



논문

대기환경연구소 기반 머신러닝 앙상블 모델과 칼만 필터를 통한 PM_{2.5} 예측 성능 향상 연구

A Study on Improving PM_{2.5} Prediction Performance through Machine Learning Ensemble Models and Kalman Filter Based on Air Quality Research Center

이성일[†], 이동훈[†], 송호종*, 안준영¹⁾, 박승명¹⁾, 이재윤¹⁾

명지대학교 환경에너지공학과, ¹⁾국립환경과학원 대기환경연구소

Seong-il Lee[†], Dong Hun Lee[†], Hyo-Jong Song*, Joon-Young Ahn¹⁾,
Seung-Myung Park¹⁾, Jaeyun Lee¹⁾

Department of Environmental Engineering, Myongji University, Yongin, Republic of Korea

¹⁾Climate and Air Quality Research Division, National Institute of Environmental Research, Incheon, Republic of Korea

접수일 2025년 2월 4일
수정일 2025년 5월 21일
채택일 2025년 5월 23일

Received 4 February 2025

Revised 21 May 2025

Accepted 23 May 2025

[†]These two authors contributed equally to this work as co-first authors.

*Corresponding author

Tel : +82-(0)31-330-6693

E-mail : hjsong@mju.ac.kr

Abstract This study aims to improve the prediction accuracy of PM_{2.5} concentrations by applying ensemble machine learning techniques and the Kalman Filter (KF), using detailed air pollutant data collected from the Air Quality Research Center, which are not available through general monitoring networks. The study evaluates the contribution of high-resolution chemical species data to PM_{2.5} prediction and the effectiveness of post-processing techniques in enhancing model performance. Three individual models—Linear Regression, Random Forest (RF), and XGBoost—were trained, evaluated, and combined to construct an ensemble model. To capture both linear and nonlinear data characteristics, Linear Regression was used alongside RF and XGBoost. Model training, validation, and testing were conducted using air quality data collected between 2019 and 2022 from both a general monitoring site and the Air Quality Research Center in Bulgwang-dong, Seoul. Permutation Feature Importance was applied to identify and remove variables that negatively affected prediction accuracy, thereby improving model efficiency. Seasonal performance was also evaluated to reflect Korea's climate characteristics. Additionally, the KF was employed as a post-processing method to further enhance predictive performance. The final ensemble model, built from the most optimal individual models, demonstrated more stable and accurate predictions than any single model. After applying the KF, all performance metrics improved. RMSE and MAE decreased, and positively biased NMB values were slightly corrected. In 12-hour forecasts, the baseline model's R² dropped below 0.6 after 7 hours and reached 0.46, while the KF-applied model maintained R² above 0.6 until 10 hours and reached 0.56. Similar improvements were seen in 48-hour forecasts. This study demonstrates the value of combining machine learning and data assimilation using high-resolution atmospheric data and provides a foundation for future air pollution forecasting and policy planning.

Key words: Particulate matter 2.5, Ensemble, Kalman Filter, Machine Learning, Air pollution intensive monitoring stations

1. 서론

대기오염물질은 인간의 건강, 농업 및 생태계 생산성뿐만 아니라 기후 변화, 오존 감소, 대기 산화력과

같은 범지구적인 변화 문제에 중대한 영향을 미친다 (Baklanov *et al.*, 2016). 그중 PM_{2.5}는 공기역학적 직경이 2.5 μm 이하인 물질을 의미한다 (Yang *et al.*, 2023). PM_{2.5}에 장기간 노출 시 심혈관 질환(CVD) 중 심근경

색(MI)의 위험성이 높아지며 뇌졸중 발생량이 증가한다(Krittanawong *et al.*, 2023; Zhu *et al.*, 2021; Yuan *et al.*, 2019). 또한, 만성 폐질환(COPD)과 같은 호흡기 질환과, 신경 퇴행과 같은 신경질환, 암을 발생시키고 이러한 질환 때문에 사망에 이르기까지 한다(Park *et al.*, 2021; Yu *et al.*, 2021; Shi *et al.*, 2020; Bowe *et al.*, 2019; Kioumourtzoglou *et al.*, 2016). 특히 대기오염은 사람이 많이 밀집된 대도시의 경우 더 두드러지며 도시에서 발생한 오염물질은 도시 내에서 국한되지 않고 상당히 먼 거리까지로 이동되어 넓은 지역의 오염에 기여할 수 있다(Baklanov *et al.*, 2016). 따라서 정확한 예측을 통해 미리 예방하고 분석을 통해 원인 물질에 대해 파악할 필요가 있다.

머신러닝(Machine Learning)의 기술이 발전함에 따라 의학(Rahmani *et al.*, 2021), 스마트 시티(Heidari *et al.*, 2022), 무선랜 성능개선(Szott *et al.*, 2022) 등 다양한 분야에서 적용되어 사용되고 있으며 환경 분야 또한 이에 해당한다(Liu *et al.*, 2022; Zhong *et al.*, 2021). 최근 대기질 예측 연구에서는 머신러닝과 딥러닝을 활용하여 PM_{2.5} 농도를 추정하는 다양한 방법이 개발되어왔다(Pak *et al.*, 2020; Suleiman *et al.*, 2019). 딥러닝 모델을 적용하여 시공간적 상관관계를 고려한 예측 기법이 제안되었으며(Pak *et al.*, 2020; Xiao *et al.*, 2020), 신경망을 활용하여 단기적인 PM_{2.5} 농도를 예측하는 연구도 진행되었다(Pérez *et al.*, 2000). 또한, 머신러닝 기반 연구에서는 랜덤 포레스트(Random Forest), XGBoost 등 다양한 기법을 적용하여 모델 성능을 비교·분석하며 최적의 조합을 찾고자 하였다(Vignesh *et al.*, 2023; Suleiman *et al.*, 2019). 최근에는 다양한 기상 및 환경 변수를 고려한 예측 모델이 개발되고 있으며, 특히 앙상블 기법을 적용하여 예측 성능을 향상시키는 연구가 이루어지고 있다. 그러나 기존 연구들은 복잡한 딥러닝 모델의 해석 가능성이 낮고, 장기 예측이 어렵다는 문제가 있으며, 예측의 불확실성을 줄이기 위한 자료 동화 기법이 충분히 고려되지 않았다. 또한, 대기환경연구소 성분 자료와 같은 세부적인 입자 성분 정보를 충분히 활용하지 못해 모델의 정확도 개

선에 어려움이 있다. 이에 본 연구에서는 이러한 한계를 해결하고자 대기환경연구소 성분 자료와 칼만 필터 기반 자료 동화 기법을 활용하여 PM_{2.5} 예측 모델의 성능을 향상시키는 것을 목표로 한다.

앙상블 기법은 여러 모델을 결합해 예측 성능을 향상하는 방법으로, 서로 다른 모델의 장점을 활용해 더 정확한 결과를 도출할 수 있으며 배깅(Bagging)과 부스팅(Boosting)과 같은 대표적인 방법을 통해 예측 오류를 줄이고 성능을 높일 수 있다. 이로 인해 회귀나 분류 문제에서 단일 모델보다 더 우수한 결과를 얻는 경우가 많다(Luo and Cheng, 2012; Melville and Mooney, 2005). 본 연구에서는 독립, 종속 변수가 선형관계일 때 높은 설명력을 가진 선형회귀, 분류 및 회귀 트리 알고리즘을 사용하는 여러 개의 결정 트리를 결합한 Random Forest (RF), 병렬화를 통해 효율성을 높인 XGBoost (XGB) (Vignesh *et al.*, 2023) 기법을 활용한다. 단, 선형회귀의 경우 과적합의 가능성이 높기 때문에 L1, L2와 같은 규제를 가해 모델을 추가로 구성하도록 한다(García-Nieto *et al.*, 2021). 여러 가지 모델들로 앙상블 모델을 구축하여 각 기법들이 가진 특징점을 활용할 수 있으면서 동시에 데이터가 선형성을 가져도, 비선형성을 가져도 모두 활용할 수 있는 모델을 구축하고자 하였다.

2019년부터 2022년까지의 수도권(서울) 불광동에 위치한 대기오염 대기환경연구소 데이터를 활용하였다. 해당 대기환경연구소는 국지적 배출 오염원의 영향을 평가하고, 도심지역에서 발생하는 고농도 대기오염 사례의 특성과 원인을 규명하기 위해 운영되고 있다(Yu *et al.*, 2018). 특히, 대기환경연구소 데이터는 일반 측정소와 달리 탄소성분, 이온성분, 금속성분, 흡수계수 등의 다양한 항목을 포함하고 있어 더욱 다양한 입력 데이터를 활용할 수 있다. 현재 대한민국에서는 CMAQ (Community Multiscale Air Quality Modeling System)를 활용하여 미세먼지 예보를 수행하고 있으나, CMAQ는 주로 배출량과 기상장 정보를 기반으로 대기오염물질을 모의하고 있다. 이러한 접근 방식은 PM_{2.5}의 형성 및 변동에 영향을 미치는 다양한 입

자 성분, 예를 들어 탄소성분, 이온성분, 금속성분 등 세부적인 물리·화학적 특성을 충분히 반영하지 못할 가능성을 내포하고 있다. 이에 따라 CMAQ의 예측 정확도를 높이고자 관련 연구들이 꾸준히 진행되고 있으며, 내부 프로세스 개선, 후처리 보정, 자료 동화 기법 등의 다양한 방법들이 제안되고 있다(Kang *et al.*, 2024; Kim *et al.*, 2024; Cheng *et al.*, 2021). 그러나 복잡한 대기 화학 및 물리 과정의 완전한 반영과 실시간 고농도 예측의 정확도 향상이라는 측면에서는 여전히 한계가 존재하는 실정이다. 따라서 본 연구에서는 대기환경연구소에서 제공하는 다양한 입자 성분 데이터를 적극 활용하여, 이러한 상세한 환경 정보를 포함한 머신러닝 기반 PM_{2.5} 예측 모델의 적합성을 평가하고자 한다.

서울 지역의 미세먼지 측정망이 대기오염 변동을 일관되게 반영한다는 점은 여러 연구에서 확인된 바 있다(Yoon *et al.*, 2015; Ghim *et al.*, 2014). GIS 기반 PM₁₀ 실시간 서비스를 활용한 분석에서도 대기환경연구소와 일반 측정소의 데이터가 유사한 경향을 보이는 것으로 나타났다. 이는 도시대기측정망 데이터의 일관성을 뒷받침하며, 측정소 간 유사한 대기환경이 군집 분석을 통해 검증된 결과와도 일치한다. 이를 활용하여 PM_{2.5} 예측에 적합한 데이터를 분류하고 선별된 데이터를 활용하여 예측을 진행한 뒤 나타난 개선사항에 대해서는 후처리 기법 KF를 적용하여 예측 모델이 안정적인 성능을 유지할 수 있게 하였다.

본 연구는 머신러닝 기법을 활용한 PM_{2.5} 예측에 중점을 두고, 예측에 사용되는 적합한 데이터를 분류하기 위한 분석과 동시에 후처리 기법을 통해 모델 성능을 향상하기 위한 연구를 수행하였다. 2절에서는 활용한 데이터와 사용한 머신러닝 기법들에 대해 설명하

고 3절에서는 결과와 논의를 제공하며 4절에서는 마지막으로 연구에서 도출된 결론을 제시한다.

2. 연구 방법

2.1 활용 데이터

대한민국은 백령도, 수도권, 중부권, 호남권, 영남권, 제주도, 경기권, 충청권, 전북권, 강원권의 지역에 10개의 대기환경연구소를 운영하고 있으며 그중 수도권(서울) 불광동에 위치한 측정소의 데이터를 활용하였다. 서울 정책아카이브에서 제공하는 자료에 따르면 2024년 1월 기준 인구가 약 960만 명에 달하며 이는 인구가 500만 명 이상인 경우를 지칭하는 ‘Megacity’에 해당한다. 해당 측정망의 경우 국지적 배출 오염원의 영향 평가와 도심에서 발생하는 고농도 대기오염 측정을 위해 운영되고 있다(Yu *et al.*, 2018). 인구밀집도가 높은 도시임과 동시에 대기환경연구소는 일반대기 측정소와는 달리 더욱 다양한 오염물질에 대해 측정을 진행하고 있어 데이터 활용에 있어 더욱 정교한 모델을 만들 수 있다. 항목은 표 1과 표 2에 나타내었다.

데이터 기간은 2019년도 1월 1일부터 2022년 12월 31일까지이며 2019년도부터 2021년까지 모델이 학습하고 2022년 데이터를 활용하여 예측함으로써 1년간의 예측을 진행할 수 있도록 한다.

2.2 방법

2.2.1 선형회귀

선형회귀는 종속 변수와 하나 이상의 독립 변수 간의 관계를 설명하는 통계적 기법으로 데이터를 학습

Table 1. Measurement items of general air quality monitoring stations and air pollution intensive monitoring stations.

Monitoring station	Measurement
General	SO ₂ , CO, NO ₂ , O ₃ , PM ₁₀ , PM _{2.5} , etc.
Intensive	SO ₂ , CO, NO ₂ , PM ₁₀ , PM _{2.5} , O ₃ , Wind direction, Wind speed, Temperature, Humidity, Carbonaceous composition, Ionic composition, Metallic composition, Absorption coefficients, Scattering coefficients, Particle size distribution

Table 2. Measurement items of air pollution intensive monitoring stations.

Composition	Species
Carbonaceous	OC, EC
Ionic	SO ₄ ²⁻ , NO ₃ ⁻ , Cl ⁻ , Na ⁺ , NH ₄ ⁺ , K ⁺ , Mg ²⁺ , Ca ²⁺
Metallic	S, K, Ca, Ti, V, Pb, Cr, Cu, Mn, Fe, Ni, As, Zn, Se, Br
Absorption coefficients	370, 470, 520, 590, 660, 880, 950 nm
Scattering coefficients	450, 550, 700 nm
Particle size distribution	10~20,000 nm

할 때 예측 함수, 즉 선 또는 평면에서 실제 관측값(훈련 데이터)까지의 평균 수직 거리를 최소화하는 방식으로 최적의 적합성을 찾는다. 간단하고 선형적인 관계를 가진 데이터라면 적절한 방법이지만 데이터가 선형적인 관계를 가지지 않는다면 오히려 모델 결과의 신뢰성을 잃는다는 단점을 지니고 있다. 선형회귀의 기본적인 형태는 다음의 식과 같다.

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n + b_0$$

y = 종속 변수(예측하려는 값)

x = 독립 변수(입력 변수)

b_0 = 절편(intercept)

a_1, a_2, \dots, a_n = 가중치(계수)

선형회귀는 데이터의 선형성을 잘 반영하면서 비교적 모델이 간단하다는 장점을 가지고 있지만 과적합의 위험이 있다는 단점을 가지고 있다. 따라서 이를 규제하기 위해 L2 (Ridge), L1 (Lasso), L1 + L2 (ElasticNet) 정규화 기법을 사용하여 만든 모델을 추가하여 앙상블 모델을 구축할 수 있도록 한다.

L2 (Ridge)는 선형회귀 모델에서 과적합을 방지하고 모델의 일반화 성능을 향상시키기 위해 도입된 기법으로 L2 정규화 기법을 사용하여 계수들의 제곱합을 패널티 항으로 추가함으로써 손실함수를 최소화한다. 이때 패널티 항은 계수들을 0에 가깝게 축소하여 다중 공선성(multicollinearity)을 완화하고, 모델의 안

정성을 높이는 데 도움을 준다. 그러나 계수를 0으로 만들지는 않기 때문에 모든 변수가 모델에 포함된다. 따라서 다중 공선성이 있는 경우나 예측 정확도를 높이는 데 효과적이다.

L1 (Lasso)는 Ridge 회귀의 한계를 보완하고 모델 해석력을 향상시키기 위해 개발되었다. 계수들의 절댓값 합을 손실 함수에 추가하며, 이는 특정 계수를 정확히 0으로 만들어 변수 선택 기능을 제공한다. 이로 인해 불필요한 변수를 제거하고, 모델을 더욱 간결하게 만들어 해석 가능성을 높이는 데 유리하다. 단, 다중 공선성이 높은 경우에는 불안정할 수 있다.

ElasticNet은 L1 정규화(Lasso)와 L2 정규화(Ridge)를 가중합으로 결합하여 두 방식의 특성을 모두 고려하는 회귀 분석 기법이다. 이 방식은 L1 정규화의 특성인 변수 선택을 활용하면서, L2 정규화의 특성인 계수의 분산 축소를 적용하여 더 강력하고 안정적인 모델을 만든다. 변수 선택과 규제를 동시에 수행하므로 모델이 더 정확하고 해석 가능해진다는 특성을 지니고 있다.

2.2.2 Random Forest (RF)

RF (Random Forest)는 여러 개의 결정 트리를 결합하는 앙상블 학습 방식으로, 각 트리는 트리 알고리즘을 기반으로 학습된다. 이 모델은 부트스트랩 샘플링과 무작위 변수 선택(random feature selection)을 통해 다양한 트리를 생성하고, 이들로부터 다수결 투표(분류) 또는 평균화(회귀)를 수행하여 최종 예측값을 도출한다(Park *et al.*, 2019). 이러한 bagging 기반 구조는 단일 결정 트리에서 발생할 수 있는 과적합을 방지할 수 있으며, 특히 비선형적인 데이터에서도 높은 예측 성능을 보이는 장점이 있다. 또한, 적절한 무작위성을 주입함으로써 대수의 법칙에 따라 모델의 분산을 줄이고, 보다 안정적이고 정확한 예측이 가능하다(Ryu *et al.*, 2022). 기본적인 형태의 Random Forest는 그림 1에 나타났다.

2.2.3 XGBoost (XGB)

XGB는 병렬 처리를 통해 효율성을 높인 알고리즘

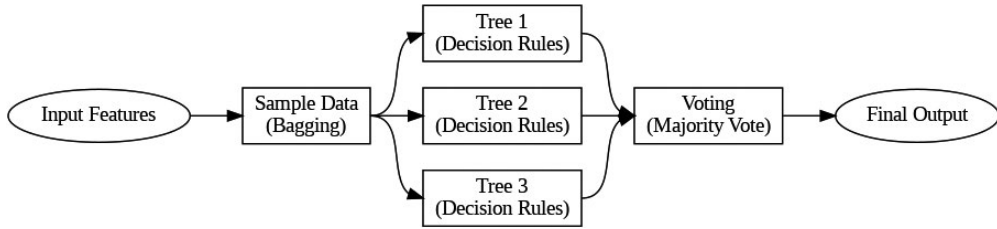


Fig. 1. The basic form of Random Forest.

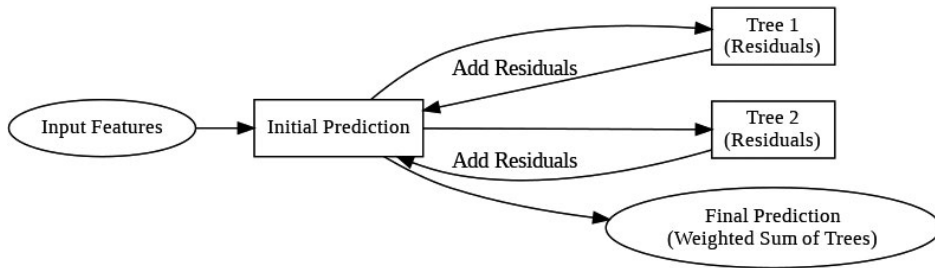


Fig. 2. The basic form of XGBoost.

으로 분류와 회귀에 둘 다 사용할 수 있다는 특징을 지니고 있다. 이 알고리즘은 Tianqi Chen이 제안한 그레디언트 부스팅 알고리즘을 기반으로 하며, 그레디언트 하강법을 사용하여 학습을 진행하기 때문에 부스팅 알고리즘으로 알려졌다. 특히 XGBoost가 이전 부스팅 알고리즘에 비해 더 정규화된 모델 형식을 사용하여 과적합을 효과적으로 제어한다고 설명한다 (Zamani Joharestani *et al.*, 2019). 또한 RF와 달리 정규화 기법을 적용하여 과적합 방지에 더욱 효과적이다. 기본적인 XGBoost의 형태는 그림 2에 나타났다.

2.2.4 Ensemble

앙상블 방법은 여러 예측 모델을 결합해 새로운 데이터에 대한 예측 성능을 개선하기 위한 기술로, 서로 다른 모델들의 장점을 결합해 더 정확한 결과를 도출하는 데 목적이 있다. 앙상블은 상호 보완적이며 높은 성능을 보이는 여러 학습 모델들을 활용해, 다양한 예측을 통합하여 개선된 결과를 얻는다. 이를 위해 학습 데이터에서 서로 다른 오류 패턴을 가진 예측 변수를 만들어 다양한 학습 모델의 결정을 결합하는 방식으

로 작동한다. 이처럼 구성된 앙상블은 단일 모델에 비해 회귀나 분류 문제에서 더 뛰어난 성능을 발휘하는 경우가 많다. 특히 위성 기반 PM_{2.5} 농도 예측에서 앙상블 모델이 개별 모델보다 교차 검증 결과에서 더 높은 R²과 낮은 RMSE를 보여, 앙상블 기법의 우수성이 실증적으로 확인되었다 (Xiao *et al.*, 2018). 앙상블 기법은 대표적으로 배깅 (Bagging), 부스팅 (Boosting)이 있는데, 이는 분산으로 인한 오류를 줄이고 예측 성능을 높이는 접근 방식이다 (Luo and Cheng, 2012; Melville and Mooney, 2005). 배깅 (Bagging)은 여러 학습자를 독립적으로 훈련하여 결과를 결합해 과적합을 줄이고 모델의 분산을 낮추며, 부스팅 (Boosting)은 학습자들이 순차적으로 오류를 보완하며 학습해 높은 예측 성능을 달성하지만 과적합 위험이 있다. 데이터의 변동성에 강하고 과적합을 줄여 예측의 안정성을 높이기 위해 앙상블 기법인 배깅 (Bagging)을 선택하여 진행하였다. 본 연구에서 만들어진 앙상블 모델의 구조는 그림 3에 나타났다.

본 연구에서는 선형회귀 (Linear Regression)를 활용하여 데이터 간의 선형 관계를 반영하고, RF와 XG

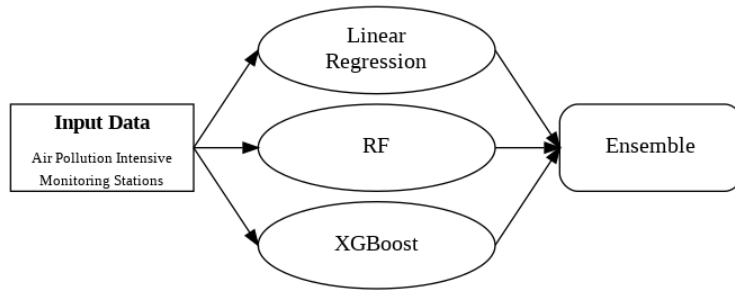


Fig. 3. The ensemble model proposed in the study.

Boost 기법을 통해 선형회귀로는 설명하기 어려운 비선형적 특성을 보완하였다. 이를 통해 선형성과 비선형성 모두를 효과적으로 고려할 수 있는 앙상블 모델을 구축하였다. 데이터의 복잡한 구조를 보다 정밀하게 반영하기 위해 다양한 기법을 고려할 수 있으나, 계산 자원 및 시간 효율성을 감안하여 본 연구에서는 선형회귀, RF, XGBoost를 기반으로 한 앙상블 기법을 최종적으로 선택하였다.

2.2.5 Feature Importance

머신러닝 기법은 다양한 분야에 활용됨에 따라, 모델의 해석 가능성을 높이기 위해 입력 변수들의 중요도를 파악하는 작업이 필수적으로 요구된다(Debeer and Strobl, 2020; Altmann *et al.*, 2010). 특히, 입력 변수의 상대적 중요도는 모델 해석뿐 아니라 의사결정 과정에서 핵심적인 정보를 제공한다.

Random Forest (RF)와 XGBoost는 자체적으로 특성 중요도 (Feature Importance)를 평가하는 기능을 제공하지만, 선형회귀 (Linear Regression)는 이러한 내장 기법이 부재하여 동일한 방식으로 중요도를 비교하기 어렵다. 이에 따라 본 연구에서는 모델 불가지론적 (model-agnostic) 기법을 활용하여 모든 모델에 대해 일관된 기준으로 특성 중요도를 평가하였다.

대표적인 모델 불가지론적 기법에는 SHAP (SHapley Additive exPlanations)와 LIME (Local Interpretable Model-agnostic Explanations)이 있으며, SHAP은 게임이론 기반의 접근으로 각 특성이 개별 예측값에 얼마나 기여했는지를 정량화하며, LIME은 국소적인 선형

모델을 학습시켜 특정 예측에 영향을 미친 특성을 설명한다. 이 두 기법은 모델 해석에 유용하지만, 연산량이 많고 구현 복잡성이 상대적으로 높다는 한계가 있다.

본 연구에서는 이러한 점을 고려하여, Permutation Feature Importance 기법을 적용하였다. 이 방법은 특정 특성의 값을 무작위로 섞은 뒤 모델의 예측 성능이 얼마나 감소하는지를 평가함으로써 중요도를 산정한다. Permutation Importance는 모든 모델에 적용 가능한 불가지론적 기법이며, 계산이 간단하고 모델 전체 성능 기반의 중요도를 제공하므로, 모델 예측 성능 중심의 본 연구 목적에 적합하고 연산 자원이 제한적인 상황에서도 효과적으로 활용 가능하다는 장점이 있다 (Khan *et al.*, 2025).

2.2.6 Kalman Filter (KF)

Kalman Filter는 두 가지 단계로 나뉘어 적용되며, 이를 ‘예측(Prediction) 단계’, ‘업데이트(Update)’ 단계로 구분할 수 있다. 예측 단계에서는 시스템의 이전 상태와 제어 입력을 바탕으로 다음 상태를 예측하고, 예측된 상태는 오차 공분산을 사용하여 신뢰도를 측정한다. 이후, 업데이트 단계에서 측정값을 반영하여 상태 추정을 보정하며, 오차 공분산을 갱신하여 추정값의 신뢰도를 향상시킨다.

2.2.6.1 Kalman Filter (KF) 예측(Prediction) 단계

예측 단계에서는 시스템의 이전 상태 \hat{x}_{k-1} 와 제어 입력 u_k 을 바탕으로 다음 상태 \hat{x}_k 를 예측한다. 이때

상태 전이 행렬 F_k 과 제어 입력 행렬 B_k 를 사용하여 다음 상태를 계산한다:

$$\hat{x}_{\bar{k}} = F_k \hat{x}_{k-1} + B_k u_k$$

또한, 오차 공분산 $P_{\bar{k}}$ 는 예측된 상태의 불확실성을 나타내며, 이 값은 이전 오차 공분산 P_{k-1} 과 시스템의 프로세스 잡음 Q_k 을 바탕으로 계산된다:

$$P_{\bar{k}} = F_k P_{k-1} F_k^T + Q_k$$

2.2.6.2 Kalman Filter (KF) 업데이트(Update) 단계

업데이트 단계에서는 실제 측정값 z_k 와 예측된 측정값 $H_k \hat{x}_{\bar{k}}$ 의 차이를 이용해 상태를 보정한다. 이 차이를 기반으로 칼만 이득 K_k 를 계산하며, 이는 예측된 상태를 측정값과 얼마나 결합할지 결정한다:

$$K_k = P_{\bar{k}} H_k^T (H_k P_{\bar{k}} H_k^T + R_k)^{-1}$$

이 후, 상태 추정값은 칼만 이득을 사용하여 보정되며, 다음과 같이 업데이트된다:

$$\hat{x}_k = \hat{x}_{\bar{k}} + K_k (z_k - H_k \hat{x}_{\bar{k}})$$

마지막으로, 오차 공분산 P_k 는 상태 추정값의 신뢰도를 나타내며, 칼만 이득을 사용하여 갱신된다:

$$P_k = (I - K_k H_k) P_{\bar{k}}$$

이와 같은 반복적인 예측과 업데이트 과정을 통해 칼만 필터는 점차적으로 시스템의 상태를 정확히 추정하게 되며, 특히 잡음이 포함된 측정값에서 신뢰할 수 있는 추정값을 도출할 수 있다. 이 필터는 그 효율성과 정확성 덕분에 다양한 응용 분야에서 널리 사용되고 있다(Zhang *et al.*, 2024).

본 연구에서는 예측값으로 대기환경연구소 데이터와 머신러닝 기법을 기반으로 만들어진 앙상블 모델을 통해 나온 결과를 활용하고 칼만 필터를 통해 보정된 결과를 분석하고자 했다. $\hat{x}_{\bar{k}}$ 는 앙상블 모델을 통해 나온 PM_{2.5} 예측값을 의미하며, \hat{x}_k 는 $\hat{x}_{\bar{k}}$ 값을 칼만 필터를 이용하여 보정된 최종값이다. 그러나 연구를 진행할 때, 예측 모델의 노이즈 공분산 Q_k 와 관측 기기의 노이즈 공분산 R_k 을 정확히 알 수 없는 한계가 있다. 관측기기는 물리적 특성, 보정 상태 등과 기상 상

태에 따라 오차가 발생하므로 명확한 노이즈 공분산 값을 알기 어렵다(Shi *et al.*, 2007). 또한, 직접 관찰할 수 없는 프로세스 노이즈도 정확한 노이즈 공분산을 알기 어려운 특성을 가진다. 이에 따라 KF를 적용할 때 필터 파라미터를 조정하여 가장 적절한 성능을 찾을 수 있다. 본 연구에서는 경험적인 실험을 통해 가장 적절한 값을 찾아 적용하였으며, 이때의 Q_k 값은 $1e^{-3}$, R_k 값은 $1e^{-1}$ 로 설정하였다. 다만, 임의로 설정된 값이 모델의 성능에 미칠 수 있는 영향을 충분히 인지하고 있으며, 후속 연구에서는 더욱 정교한 값 설정을 통해 KF의 성능을 개선할 필요가 있다(Welch, 1995).

2.2.7 평가 변수

예측 모델의 성능을 평가하기 위해, 예측 결과를 분석하고 다양한 통계적 평가지표를 활용할 필요가 있다. 본 연구에서는 예측 모델의 정확성을 평가하기 위해 RMSE, MAE, NMB, R²과 같은 통계적 평가지표를 사용하였으며, 미세먼지 등급 분류의 성능 평가는 ACC, POD, FAR, F1-Score를 통해 수행하였다. 이러한 평가지표들의 계산식은 표 3과 표 4, 그리고 그림 4에 제시되어 있다.

RMSE (Root Mean Square Error)는 예측값과 실제 값 사이의 차이(오차)를 제공한 후 평균을 내고, 다시 제곱근을 취한 지표로 예측 오차의 크기를 종합적으로 나타낸다. 이 지표는 오차가 클수록 제곱에 의해 더

Table 3. Statistical evaluation metrics.

Metric	Formula
RMSE	$\sqrt{\frac{1}{N} \sum (P_i - O_i)^2}$
MAE	$\frac{1}{n} \sum_{i=1}^n \hat{y}_i - y_i $
NMB	$\frac{\sum_{i=1}^n (\hat{y}_i - y_i)}{\sum_{i=1}^n y_i} \times 100$
R ²	$1 - \frac{\sum (P_i - O_i)^2}{\sum (P_i - \bar{O} + O_i - \bar{O})^2}$

Prediction \ Observation	Good	Moderate	Unhealthy	Very Unhealthy	
Good	a1	b1	c1	d1	I
Moderate	a2	b2	c2	d2	II
Unhealthy	a3	b3	c3	d3	III
Very Unhealthy	a4	b4	c4	d4	IV

Fig. 4. PM_{2.5} grade classification criteria.

Table 4. Statistical evaluation metrics.

Metric	Formula
ACC	$\frac{(a1 + b2 + c3 + d4)}{N} \times 100$
POD	$\frac{(IV)}{(III + IV)} \times 100$
FAR	$\frac{(II)}{(II + IV)} \times 100$
F1-Score	$2 \times \frac{\{POD \times (100 - FAR)\}}{\{POD + (100 - FAR)\}}$

크게 반영되기 때문에 이상치에 민감하며, 예측 모델의 정확도와 더불어 분산까지 고려하는 데 유용하다. RMSE 값이 작을수록 예측값이 실제값과 가까움을 의미하며, 단위는 원래 데이터와 동일하다.

MAE (Mean Absolute Error)는 예측값과 실제값 사이의 절대적인 차이의 평균을 나타내는 지표로, 예측값이 실제값으로부터 얼마나 떨어져 있는지를 직관적으로 표현한다. RMSE와 달리 오차를 제공하지 않기 때문에 이상치의 영향을 적게 받는다는 장점이 있으며, 예측값의 정확도를 평가할 때 자주 사용된다. 값이 작을수록 전반적인 예측 오차가 작음을 의미한다.

NMB (Normalized Mean Bias)는 예측값과 실제값 간의 평균적인 편향(Bias)을 정규화하여 나타낸 지표이다. 예측값과 실제값의 차이를 전체 실제값의 총합

으로 나누어 계산하며, 예측이 실제값보다 과대 혹은 과소 추정되는 경향을 정량적으로 평가할 수 있다. NMB가 0에 가까울수록 편향이 적고, 양의 값은 과대 예측(overestimation), 음의 값은 과소 예측(underestimation)을 의미한다.

R² (Coefficient of Determination) 회귀 분석 모델의 적합도를 나타내는 지표로 종속 변수의 총 변동 중에서 모델로 설명되는 변동의 비율을 나타내며 1에 가까울수록 모델이 데이터를 잘 설명한다는 것을 의미한다. 단, R²이 높은 값이라고 해서 반드시 좋은 예측 성능을 의미하지 않으므로 다른 지표들과 함께 해석하는 것이 중요하다.

ACC (Accuracy)는 모델이 전체 예측 중에서 실제 관측값과 일치하게 분류한 비율을 의미한다. 이는 모든 클래스(좋음, 보통, 나쁨, 매우나쁨)에 대한 전반적인 예측 능력을 포괄적으로 나타낸다.

POD (Probability of Detection)는 실제로 고농도 상태(즉, PM_{2.5} 농도가 36 µg/m³ 이상인 경우)에 있었던 사례 중, 모델이 이를 정확히 예측한 비율을 의미한다. 이는 고농도 상황에서 모델의 탐지 민감도를 측정하는 지표로 특히 고농도 미세먼지로 인한 건강 피해를 최소화하기 위해 중요한 역할을 하며, 누락 없이 고농도 사례를 잘 탐지했는지를 평가한다.

FAR (False Alarm Rate)은 모델이 고농도(“나쁨” 이상) 상태라고 잘못 예측한 비율로 정의된다. 실제로는

저농도 상태(‘좋음’ 또는 ‘보통’)였으나 모델이 고농도로 예측한 경우의 비율을 의미하며 모델이 얼마나 자주 불필요한 경고를 발생시키는지를 나타내며, 실질적 정책 결정 또는 대기질 경고 시스템의 신뢰도와 직결된다.

F1-Score는 POD와 Precision (정밀도)의 조화 평균으로, 고농도 예측의 민감도와 정확도 간의 균형을 평가하는 지표로 특히 POD와 FAR 간 균형이 중요할 때 유용하며, 고농도 상황 예측의 종합적 성능을 평가하는 데 적합하다.

3. 결과 및 고찰

3.1 단기(12H)예측: 데이터 특성별 모델 성능 비교

일반측정망과 대기환경연구소에서 수집된 데이터를 기반으로 다양한 조건을 조합하여 총 6개의 예측 모델을 구성하였으며, 이를 비교 분석하였다. 각 모델에 사용된 입력데이터 구성은 표 5에 정리하였으며 이에 따른 예측 성능은 그림 5와 그림 6에 시각화하였다. 모델 성능 평가는 2022년 한 해 동안의 시간별 PM_{2.5} 농도를 기준으로 하였으며 각 시점을 기준으로 12시간을 예측하는 방식으로 수행되었다. 매시간 시점을 기준으로 12시간 후까지의 PM_{2.5} 농도를 예측하고, 이 예측값들을 전체 시간 구간에 대해 평균하여 성능 지표를 도출하였다. 이를 통해 다양한 입력조건을 바탕으로 입력 데이터가 예측 정확도에 미치는 영향을 평가하고, 각 모델이 단기 예측에 있어 어떤 상

대적 성능을 보이는지 비교하였다.

평가지표인 RMSE와 MAE, R², NMB를 분석한 결과, ‘Ensemble V3’와 ‘Ensemble V3_FI’ 모델은 전반적으로 낮은 성능을 보였다. 이는 “입자 크기별 수농도” 데이터가 PM_{2.5} 예측에 기여하지 못하거나 오히려 불필요한 영향을 주었음을 시사한다. 반면, ‘AirKorea (일반대기성분)’ 모델은 RMSE와 MAE에서 가장 낮은 값을 보여 예측 오차가 가장 작았으며, 이는 단기 예측 관점에서 가장 양호한 성능을 나타낸 것으로 해석된다. 그러나 R² 값은 다른 모델에 비해 상대적으로 낮아 데이터의 변동성을 충분히 설명하지 못하고 있으며, NMB 지표에서는 초기에는 낮은 편향을 보였으나 예측 시간이 길어질수록 점차 양의 편향이 커지는 경향을 보였다. ‘Ensemble V2’ 및 ‘Ensemble V2_FI’는 오차 측면에서는 ‘AirKorea’보다 다소 높은 값을 보였으나, R²과 NMB에서 보다 우수한 성능을 보였다. 특히 ‘Ensemble V2_FI’는 시간에 따른 예측 편향을 줄이는 데 효과적이며, 예측값의 분산을 잘 설명하고 있는 것으로 나타났다. 또한, V2 및 V3 모델에 활용된 흡수계수(Absorption coefficient) 및 산란계수(Scattering coefficient)는 ‘AirKorea’ 모델에서 시간이 지남에 따라 편향이 증가하던 현상과 달리, 일정 시간 이후에는 편향이 다시 감소하는 양상을 보여, 해당 변수들이 장기적인 예측에서의 편향을 보정하는 데 기여할 수 있음을 시사한다.

그림 6은 PM_{2.5} 등급 분류 기준을 적용하여 각 모델의 예측 결과를 분류 성능 관점에서 평가한 것이다. 정확도(Accuracy, ACC)를 기준으로 보았을 때, ‘Air

Table 5. Data used for each model.

Model name	Used data
Airkorea	General ait pollutants
Ensemble V1	General ait pollutants, Carbonaceous composition, Ionic composition, Metallic composition
Ensemble V2	Ensemble V1 + Absorption coefficients, Scattering coefficients
Ensemble V2_FI	Apply permutation importance to Ensemble V2
Ensemble V3	Ensemble V2 + Particle size distribution
Ensemble V3_FI	Apply permutation importance to Ensemble V3

The representation of each component is shown in Table 3. ‘Airkorea’ is a component available from the Air Korea website.

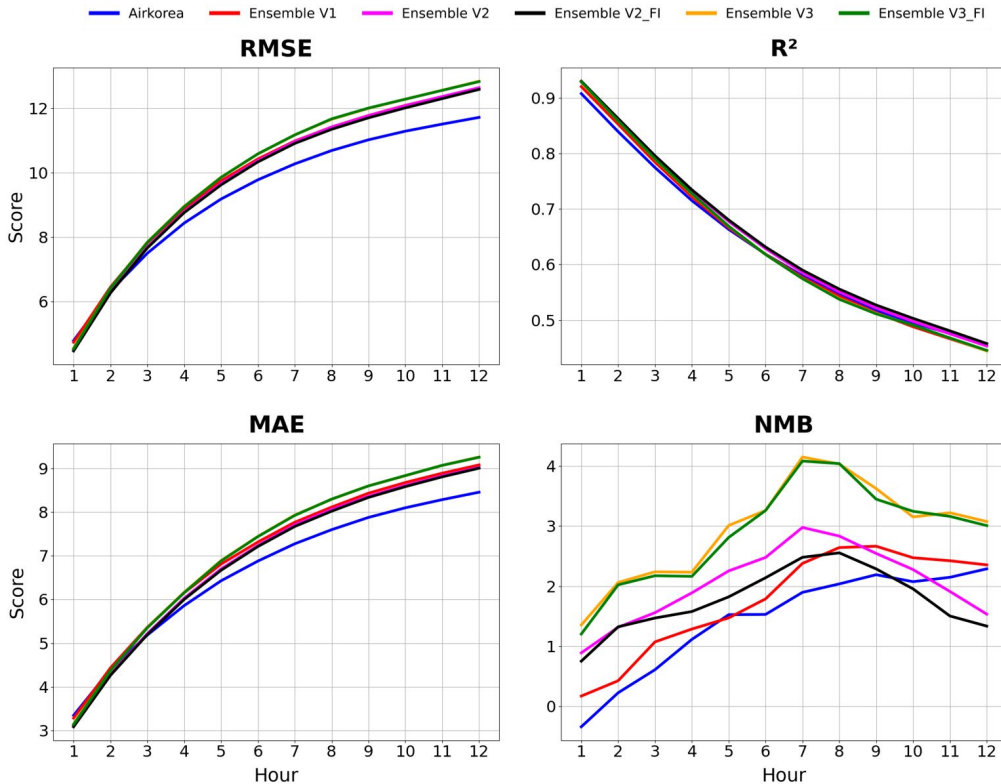


Fig. 5. Comparison of 12-hour $PM_{2.5}$ forecasting performance: model accuracy based on input data composition (including FI). This figure compares the 12-hour $PM_{2.5}$ forecasting performance of several models developed using different input data compositions, with and without feature importance (FI) integration. Performance metrics including RMSE, R^2 , MAE, and NMB are used to assess the accuracy for each forecast hour. The results visually demonstrate the impact of input data selection and FI application on model performance, with certain models showing improved accuracy when FI is incorporated.

Korea' 모델은 가장 높은 정확도를 보이며 전반적으로 우수한 분류 성능을 나타내는 것처럼 보인다. 또한 오경보율(FAR)에서도 비교적 낮은 값을 보여 준수한 성능을 나타낸다. 그러나 고농도 상황을 탐지하는 지표인 POD (Probability of Detection)에서는 매우 낮은 값을 보여, 고농도 $PM_{2.5}$ 상황을 제대로 예측하지 못하는 것으로 해석된다. 이러한 경향은 F1-Score에서도 반영되어, 'AirKorea' 모델은 전반적으로 낮은 F1-Score를 기록하였다. 이는 앞서 그림 5에서 분석된 바와 같이 R^2 값이 낮아 데이터의 변동성을 충분히 반영하지 못한 결과로 판단된다.

한편, POD가 가장 높은 성능을 보인 'Ensemble V3' 및 'Ensemble V3_FI' 모델은 전반적으로 예측값이 실

제값보다 과도하게 높은 양의 편향을 가지는 경향이 있으며, 이에 따라 오경보율(FAR)도 함께 높아지는 현상을 보였다. 이는 예측에 있어 과대 예측이 발생함을 의미한다.

FAR이 상대적으로 낮으면서 POD가 높은 성능을 보인 모델은 'Ensemble V2_FI'로, 예측의 균형 측면에서 우수한 결과를 보였다. 특히 시간에 따라 FAR이 다소 증가하는 경향을 보이나, POD와의 조화평균으로 계산되는 F1-Score에서 가장 높은 점수를 기록하여 전반적인 분류 성능이 가장 우수한 것으로 평가된다.

이러한 결과를 종합하면, $PM_{2.5}$ 의 농도 수준을 분류하고 예측하는 데 있어 데이터의 변동성을 잘 설명하고 예측 편향과 오차 간 균형이 가장 적절하게 이루어

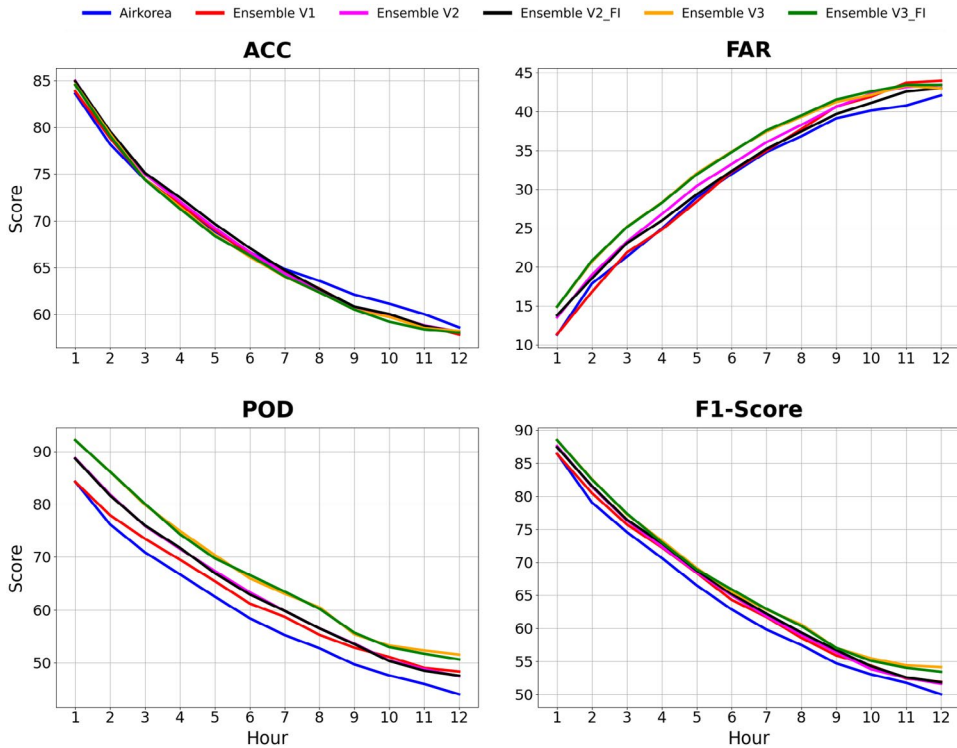


Fig. 6. Comparison of 12-hour PM_{2.5} grade classification performance: model accuracy based on input data composition (including FI). This figure presents a comparison of PM_{2.5} grade classification performance across various models using ACC, FAR, POD, and F1-score over 12 forecast hours. The models were developed with different input data compositions, with and without feature importance (FI). The figure highlights the effect of FI and input features on classification accuracy.

진 모델은 ‘Ensemble V2_FI’로 판단된다. 이는 일반대기성분만을 활용한 ‘AirKorea’ 모델보다 대기환경연구소에서 수집된 탄소, 이온, 금속성분과 흡수계수 및 산란계수 등의 데이터를 추가로 활용함으로써 예측 성능을 향상시킬 수 있음을 시사한다.

3.2 단기(12H)예측: 평가지표 기반 성능 분석과 앙상블 모델 최적화

최종적으로 선정된 모델인 ‘Ensemble V2_FI’는 다른 모델의 결과와 비교하였을 때 전반적으로 우수한 성능을 보였으나, 여전히 개선이 필요한 부분이 존재한다. 특히 그림 5에 나타난 ‘AirKorea (일반대기성분)’ 결과와 비교하였을 때, 예측값의 편향이 상대적으로 크게 나타나는 양상이 확인되었다. 이에 따라, 앙상블

모델의 성능을 더욱 향상시키기 위해 ‘Ensemble V2_FI’를 구성하는 개별 모델들의 예측 결과를 분석할 필요가 있으며, 이는 그림 7에 제시하였다.

모든 모델의 결과를 통합하여 산출한 앙상블 결과는 전반적으로 안정적인 예측 성능을 나타냈다. 그러나 NMB 지표에서는 대부분의 모델이 큰 변화 없이 유사한 경향을 보인 반면, XGBoost는 예측 시간이 증가할수록 급격히 양의 방향으로 치우치는 편향을 나타냈다. 이로 인해 RMSE, MAE, R² 지표에서도 상대적으로 낮은 성능이 나타났다. 이러한 분석을 바탕으로, 앙상블 모델의 최적화를 위해 편향이 과도하게 나타나는 모델인 XGBoost를 제외한 결과를 도출하여 성능의 변화를 분석하였으며, 이는 그림 8에 제시하였다.

그림 8은 기존 모델인 ‘Ensemble V2_FI’와 XGBoost

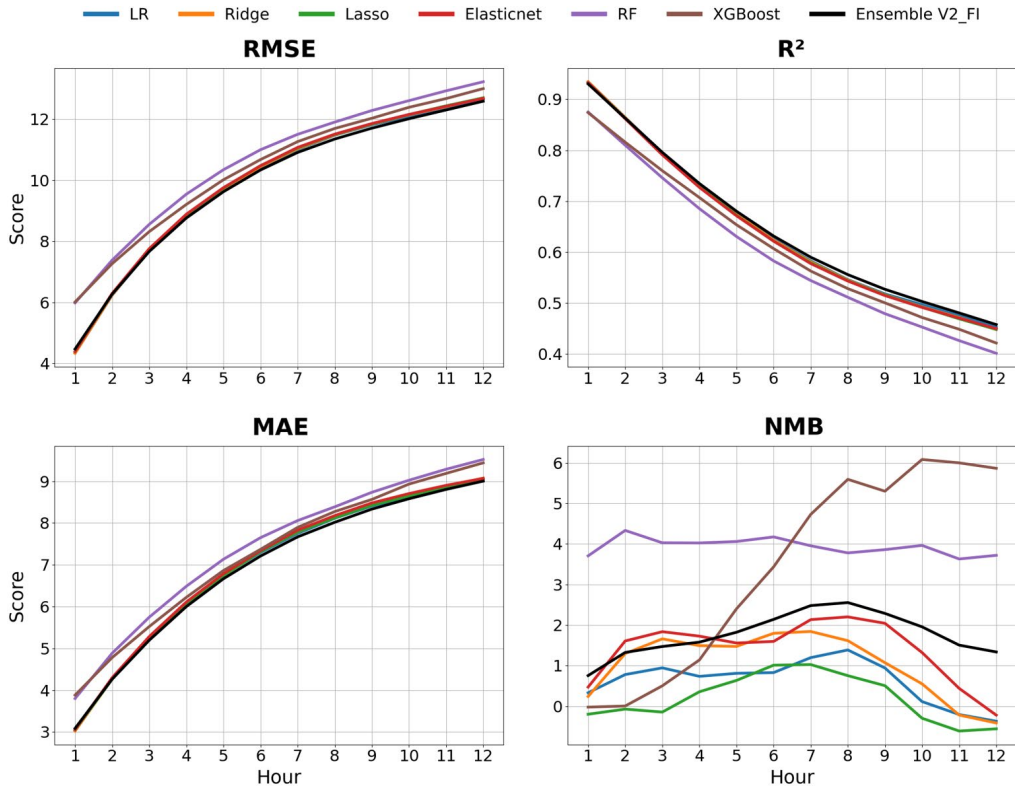


Fig. 7. Forecasting performance of Ensemble V2_FI and its base models for 12-hour $PM_{2.5}$ prediction. RMSE, MAE, R^2 , and NMB are used to evaluate and compare the accuracy of Ensemble V2_FI with its constituent models (LR, Ridge, Lasso, ElasticNet, RF, XGBoost).

결과를 제외한 ‘Ensemble V2_FI_Select’의 성능을 비교한 결과이다. RMSE, MAE, R^2 지표에서는 기존 앙상블 결과와 유사한 수준의 성능을 유지한 반면, NMB 지표에서는 기존의 과대모의가 일부 완화되어 전반적인 성능 향상이 확인되었다. 이는 연구에 사용된 데이터가 포함하고 있는 특성을 XGBoost 모델이 적절히 반영하지 못한 것으로 해석되며, 오히려 예측 성능 저하를 초래했음을 시사한다.

따라서 앙상블 모델의 성능을 극대화하기 위해서는 개별 모델의 성능뿐만 아니라, 데이터의 구조적 특성과 분포를 정확히 파악하는 과정이 선행되어야 한다. 특히 대기오염과 같이 시공간적 이질성과 비선형성이 내포된 데이터의 경우, 클러스터링과 같은 데이터 분류 기법을 통해 데이터의 특성을 세분화하고, 이에 적

합한 모델을 선택하거나 구성하는 방식이 필요하다. 이러한 전처리 과정을 통해 모델 간 상호보완적인 조합이 가능해지며, 앙상블 모델의 예측 성능을 보다 효과적으로 향상시킬 수 있다.

3.3 단기(12H)예측: 칼만 필터(Kalman Filter) 적용을 통한 성능 향상 분석

앞서 Permutation Feature Importance를 활용하여 주요 변수를 식별하고, 앙상블 모델의 구조를 최적화함으로써 예측 편향을 일부 개선하였으나, 여전히 오차 및 시스템적 편향이 잔존하는 한계가 확인되었다. 이러한 한계를 보완하기 위한 후처리 기법으로 본 연구에서는 칼만 필터(Kalman Filter, KF)를 적용하였다.

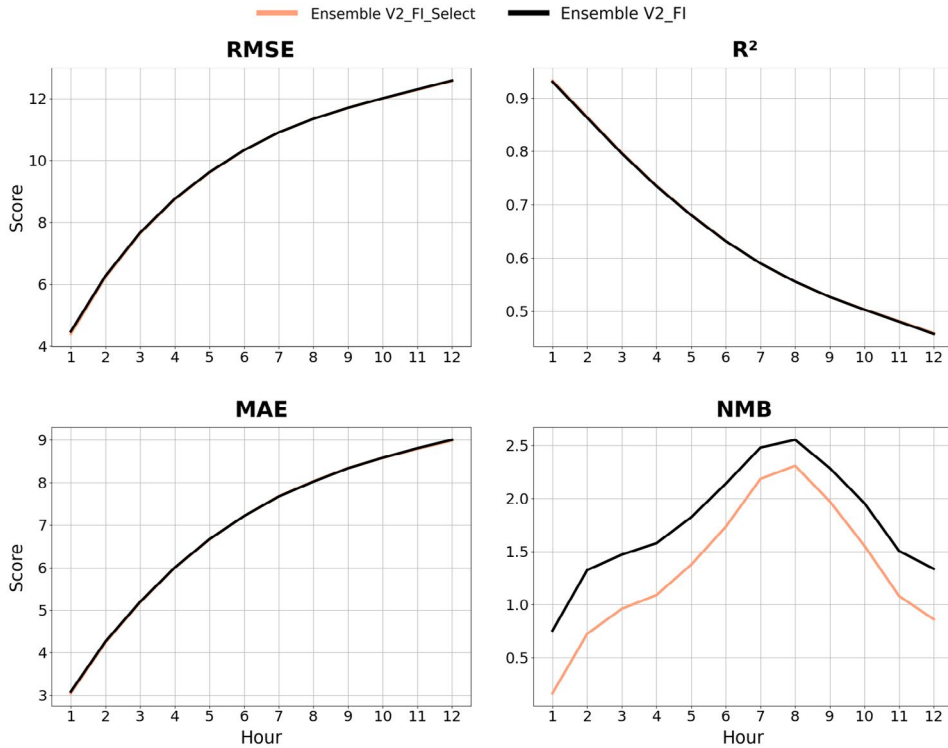


Fig. 8. Forecasting performance of the ensemble model with selected base models (excluding XGBoost). The comparison between Ensemble V2_FI and Ensemble V2_FI_Select highlights a reduction in NMB, suggesting that excluding XGBoost can mitigate bias without sacrificing overall accuracy.

칼만 필터는 자료동화(data assimilation)를 기반으로 한 재귀적 상태 추정 알고리즘으로, 시간에 따라 변화하는 시스템의 동적 상태를 예측하고 갱신하는데 효과적이다. 이 알고리즘은 시스템 모델을 통한 사전 예측값(prior estimate)과 잡음을 포함한 관측값(observation)을 통합하여, 상태 변수의 사후 추정값(posterior estimate)을 반복적으로 갱신함으로써 예측의 정확도와 신뢰성을 향상시킨다.

이와 같은 칼만 필터의 원리를 활용하여 기존 예측 결과에 후처리를 적용하였으며, 그 적용 전후의 성능 비교 결과는 그림 9에 제시하였다.

KF를 적용한 결과('Ensemble V2_FI Apply KF'), 예측값의 오차 및 편향이 감소함과 동시에 R² 지표에서도 성능 향상이 나타났다. MAE (Mean Absolute Error) 및 NMB (Normalized Mean Bias) 지표 역시 다소 개선

된 경향을 보였으며, 특히 RMSE (Root Mean Square Error)의 감소는 고농도 PM_{2.5} 상황과 같은 이상치에 대한 예측 정확도 향상으로 이어진 것으로 판단된다.

이러한 결과는 칼만 필터가 머신러닝 모델이 가지는 예측상의 불완전성을 보완하는 후처리 기법으로서 효과적으로 작용했음을 시사한다. 본 연구에서는 칼만 필터를 통해 단기 PM_{2.5} 예측에서의 성능 개선 가능성을 확인하였으며, 이 기법은 다양한 예측 모델에서의 오차 보정 및 신뢰도 향상을 위한 확장 가능한 방법론으로 활용될 수 있을 것으로 기대된다.

3.4 단기(12H)예측: 계절별 모델 성능 비교 분석
대한민국은 사계절이 뚜렷한 기후적 특성을 지니고 있어, 계절별 대기오염의 발생 양상 및 농도 수준에 차이가 존재한다. 따라서 앞서 머신러닝 앙상블, Per-

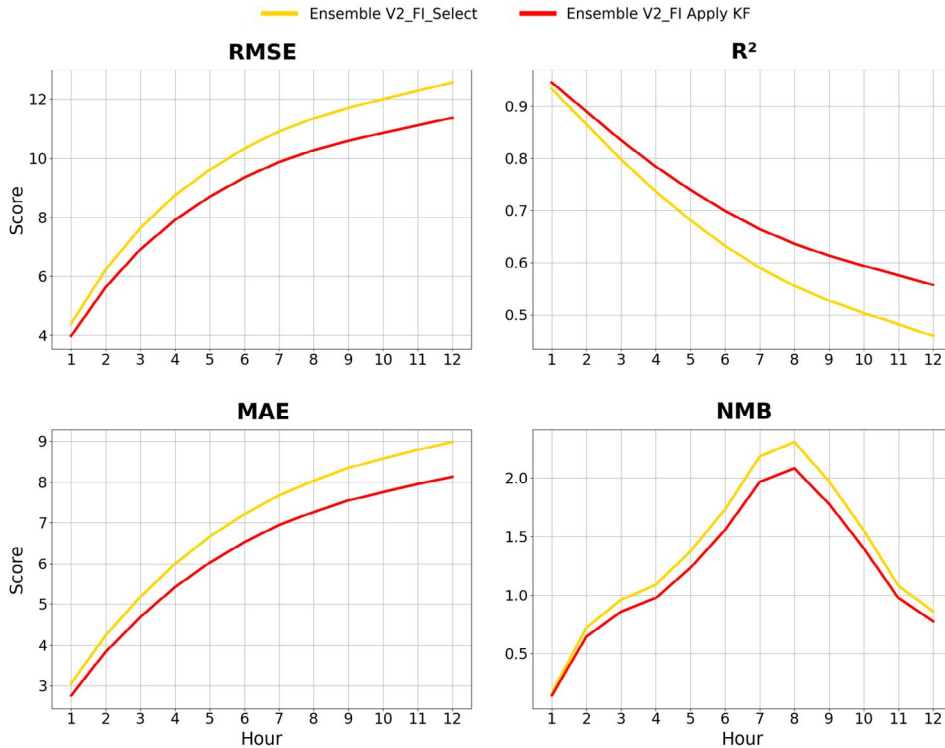


Fig. 9. Performance comparison before (gold) and after (red) applying Kalman Filter to the ensemble model. Kalman Filter application leads to improved prediction performance in RMSE, MAE, and R² metrics, contributing to enhanced model stability and reliability.

mutation Feature Importance, Kalman Filter가 적용된 최종 예측 결과를 계절별로 분류하여 분석하였다. 계절별 예측 성능은 그림 10에 제시하였다.

예측 오차를 나타내는 지표인 RMSE (Root Mean Square Error) 및 MAE (Mean Absolute Error)는 겨울과 봄에서 상대적으로 높은 값을 보였으며, 이는 두 계절이 고농도 PM_{2.5}가 자주 발생하는 미세먼지 계절관리제 시행 기간과 중첩되기 때문이다. 고농도 상황에서는 예측 오차가 크게 증가하는 경향을 보이며, 이로 인해 해당 기간의 예측 정밀도가 저하되는 것으로 판단된다. 반면, 가을은 전반적으로 안정적인 결과를 보여 타 계절 대비 상대적으로 낮은 오차를 나타냈다.

여름의 경우 RMSE와 MAE는 낮은 수치를 보였으나, NMB (Normalized Mean Bias)가 음의 방향으로 가장 크게 나타나고 R² 역시 가장 낮은 값을 기록하였다.

이는 여름철 빈번한 강수로 인해 실제 PM_{2.5} 농도가 급격히 낮아지는 현상이 반영되지 못한 결과로 해석된다. 한편, 겨울은 전반적인 오차 및 편향이 크게 나타났음에도 불구하고 R²은 높은 값을 보였는데, 이는 일정 수준의 분산 설명력이 확보되었음을 시사한다.

이러한 수치 기반 평가지표만으로는 실제 예보로서의 신뢰도를 충분히 판단하기 어려우므로, 본 연구에서는 PM_{2.5} 등급 분류표를 기반으로 예측 결과의 등급 분류 정확도를 추가적으로 분석하였다. 해당 분석 결과는 그림 11에 제시하였다.

등급 분류 성능 분석 결과, 봄과 가을은 상대적으로 양호한 정확도(ACC)를 보였으나, 겨울은 ACC가 현저히 낮은 수준으로 나타났다. 반면, 겨울철 POD (Probability of Detection)는 매우 높은 값을 기록하였으나, 이는 양의 방향으로 과도한 편향(NMB)에 의해 발생

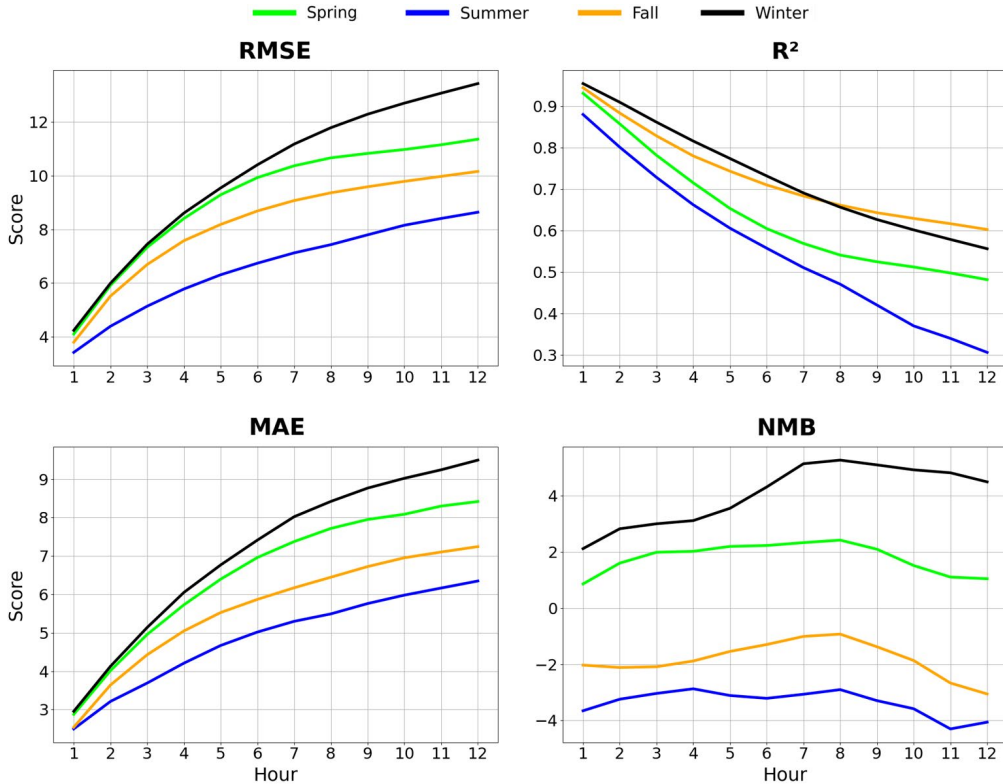


Fig. 10. Seasonal model results (2022). Seasonal PM_{2.5} prediction results in 2022 using the ensemble model excluding XGBoost and applying the Kalman Filter. Summer shows the best performance in RMSE and MAE but the lowest R², whereas winter exhibits the opposite.

한 결과로 판단된다. 여름의 경우 전체적으로는 ACC가 양호해 보이지만, 강수에 의한 낮은 PM_{2.5} 농도를 과소 예측함에 따라 POD가 낮아지고, 8시간 이후 시점부터 FAR (False Alarm Rate)이 급격히 증가하는 경향을 보인다.

결과적으로, 봄은 오차가 다소 높은 편이지만 일부 등급 예측은 잘 수행되었으며, 가을은 낮은 오차와 음의 편향을 보이면서도 전체적으로 우수한 성능을 나타냈다. 반면 겨울과 여름은 각 계절의 기상학적 특성으로 인해 예측 편향이 크고 정확도가 낮아지는 한계가 관찰되었다. 이러한 분석은 계절별 모델 보정 필요성을 시사하며, 향후 계절 특성을 반영한 정교한 모델링이 진행되어야 할 필요가 있음을 보여준다.

계절별로 대기질에 영향을 미치는 주요 변수가 상

이할 것으로 예상됨에 따라, 본 연구에서는 연간 전체 데이터를 기반으로 예측을 수행하는 대신, 변수 중요도 분석 결과를 바탕으로 각 계절에 적합한 입력 변수를 선별하여 계절별로 예측 모델을 구축하였다. 이를 통해 계절별 대기 특성을 보다 효과적으로 반영하고자 하였다. 특히, 강수 영향으로 PM_{2.5}가 과소모의되는 경향이 있는 여름철과, 반대로 과대모의되는 경향이 있는 겨울철의 경우, 계절별 입력 변수 최적화를 통해 예측 성능이 개선될 것으로 기대된다.

3.5 중기(48H)예측: 동일 모델 조건 예측 시간 증가에 따른 성능변화 분석

본 연구는 데이터의 특성을 반영할 수 있는 앙상블 모델을 구축하고, Permutation Feature Importance 기

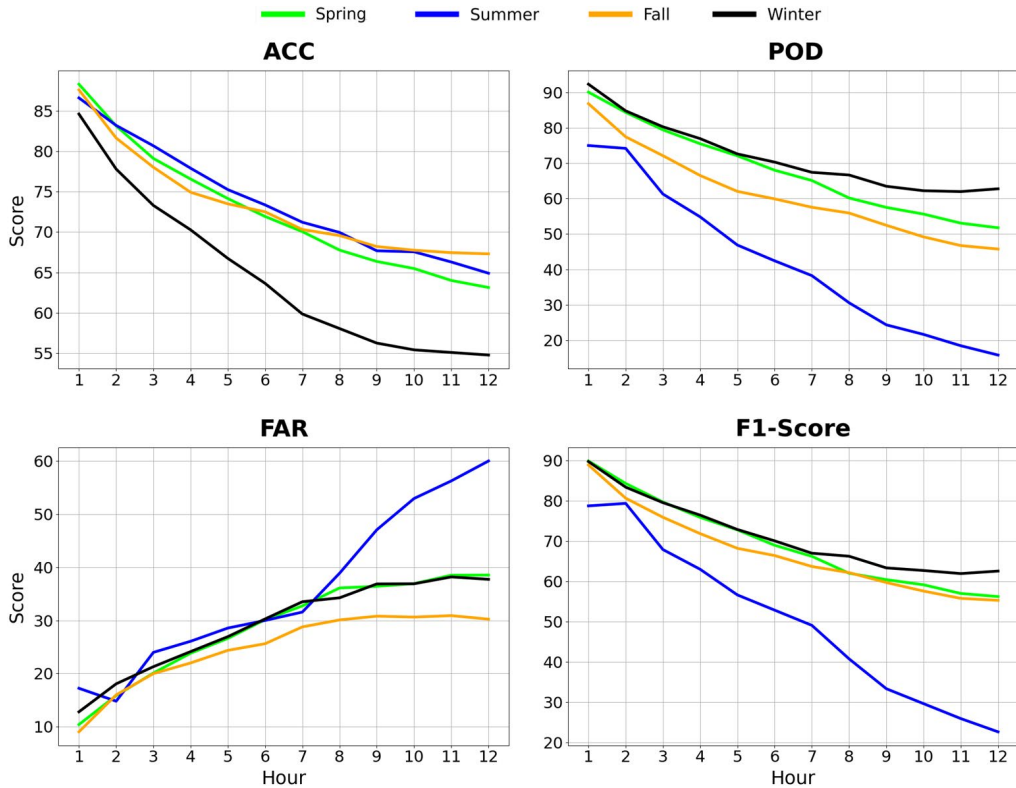


Fig. 11. Seasonal evaluation of PM_{2.5} grade classification results (2022). Seasonal evaluation of PM_{2.5} grade classification using the ensemble model with Kalman Filter applied (2022 data). Summer shows high FAR and low F1-score, while spring and fall demonstrate relatively stable classification performance.

법을 활용하여 예측 성능에 부정적인 영향을 미치는 변수를 제거함으로써 성능을 향상시키고, 이후 KF (Kalman Filter)를 후처리 기법으로 적용하여 예측의 정밀도를 추가로 개선하고자 하였다.

현재 대한민국의 대기질 예보 시스템(AirKorea 기준)은 “오늘”, “내일”, “모레” 수준의 예측을 제공하고 있으며, 에이 따라 본 연구 또한 단기(12시간) 수준을 넘어 중기(최대 48시간)까지의 예측 가능성을 검토하였다. 모델 구성과 입력 변수 조건은 단기 예측 실험과 동일하게 설정하였으며, 일반대기성분, 탄소성분, 이온성분, 금속성분, 흡수계수, 산란계수의 데이터 기반으로 Permutation Importance를 적용하여 주요 변수를 선별한 후, 해당 변수들만을 입력값으로 하여 앙상블 모델을 구축하고 48시간 예측을 수행하였다. 결과

는 그림 12, 그림 13에 나타내었다.

예측 시간의 증가에 따라 전반적인 성능 저하가 관찰되었으며, 특히 결정계수(R^2)의 경우 예측 시간 10시간을 기준으로 0.5 이하로 하락하는 양상을 보였다. 하지만 KF를 적용한 결과, 성능 저하 시점이 예측 16시간 이후로 지연되었으며, 이에 따라 중기 구간에서도 상대적으로 향상된 R^2 값을 확인할 수 있었다. 또한 RMSE와 MAE 오차 지표에서도 KF 적용 모델이 전반적으로 낮은 값을 보이며 개선된 예측 정확도를 나타냈다. NMB의 경우 상대적으로 큰 변화는 없었으나, 기존 모델에서 나타났던 과소 혹은 과대모의 경향은 일정 부분 완화되었음을 확인하였다. 미세먼지 등급 분류 기준을 적용한 성능 분석에서도 ACC와 POD는 KF 적용 시 향상된 성능을 보였으며, 특히 FAR의 경

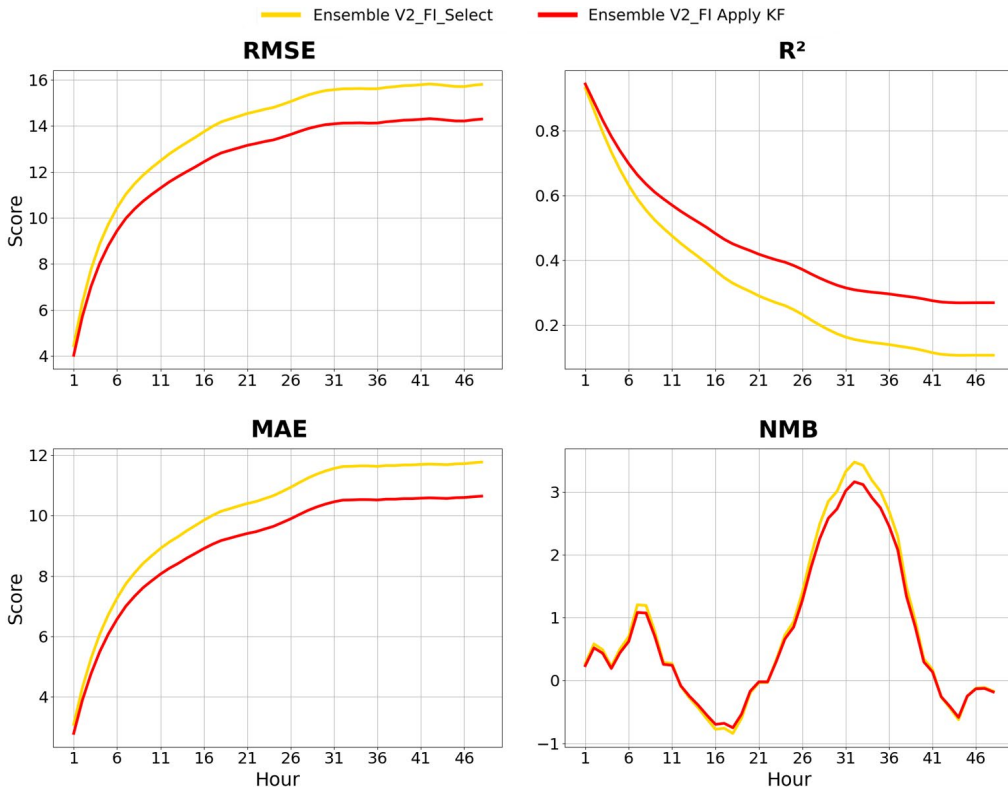


Fig. 12. Results before (gold) and after (red) applying the KF -48H. Comparison of 48-hour PM_{2.5} forecast performance before (gold) and after (red) applying the Kalman Filter to the ensemble model. The results show improved accuracy with reduced RMSE/MAE and increased R² after Kalman Filter application.

우 가장 뚜렷한 개선이 나타나, 이는 예측 오차 감소의 결과로 해석된다.

4. 결 론

본 연구는 PM_{2.5} 농도를 예측의 정확도를 향상시키기 위해 다양한 대기오염물질 데이터를 활용하고, 머신러닝 기반의 앙상블 기법과 칼만 필터(Kalman Filter, KF)를 결합한 예측 모델을 제안하였다. 특히 대기환경연구소에서 수집된 데이터를 통해 일반 측정망으로는 확보하기 어려운 세부 대기화학 성분을 포함함으로써, 보다 정밀한 예측이 가능함을 검토하였다.

모델링 과정에서는 선형회귀(LinearRegression)와

Random Forest, XGBoost와 같은 서로 다른 특성을 지닌 모델들을 활용하여 앙상블 모델을 구성하였고, Permutation Feature Importance 기법을 통해 예측 정확도를 저하시키는 변수를 제거함으로써 과적합을 줄이고 일반화 성능을 개선하여 예측 정확도를 향상시켰다. 또한, 각 모델의 성능을 비교하여 예측력이 현저히 낮은 모델은 앙상블 구성에서 제외하는 방식으로 모델 최적화를 수행하였다.

모델의 학습 및 검증은 2019년부터 2022년까지 서울 불광동의 일반측정소 및 대기환경연구소에서 수집된 데이터를 기반으로 수행되었으며, 단기 예측(12시간)과 중기 예측(48시간)을 대상으로 모델의 적합성과 예측력을 평가하였다. 그 결과, 일반측정망 데이터를 활용한 경우보다 대기환경연구소의 데이터를 활용

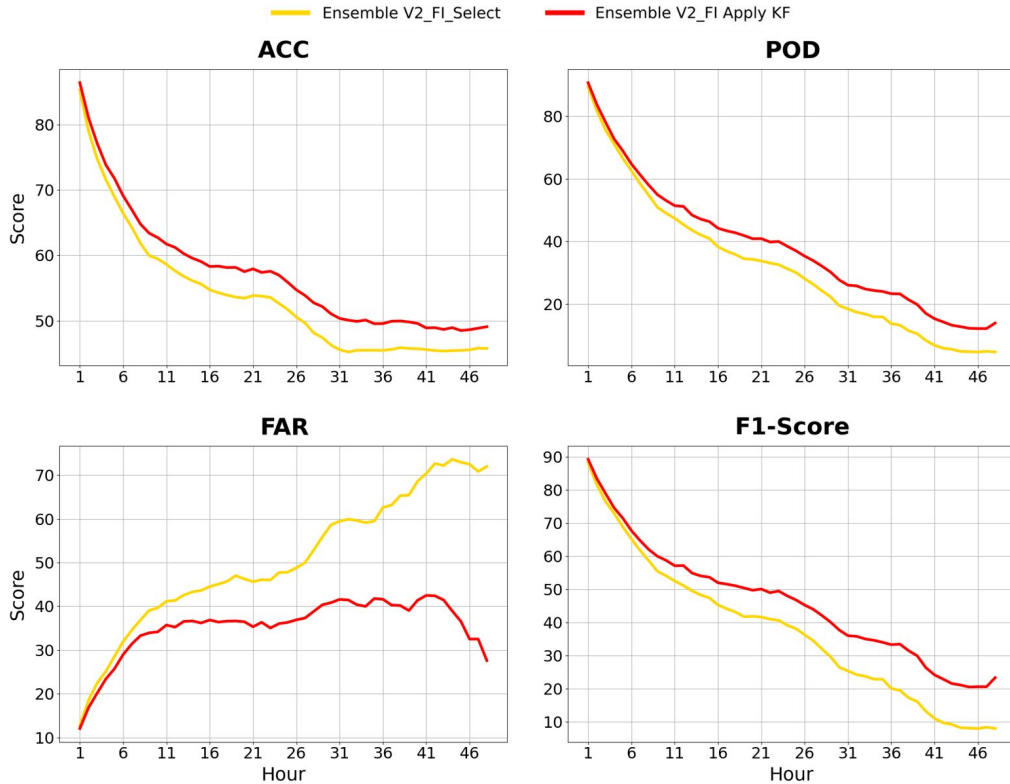


Fig. 13. Results before (gold) and after (red) applying the KF - 48H $PM_{2.5}$ grade classification. Improvements are observed in ACC (Accuracy), POD (Probability of Detection), FAR (False Alarm Rate), and F1-Score after applying the KF.

한 모델이 전반적으로 더 우수한 예측 성능을 보였으며, 특히 Permutation Feature Importance를 통한 변수 선별이 성능 향상에 기여함을 확인하였다.

후처리 기법으로 적용한 칼만 필터는 예측 시간 증가에 따라 발생하는 성능 저하 문제를 효과적으로 개선하였다. 단기 예측에서는 앙상블 최적화만으로는 편향 지표 일부만 개선되었으나, KF 적용 후 오차 (RMSE, MAE), 편향(NMB), 결정계수(R^2) 등 모든 지표에서 성능이 향상되었다. 중기 예측에서도 유사한 경향을 보이며, 특히 음·양의 편향이 모두 존재하던 모델 결과에서 칼만 필터 적용 후 편향이 효과적으로 개선되었다. 이는 후처리 기법이 과대모의 및 과소모의를 동시에 보정할 수 있음을 보여준다.

아울러, 대한민국의 뚜렷한 사계절 특성을 반영한 계절별 평가 결과, 겨울과 봄철에는 오차가 크게 나타

나고 양의 편향이 두드러졌으며, 여름철에는 전반적으로 오차는 낮았으나 고농도 $PM_{2.5}$ 에 대한 민감도가 낮게 나타났다. 특히 여름철의 음의 편향은 높은 습도 및 강수량의 영향으로 해석되며, 이는 대기질 예측에서 계절 요인을 반드시 고려해야 함을 시사한다.

종합적으로, 본 연구는 고해상도 대기성분 자료의 활용, 앙상블 모델 최적화, 그리고 칼만 필터 기반의 후처리 기법이 $PM_{2.5}$ 예측 성능을 효과적으로 향상시킬 수 있음을 입증하였다. 향후 연구에서는 칼만 필터의 설정값을 상황에 따라 조절할 수 있는 적응형 접근 방식을 적용하고, 추가적인 기상 및 대기오염 관련 변수 통합을 통해 예측 모델의 장기적 안정성과 예측력을 제고할 수 있을 것으로 기대된다. 이러한 통합적 접근은 향후 대기질 예측, 오염관리 정책 수립, 그리고 조기 경보 체계 고도화에 기초자료로 활용될 수 있을 것이다.

감사의 글

이 연구는 국립환경과학원에서 주최한 제3회 대학(원)생 미세먼지연구아이디어공모전으로 수행되었습니다(NIER-2024-03-00-005).

References

- Altmann, A., Toloşi, L., Sander, O., Lengauer, T. (2010) Permutation importance: A corrected feature importance measure, *Bioinformatics*, 26(10), 1340-1347. <https://doi.org/10.1093/bioinformatics/btq134>
- Baklanov, A., Molina, L.T., Gauss, M. (2016) Megacities, air quality and climate, *Atmospheric Environment*, 126, 235-249. <https://doi.org/10.1016/j.atmosenv.2015.11.059>
- Bowe, B., Xie, Y., Yan, Y., Al-Aly, Z. (2019) Burden of cause-specific mortality associated with PM_{2.5} air pollution in the United States, *JAMA Network Open*, 2(11), e1915834-e1915834. <https://doi.org/10.1001/jamanetworkopen.2019.15834>
- Cheng, F.Y., Feng, C.Y., Yang, Z.M., Hsu, C.H., Chan, K.W., Lee, C.Y., Chang, S.C. (2021) Evaluation of real-time PM_{2.5} forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted PM_{2.5} forecasts in Taiwan, *Atmospheric Environment*, 244, 117909. <https://doi.org/10.1016/j.atmosenv.2020.117909>
- Debeer, D., Strobl, C. (2020) Conditional permutation importance revisited, *BMC Bioinformatics*, 21, 1-30. <https://doi.org/10.1186/s12859-020-03622-2>
- García-Nieto, P.J., García-Gonzalo, E., Paredes-Sánchez, J.P. (2021) Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso and Elastic-net machine learning techniques, *Neural Computing and Applications*, 33, 17131-17145. <https://doi.org/10.1007/s00521-021-06304-z>
- Ghim, Y.S., Choi, Y., Park, J.S., Kim, C.H. (2014) An assessment study for the urban air monitoring network in Seoul, *Journal of Korean Society for Atmospheric Environment*, 30(5), 504-509, (in Korean with English abstract). <https://doi.org/10.5572/KOSAE.2014.30.5.504>
- Heidari, A., Navimipour, N.J., Unal, M. (2022) Applications of ML/DL in the management of smart cities and societies based on new trends in information technologies: A systematic literature review, *Sustainable Cities and Society*, 85, 104089. <https://doi.org/10.1016/j.scs.2022.104089>
- Kang, J.G., Lee, J.Y., Lee, J.B., Lim, J.H., Yun, H.Y., Choi, D.R. (2024) High-Resolution Daily PM_{2.5} Exposure Concentrations in South Korea Using CMAQ Data Assimilation with Surface Measurements and MAIAC AOD (2015-2021), *Atmosphere*, 15(10), 1152. <https://doi.org/10.3390/atmos15101152>
- Khan, A., Ali, A., Khan, J., Ullah, F., Faheem, M. (2025) Using Permutation-Based Feature Importance for Improved Machine Learning Model Performance at Reduced Costs, *IEEE Access*, 13, 36421-36435. <https://doi.org/10.1109/ACCESS.2025.3544625>
- Kim, D.J., Kim, T.H., Choi, J.Y., Lee, J.B., Kim, R.H., Son, J.S., Lee, D. (2024) The Impact of Vertical Eddy Diffusivity Changes in the CMAQ Model on PM_{2.5} Concentration Variations in Northeast Asia: Focusing on the Seoul Metropolitan Area, *Atmosphere*, 15(3), 376. <https://doi.org/10.3390/atmos15030376>
- Kioumourtzoglou, M.A., Schwartz, J.D., Weisskopf, M.G., Melly, S.J., Wang, Y., Dominici, F., Zanobetti, A. (2016) Long-term PM_{2.5} exposure and neurological hospital admissions in the northeastern United States, *Environmental Health Perspectives*, 124(1), 23-29. <https://doi.org/10.1289/ehp.1408973>
- Krittanawong, C., Qadeer, Y.K., Hayes, R.B., Wang, Z., Virani, S., Thurston, G.D., Lavie, C.J. (2023) PM_{2.5} and cardiovascular health risks, *Current Problems in Cardiology*, 48(6), 101670. <https://doi.org/10.1016/j.cpcardiol.2023.101670>
- Liu, X., Lu, D., Zhang, A., Liu, Q., Jiang, G. (2022) Data-driven machine learning in environmental pollution: gains and problems, *Environmental Science & Technology*, 56(4), 2124-2133. <https://doi.org/10.1021/acs.est.1c06157>
- Luo, S.T., Cheng, B.W. (2012) Diagnosing breast masses in digital mammography using feature selection and ensemble methods, *Journal of Medical Systems*, 36, 569-577. <https://doi.org/10.1007/s10916-010-9518-8>
- Melville, P., Mooney, R.J. (2005) Creating diversity in ensembles using artificial data, *Information Fusion*, 6(1), 99-111. <https://doi.org/10.1016/j.inffus.2004.04.001>
- Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., Pak, C. (2020) Deep learning-based PM_{2.5} prediction considering the spatiotemporal correlations: A case study of Beijing, China, *Science of the Total Environment*, 699, 133561. <https://doi.org/10.1016/j.scitotenv.2019.07.367>
- Park, H., Kim, K., Lee, D.K. (2019) Prediction of severe drought

- area based on random forest: Using satellite image and topography data, *Water*, 11(4), 705. <https://doi.org/10.3390/w11040705>
- Park, J., Kim, H.J., Lee, C.H., Lee, C.H., Lee, H.W. (2021) Impact of long-term exposure to ambient air pollution on the incidence of chronic obstructive pulmonary disease: a systematic review and meta-analysis, *Environmental Research*, 194, 110703. <https://doi.org/10.1016/j.envres.2020.110703>
- Pérez, P., Trier, A., Reyes, J. (2000) Prediction of PM_{2.5} concentrations several hours in advance using neural networks in Santiago, Chile, *Atmospheric Environment*, 34(8), 1189-1196. [https://doi.org/10.1016/S1352-2310\(99\)00316-7](https://doi.org/10.1016/S1352-2310(99)00316-7)
- Rahmani, A.M., Yousefpoor, E., Yousefpoor, M.S., Mehmood, Z., Haider, A., Hosseinzadeh, M., Ali Naqvi, R. (2021) Machine learning (ML) in medicine: Review, applications, and challenges, *Mathematics*, 9(22), 2970. <https://doi.org/10.3390/math9222970>
- Ryu, M., Son, S., Kim, J. (2022) Prediction and Analysis of PM_{2.5} Concentration in Seoul Using Ensemble-based Model, *Korean Journal of Remote Sensing*, 38(6), 1191-1205, (in Korean with English abstract). <https://doi.org/10.7780/kjrs.2022.38.6.1.16>
- Shi, L., Johansson, K.H., Murray, R.M. (2007) Kalman filtering with uncertain process and measurement noise covariances with application to state estimation in sensor networks, 2007 IEEE International Conference on Control Applications, pp. 1031-1036. <https://doi.org/10.1109/CCA.2007.4389369>
- Shi, L., Wu, X., Yazdi, M.D., Braun, D., Awad, Y.A., Wei, Y., Liu, P., Di, Q., Wang, Y., Schwartz, J., Dominici, F., Kioumourtoglou, M.-A., Zanobetti, A. (2020) Long-term effects of PM_{2.5} on neurological disorders in the American Medicare population: a longitudinal cohort study, *The Lancet Planetary Health*, 4(12), e557-e565 [https://doi.org/10.1016/S2542-5196\(20\)30227-8](https://doi.org/10.1016/S2542-5196(20)30227-8)
- Suleiman, A., Tight, M.R., Quinn, A.D. (2019) Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM₁₀ and PM_{2.5}), *Atmospheric Pollution Research*, 10(1), 134-144. <https://doi.org/10.1016/j.apr.2018.07.001>
- Szott, S., Kosek-Szott, K., Gawłowicz, P., Gómez, J.T., Bellalta, B., Zubow, A., Dressler, F. (2022) Wi-Fi meets ML: A survey on improving IEEE 802.11 performance with machine learning, *IEEE Communications Surveys & Tutorials*, 24(3), 1843-1893. <https://doi.org/10.1109/comst.2022.3179242>
- Vignesh, P.P., Jiang, J.H., Kishore, P. (2023) Predicting PM_{2.5} concentrations across USA using machine learning, *Earth and Space Science*, 10(10), e2023EA002911. <https://doi.org/10.1029/2023EA002911>
- Welch, G. (1995) An introduction to the Kalman filter [Technical report]. University of North Carolina at Chapel Hill. https://st.ewi.tudelft.nl/koen/cs4140/Resources/kalman_theory.pdf
- Xiao, F., Yang, M., Fan, H., Fan, G., Al-Qaness, M.A. (2020) An improved deep learning model for predicting daily PM_{2.5} concentration, *Scientific Reports*, 10(1), 20988. <https://doi.org/10.1038/s41598-020-77757-w>
- Xiao, Q., Chang, H.H., Geng, G., Liu, Y. (2018) An ensemble machine-learning model to predict historical PM_{2.5} concentrations in China from satellite data, *Environmental Science & Technology*, 52(22), 13260-13269. <https://doi.org/10.1289/isesisee.2018.o01.04.07>
- Yang, Y., Yang, T., Zhou, J., Cao, Z., Liao, Z., Zhao, Y., Su, X., He, J., Hua, J. (2023) Prenatal exposure to concentrated ambient PM_{2.5} results in spatial memory defects regulated by DNA methylation in male mice offspring, *Environmental Science and Pollution Research*, 30(12), 35142-35152. <https://doi.org/10.1007/s11356-022-24663-5>
- Yoon, H.J., Han, G.I., Cho, S.H., Jung, B.H. (2015) GIS-based PM₁₀ concentration real-time service, *Journal of Korean Society for Atmospheric Environment*, 31(6), 585-592, (in Korean with English abstract). <https://doi.org/10.5572/KOSAE.2015.31.6.585>
- Yu, G.H., Park, S.S., Ghim, Y.S., Shin, H.J., Lim, C.S., Ban, S.J., Yu, J.A., Kang, H.J., Seo, Y.K., Kang, K.S., Jo, M.R., Jung, S.A., Lee, M.H., Hwang, T.K., Kang, B.C., Kim, H.S. (2018) Difference in chemical composition of PM_{2.5} and investigation of its causing factors between 2013 and 2015 in air pollution intensive monitoring stations, *Journal of Korean Society for Atmospheric Environment*, 34(1), 16-37, (in Korean with English abstract). <https://doi.org/10.5572/kosae.2018.34.1.016>
- Yu, P., Guo, S., Xu, R., Ye, T., Li, S., Sim, M.R., Abramson, M.J., Guo, Y. (2021) Cohort studies of long-term exposure to outdoor particulate matter and risks of cancer: a systematic review and meta-analysis, *The Innovation*, 2(3). <https://doi.org/10.1016/j.xinn.2021.100143>
- Yuan, S., Wang, J., Jiang, Q., He, Z., Huang, Y., Li, Z., Cai, L., Cao, S. (2019) Long-term exposure to PM_{2.5} and stroke: a systematic review and meta-analysis of cohort studies, *Environmental Research*, 177, 108587. <https://doi.org/10.1016/j.envres.2019.108587>
- Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S. (2019) PM_{2.5} prediction based on random for-

- est, XGBoost, and deep learning using multisource remote sensing data, *Atmosphere*, 10(7), 373. <https://doi.org/10.3390/atmos10070373>
- Zhang, X., Ding, C., Wang, G. (2024) An autoregressive-based Kalman filter approach for daily PM_{2.5} concentration forecasting in Beijing, China, *Big Data*, 12(1), 19-29. <https://doi.org/10.1089/big.2022.0082>
- Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B.M., Xiao, X., Yu, X., Zhu, J.-J., Zhang, H. (2021) Machine learning: new ideas and tools in environmental science and engineering, *Environmental Science & Technology*, 55(19), 12741-12754. <https://doi.org/10.1021/acs.est.1c01339>
- Zhu, W., Cai, J., Hu, Y., Zhang, H., Han, X., Zheng, H., Wu, J. (2021) Long-term exposure to fine particulate matter relates with incident myocardial infarction (MI) risks and post-MI mortality: A meta-analysis, *Chemosphere*, 267, 128903. <https://doi.org/10.1016/j.chemosphere.2020.128903>

Authors Information

- 이성일 (명지대학교 환경에너지공학과 박사과정)
(si_lee@mju.ac.kr)
- 이동훈 (명지대학교 환경에너지공학과 석박사통합과정)
(dong2979@mju.ac.kr)
- 송효중 (명지대학교 환경에너지공학과 부교수)
(hjsong@mju.ac.kr)
- 안준영 (국립환경과학원 대기환경연구과 연구관)
(nierair@korea.kr)
- 박승명 (국립환경과학원 대기환경연구과 전문위원)
(psm5555@korea.kr)
- 이재윤 (국립환경과학원 대기환경연구과 전문위원)
(jaeyun2368@korea.kr)