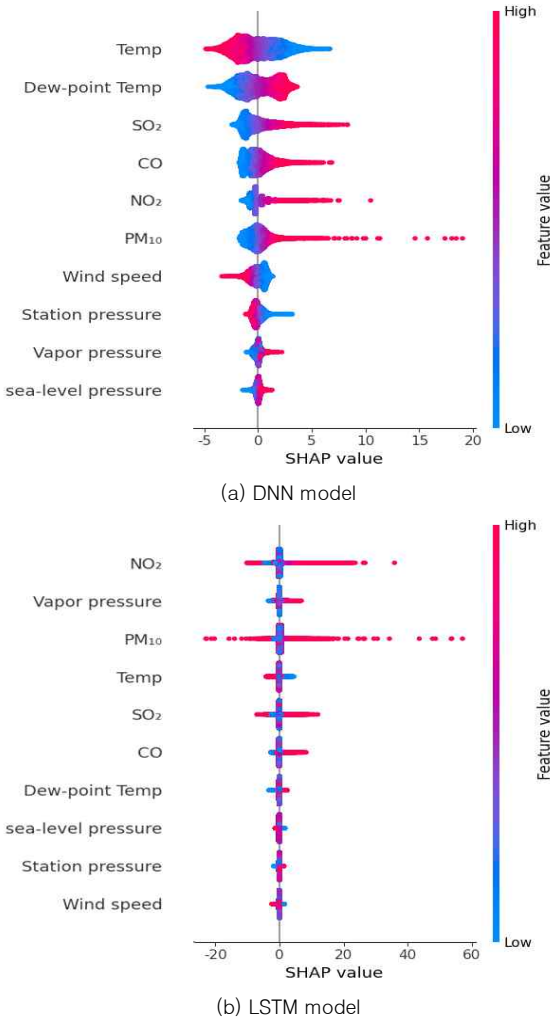


그림 3. SHAP value 기반 dot plot
Fig. 3. SHAP value based dot plot.

Ⅴ. 결 론

본 논문에서는 $PM_{2.5}$의 예측 모델에 변수들이 어떠한 영향을 주는지 확인하기 위해 SHAP을 이용하여 분석을 진행하였다. 딥러닝 알고리즘 중 DNN 알고리즘과 LSTM 알고리즘을 사용하여 $PM_{2.5}$를 예측하기 위한 두 개의 모델을 구축하였다. 예측 모델의 학습을 위한 데이터는 천안시와 오창읍에서 2019년부터 5년간 수집한 기상데이터와 대기오염물질 데이터를 사용하였으며, 상관분석을 통해 학습 및 테스트 데이터를 구성하였다. 모델의 성능 평가는 오차에 대한 지표인 RMSE를 사용하였으며, 정확도는 AQI를 참고하여 각각의 범주에 대한 정확도를 알아보았다. RMSE로 확인한 결과, DNN 예측 모델이 4.346, LSTM 예측 모델이 3.99로 LSTM이 더 나은 성능을 보였으며, 정확도의 경우, DNN 예측 모델이 85.99%, LSTM 예측 모델이 87.51%로 LSTM이 더 나은 성능을 보였다. SHAP을 이용한 변

수들의 영향력을 분석한 결과, 두 모델에 대해 공통적으로 대기오염물질 데이터가 $PM_{2.5}$를 예측하는 과정에서 기여도가 높은 것을 확인하였으며, 기상 데이터 중 온도가 두 모델의 예측 과정에서 기여도가 높은 것을 확인하였다. 변수들의 영향력에서는 온도, 풍속, 해면기압이 값이 높을 때 예측 값을 감소시키는 영향을 주며, 값이 낮을 때 예측 값을 증가시키는 영향을 주었다. 이와 반대의 영향력을 준 변수는 증기압, CO, 이슬점온도, 현지기압이다. NO_2, PM_{10}, SO_2의 경우 DNN 예측 모델과는 달리 LSTM 예측 모델에서는 값이 높을 때 예측 값의 양방향으로 영향을 주는 것을 확인하였으며, 이는 두 알고리즘의 학습 방식에 대한 차이로 해석할 수 있다. 향후, 다양한 모델의 해석 기법을 적용하여 각 모델의 특징과 모델에 따른 변수들의 특징을 확인할 예정이다.

Acknowledgments

This paper was supported by the Education and Research Promotion Program of KOREATECH in 2023.

References

[1] Korea Disease Control and Prevention Agency. Health effects of fine dust [Internet]. Available: https://www.kdca.go.kr/contents.es?mid=a20205030301.

[2] D. S. Kim and K. S. Ban, *Nearly Everything about the Fine Dust*, Seoul : Prisma, 2019.

[3] F. Keith, D. S. Krantz, R. Chen, K. M. Harris, C. M. Ware, A. K. Lee, … , S. S. Gottlieb, “Anger, hostility, and hospitalizations in patients with heart failure,” *Health Psychology*, Vol. 36, No. 9, pp. 829-838. Sep. 2017. DOI: 10.1037/hea0000519.

[4] E. J. Bang and Y. H. Choi, “Recent understanding in particular matter-mediated aging and age-related diseases,” *Journal of Life Science*, Vol. 34, No. 1, pp. 68-77, Jan. 2024. DOI: 10.5352/JLS.2024.34.1.68.

[5] World Health Organization (WHO), Health effects of particulate matter: Policy implications for countries in eastern europe, caucasus and central asia [Internet]. Available: https://iris.who.int/handle/10665/344854.

[6] H. J. Choi, “The effect of fine dust risk perception on indoor and outdoor tourists: focusing on planned behavior theory (TPB),” *Journal of Korea Entertainment Industry Association*, Vol. 17, No. 8, pp. 25-37, Dec. 2023. DOI: 10.21184/jkeia.2023.12.17.8.25.

[7] T. G. Kwon, The effect of atmospheric environment on the fan attendance in KBO League : focusing on fine dust, M. A. dissertation, Hanyang University, Republic of Korea, 2019. Retrieved from https://www.riss.kr/link?id=T15035580.

[8] D. H. Kim and H. B. Kim, “Perception of participants in outdoor physical activity for particulate matter : focusing on the university students,” *The Korean Journal of Sport*, Vol. 18, No. 1, pp. 369-378, Mar. 2020. Retrieved from https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002573513.

[9] J. Li, C. Hua, L. Ma, K. Chen, F. Zheng, Q. Chen, … , Y. Liu, “Key drivers of the oxidative potential of PM2.5 in Beijing in the context of air quality improvement from 2018 to 2022,” *Environment International*, Vol. 187, May. 2024. DOI: 10.1016/j.envint.2024.108724.

[10] X. Gao, Z. Ruan, J. Liu, Q. Chen, and Y. Yuan, “Analysis of atmospheric pollutants and meteorological factors on PM2.5 concentration and temporal variations in Harbin,” *Atmosphere*, Vol. 13, No. 9, Sep. 2022. DOI: 10.3390/atmos13091426.

[11] Z. Guo, X. Wang, and L. ge, “Classification prediction model of indoor PM2.5 concentration using CatBoost algorithm,” *Frontiers in Built Environment*, Vol. 9. Jul. 2023. DOI: 10.3389/fbuil.2023.1207193.

정 용 진 (Yong-Jin Jung)

2016년 2월 한국기술교육대학교 전기전자통신공학과 (공학석사)
2018년 3월 ~ 현재 한국기술교육대학교 전기전자통신공학과 박사과정
2014년 2월 공주대학교 전기전자제어공학부 전자공학·나노정보공학전공 전자공학트랙 공학사
※관심분야 : 미세먼지 예측, 기계 학습, 인공신경망, 심층신경망

오 창 헌 (Chang-Heon Oh)

1988년 2월 한국항공대학교 항공통신공학과 (공학사),
1990년 2월 한국항공대학교 항공통신정보공학과 (공학석사)
1996년 2월 한국항공대학교 항공전자공학과 (공학박사),
1990년 2월 ~ 1993년 8월 한진전자(주) 기술연구소 전임연구원
1993년 10월 ~ 1999년 2월 삼성전자(주) CDMA 개발팀 선임연구원
2006년 8월 ~ 2007년 7월 방문교수(University of Wisconsin-Madison)
1999년 3월 ~ 현재 한국기술교육대학교 전기전자통신공학부 교수
※관심분야 : 무선/이동통신, IoT, 기계학습 기반 통신시스템