Atmosphere. Korean Meteorological Society
Vol. 36, No. 1 (2026) pp. 1-12
https://doi.org/10.14191/Atmos.2026.36.1.001
pISSN 1598-3560 eISSN 2288-3266

연 구 논 문
(Article)

기계학습을 이용한 태양광 발전 예측: 영암 F1 태양광 발전소 사례

심정우[1] · 김예림[2] · 손석우[1),2),3)*] · 최정[1]
[1]서울대학교 자연과학대학 지구환경과학부, [2]서울대학교 인공지능 협동과정,
[3]서울대학교 에너지 이니셔티브

(접수일: 2025년 9월 11일, 수정일: 2026년 1월 6일, 게재확정일: 2026년 1월 30일)

Solar Photovoltaic Power Prediction Using Machine Learning: A Case Study of the Yeongam F1 Solar Power Plant

Jeongwoo Shim[1], Yelim Kim[2], Seok-Woo Son[1),2),3)*], and Jung Choi[1]
[1]*School of Earth and Environmental Sciences, Seoul National University, Seoul, Korea*
[2]*Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea*
[3]*SNU Energy Initiative, Seoul National University, Seoul, Korea*

(Manuscript received 11 September 2025; revised 6 January 2026; accepted 30 January 2026)

Abstract Accurately estimating and predicting solar photovoltaic (PV) power generation is essential for ensuring a stable renewable energy supply. This study proposes a machine learning-based approach to estimate and predict solar PV power generation. A linear model and three machine learning models were trained using hourly datasets from the Yeongam F1 solar power plant from 2019 to 2022. To ensure model stability, a year-based 4-fold cross-validation was performed. Of the tested models, XGBoost model performed the best, achieving a normalized mean absolute error (nMAE) of 4.16% with a standard deviation of 0.07%. Furthermore, in a practical forecasting scenario where models were trained on data from 2019 to 2021, XGBoost model successfully predicted next-day solar PV power generation for 2022 using LDAPS forecasts initialized every 2100 KST, achieving a nMAE of 7.89%. These results demonstrate that combining short-term numerical weather forecasts with machine learning algorithms can enable reliable prediction of solar PV power generation one day in advance.

Key words: Solar PV power, Yeongam F1 solar power plant, Machine learning, Bias correction

1. 서 론

전 세계적으로 기후변화의 심각성이 대두되면서 온실가스 감축을 위한 신재생 에너지의 중요성이 급격히 부각되고 있다. 특히 우리나라는 2050년 탄소중립 달성을 목표로 설정하였으며, 이를 달성하기 위해 풍력, 태양광 발전 등 청정에너지원의 확대를 적극적으로 추진하고 있다. 정부는 2030년까지 신재생에너지 발전 비중을 20%로 확대하겠다는 정책 목표를 제시하였으며, 2024년에는 신재생에너지 발전 비중이 처음으로 10%를 돌파하는 성과를 달성하였다(MOTIE, 2025). 한국에너지공단의 2023년도 신재생에너지 보급통계에 따르면, 현재 신재생에너지 설비용량의 55%를 태양광 발전이 차지하고 있어 국내 청정에너지 전환의 핵심 동력으로 자리잡고 있다(KEA, 2024).

태양광 발전은 발전 과정에서 탄소를 배출하지 않으며, 무한하고 지속가능한 태양 에너지를 활용한다는 점에서 미래 주력 청정 에너지원으로 각광받고 있다. 그러나 태양광 발전은 일사량, 운량, 기온 등의 기상 조건에 따라 발전량이 민감하게 변동하는 내재적

*Corresponding Author: Seok-Woo Son, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea.
Phone: +82-2-880-8147, Fax: +82-2-883-4972
E-mail: seokwooson@snu.ac.kr

한계를 가진다. 급변하는 기상 상태는 태양광 발전 전력 계통의 안정성을 저해하는 주요 요인으로, 기상조건에 기반한 정확한 발전량 예측의 중요성이 최근 강조되고 있다. 단기적으로는 실시간 전력 수급 균형 유지를, 장기적으로는 전력 인프라 투자 계획 수립을 위해 신뢰성 높은 태양광 발전량 예측 모델은 필수적이다(Alkabbani et al., 2021).

태양광 발전량은 다양한 방법을 통해 예측할 수 있다. 초기에는 기상 변수에서 유도된 물리적 모델이나 통계-경험적 모델들을 중심으로 연구가 진행되었다(Antonanzas et al., 2016; Sobri et al., 2018; Ahmed et al., 2020). 이러한 방식은 주로 선형 모델에 기반하며, 비선형적인 변화를 예측하는데 한계가 있다. 이를 극복하기 위해 최근에는 기계학습과 심층학습 모델이 도입되었다(Markovics and Mayer, 2022; Gaboitaolelwe et al., 2023). 심층학습 모델은 기계학습 모델과 비교하여 복잡하고 규모가 큰 자료에서 더 높은 예측 정밀도를 달성할 수 있다는 장점이 있지만, 결과에 대한 해석이 어렵다는 단점이 있다. 반면 기계학습 모델은 단순하고 규모가 적은 자료에서도 활용이 가능하며 예측 결과의 해석이 용이하다(Iheanetu, 2022).

기계학습을 활용한 태양광 발전량 예측 연구는 국내외에서 활발하게 진행되고 있다. Gottwald et al. (2024)는 기상 변수를 입력으로 하는 선형모델과 다양한 기계학습 모델의 성능을 비교하여 Random Forest와 XGBoost 모델의 우수한 예측 성능을 확인하였다. Nguyen et al. (2025)는 여러 기계학습 알고리즘 중 CatBoost 모델이 최적의 성능을 보임을 보고하였다. 그러나 기존 연구들은 주로 관측 자료를 활용해 태양광 발전량을 추정하는 데 국한되어 있다. 기계학습 모델을 실시간 예보에 적용한 연구들은 많지 않은 실정이다(Khaire et al., 2023). 특히 국내 사례는 매우 제한적이다.

Kim et al. (2019)는 기상청 수치예보모델인 LDAPS를 활용하여 일사량 예보 성능을 평가하였다. 그러나 태양광 발전량 예측은 시도되지 않았다. Lee et al. (2021)은 국내 태양광 발전소를 대상으로, 한국에너지기술연구원에서 제공하는 수치예보모델 자료를 활용하여 다양한 기계학습 및 심층학습 기반의 태양광 발전량 예측 모델의 성능을 평가하였다. 이 연구에서는 수치예보모델 자료를 결합하여 모델을 학습 및 검증하였으나, 제공된 수치예보모델 자료 기간이 상대적으로 짧아 실제 학습에 1년치 자료만 사용되었다. 이로 인해 장기적인 예측 신뢰도 평가나 계절변동성 등에 대한 모델 성능 검증에는 한계가 있었다. Shim et al. (2024)는 기상청 동네예보 자료를 활용하여 태양광 발전량 예측을 수행하였으나, 예측 범위가 도시 단위의 일일 발전량으로 한정되었으며 예측에 가장 핵심적인 변수인 일사량을 포함하지 않았다. 따라서 실제 현업에서 요구하는 세밀한 시간 해상도나 개별 발전소 수준의 예측에는 적용에 한계가 있었다.

본 연구는 국내 예보 환경에 최적화된 기계학습 기반의 태양광 발전량 추정 모델을 개발하고, 이를 기상청 수치예보모델 예측자료에 적용하여 태양광 발전량의 24시간 예측 가능성을 평가하였다. 구체적으로 전라남도 영암의 F1 자동차경주장 태양광 발전소(이하 영암 F1 태양광 발전소) 발전량을 분석하고, 이를 인근 기상관측소의 관측 자료를 활용해 추정하였다. 발전량 추정에는 선형 모델뿐만 아니라 3가지 기계학습 모델을 활용하였다. 개발된 모델들은 현업 국지예보모델 예측자료에 적용되었으며, 이를 통해 태양광 발전량의 24시간 예측 가능성을 평가하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 연구에 사용된 영암 F1 태양광 발전소 발전량 자료, 목포 기상대의 기상 관측 자료, 기상청 국지예보모델 자료를 기술하였다. 제3장에서는 자료 전처리, 모델 구축, 성능 평가 방법을 서술하였다. 제4장에서는 태양광 발전량 예측 모델을 활용한 실험 결과를 제시하였으며, 마지막 제5장에서는 결론과 향후 과제에 대해 논의하였다.

2. 자 료

2.1 영암 F1 태양광 발전소 발전량

본 연구에서는 전라남도 영암군에 위치한 영암 F1 자동차경주장 태양광 발전소(34.745°N, 126.411°E) 자료를 활용하였다. 해당 발전소는 영암 F1 자동차경주장 내 주차장에 설치되어 있으며, 총 설비용량은 13.296 MW이다. 2019년부터 2022년까지 총 4년간의 시간별 발전량자료를 분석하였다. 2018년 이전 자료는 데이터 품질 검토 결과를 토대로 분석에서 제외하였다. 자료는 공공데이터포털(data.go.kr)을 통해 수집하였다.

2.2 목포 기상대 관측 자료

기상 관측 자료는 영암 F1 태양광 발전소와 지리적으로 가장 인접한 목포 기상대(지점번호: 165, 34.817°N, 126.381°E)의 종관기상관측(automated synoptic observing system, ASOS) 자료를 활용하였다. 목포 기상대는 발전소로부터 북서쪽으로 약 10 km 거리에 위치한다(Fig. 1).

2.3 기상청 국지예보모델 자료

기계학습 모델의 현업 적용성을 평가하기 위해, 기상청에서 운영하는 국지예보모델(local data assimilation and prediction system, LDAPS)의 예측 자료를 활용하

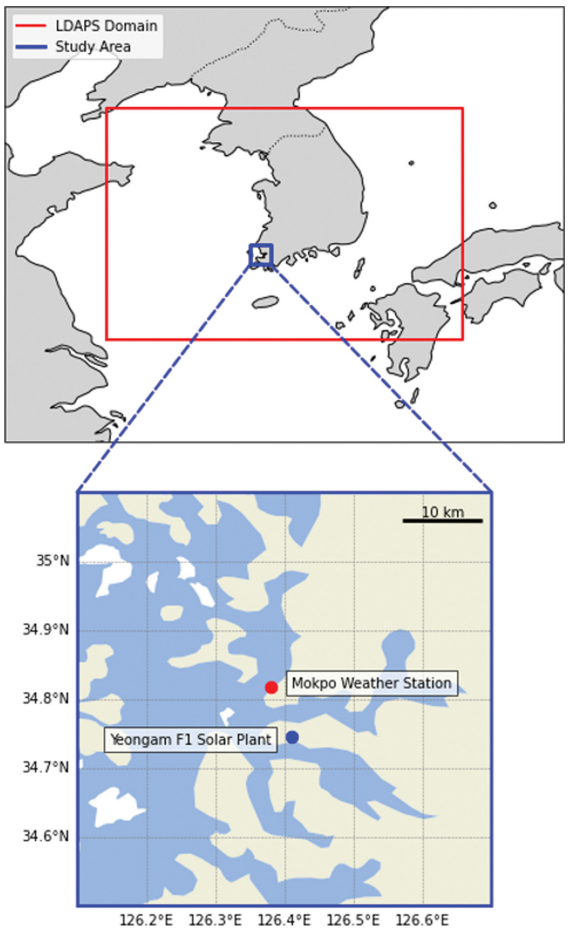

Fig. 1. LDAPS modeling domain (red box) and the study area (blue box). Locations of the Mokpo weather station (red) and the Yeongam F1 solar power plant (blue) are indicated in the bottom panel.

였다. LDAPS는 한반도 지역의 상세한 기상 예측을 위해 설계된 고해상도 수치예보모델로, 수평으로는 1.5 km의 해상도, 연직으로는 약 40 km 고도까지 70개 층으로 구성되어 있다. LDAPS 모델의 예측 영역은 Fig. 1에 표시하였다. 해당 모델은 하루 8회 3시간 간격으로 전지구모델로부터 경계장을 제공받아 운영되는데, 주요 시간대인 0900, 1500, 2100, 0300 KST (0000, 0600, 1200, 1800 UTC)에는 48시간 예측을, 1200, 1800, 0000, 0600 KST (0300, 0900, 1500, 2100 UTC)에는 3시간 예측을 수행한다. 3차원변분 자료동화 기법으로 자체 분석, 예측의 순환체계로 운영된다 (Oh et al., 2024; KMA, 2025). 발전량 예측에는 매일 2100 KST에 초기화된 48시간 예측 자료를 활용하였다. 2022년 1년 동안 예측 자료를 이용했으며, 발전소에 가장 인접한 격자점(34.745°N, 126.411°E)의 예측값을 추출하여 사용하였다. 입력 변수는 ASOS 변수와 동일하게 구성하였다.

3. 연구 방법

3.1 태양광 발전량 추정 모델

본 연구에서 활용한 태양광 발전량 추정 모델을 Table 1에 제시하였다. 가장 기본적으로 다중 선형 회귀(multiple linear regressor, MLR) 모델을 활용하였다. MLR는 변수 사이의 선형 관계를 모델링하며, 계산 복잡도와 과적합 위험이 낮은 반면 변수 사이의 복잡한 비선형 관계는 포착하기 어렵다는 단점이 있다. 그 외 3가지 기계학습 모델을 활용하였다. 구체적으로 가장 기본적인 기계학습 모델인 Random Forest (RF), 높은 성능으로 알려진 XGBoost (eXtreme Gradient Boosting, XGB), 그리고 XGB와 비슷한 성능이지만 학습 속도가 빠른 LightGBM (Light Gradient Boosting Machine, LGB)를 사용했다. RF는 Breiman (2001)이 제안한 기계학습 모델로, 의사결정 나무를 독립적으로 여러 번 학습한 뒤 이 들의 예측 결과를 평균하여 최종 예측 값을 도출하는 방법이다. 의사결정 나무를 독립적으로 학습시키는 RF와 달리 Gradient Boosting은 이전 학습에서의 오류를 보완하도록 순차적으로 학습시키는 기법이다. 이러한 방법 중 하나인 XGB는 Chen and Guestrin (2016)이 제안한 모델로 RF 보다 성능이 뛰어난 것으로 알려져 있다. LGB는 Microsoft에서 개발한 Gradient Boosting 기반 모델로, XGB와 비슷한 성능을 유지하면서 계산 효율을 높여 학습 시간을 단축한 알고리즘이다.

각 모델은 3개에서 9개의 변수를 활용해 태양광 발전량을 추정한다. 사용된 변수 구성을 명확히 구분하기 위해, 모델 약어 뒤에 입력 변수의 개수를 함께 표기하였다. 예를 들어 9개의 입력 변수를 사용한 XGB 모델은 XGB9으로 표기하였다. 또한 기계학습 모델들은 모두 하이퍼파라미터 최적화 라이브러리인 Optuna (Akiba et al., 2019)를 적용하여, 검증 오차를 최소화하는 최적의 하이퍼파라미터 조합을 탐색한 후 모델 학습에 활용하였다. 사용한 하이퍼파라미터는 Table 2

Table 1. List of the algorithms for estimating solar PV power generation and their abbreviations used in this study.

Linear model	Multiple Linear Regression (MLR)
Machine learning model	Random Forest (RF)
	LightGBM (LGB)
	XGBoost (XGB)

Table 2. Hyperparameters of the machine learning algorithms used in this study. The number following the model abbreviation indicates the number of input variables. See the main text.

Model	Hyperparameter	Value
RF9	n_estimators	534
	max_depth	11
	min_samples_split	0.01
	min_samples_leaf	4
LGB9	n_estimators	927
	max_depth	6
	learning_rate	0.02
	num_leaves	29
	colsample_bytree	0.92
	subsample	0.96
XGB9	n_estimators	686
	max_depth	6
	learning_rate	0.02
	gamma	5.62
	colsample_bytree	0.91
	subsample	0.84

에 제시하였다.

3.2 자료 전처리

야간 및 발전량이 거의 없는 시간대의 영향을 최소화하고 모델이 의미 있는 발전 패턴 학습에 집중하도록 자료를 정제하였다. 이를 위해 시간대별 설비이용률(capacity factor, CF)이 1% 이하인 경우는 학습에서 제거하였다. 여기서 설비이용률은 시간별 발전량을 설비용량 기준 최대 발전량(13.296 MWh)으로 나눈 백분률 값으로 정의된다. 설비이용률 1% 임계값은 실질적인 발전이 발생했다고 간주할 수 있는 최소한의 기준으로, 전체 학습 자료의 약 70% 정도를 차지하였다. 기계학습 모델의 평가 단계에서는 학습 자료와 동일한 기준을 적용하기 위해 XGBoost 분류 모델을 구축하여 특정 시간대의 발전 가능 여부(설비이용률 1% 초과/이하)를 예측하도록 하였다. 그리고 분류모델이 '1% 초과'로 예측하는 경우에만 태양광 발전량을 추정하거나 예측하였다.

ASOS 자료로 훈련된 모델은 예보 자료에 적용되었다. 2022년에 대해 2100 KST에 초기화된 LDAPS 예측 자료를 활용하였으며, 다음날 0500 KST부터 2000 KST까지 발전량을 예측하였다. 이 과정에서 LDAPS 자료와 ASOS 자료 간의 체계적 편차를 보정하기위해 분위사상법(Quantile Mapping)을 적용하였다. 각 기상 변수별로 ASOS 자료와 LDAPS 자료의 경험적 누적분포함수(empirical cumulative distribution function, CDF)를 산출하고, 동일한 누적확률값을 갖도록 LDAPS 자료를 ASOS 자료의 분포 특성에 매핑하여 보정된 값을 도출하였다(Gudmundsson et al., 2012). 구체적으로, LDAPS 자료의 각 값에 대해 해당하는 누적확률을 계산한 후, 동일한 누적확률에 대응하는 ASOS 훈련 자료의 값으로 변환하는 방식을 사용하였다. 이를 통해 LDAPS 자료의 분포가 ASOS 자료의 통계적 특성과 일치하도록 조정하였다.

3.3 성능 평가 방법

본 연구에서는 개발된 모델들의 성능을 객관적으로 평가하기 위해 평균 제곱근 오차(root mean square error, RMSE)와 정규화된 평균 절대 오차(normalized mean absolute error, nMAE)를 계산하였다. 의미 있는 평가 지표 계산을 위해 실질적인 발전이 발생하는 일조 시간대(0500~2000 KST)의 자료에 한정하여 분석을 수행하였으며, 2100 KST부터 0400 KST까지는 오차 계산에서 제외하였다.

RMSE는 예측값과 실제값 사이의 오차를 직관적으로 이해하기 쉬운 '거리(distance)'의 형태로 제공한다. 이는 개별 오차를 제곱한 후 평균을 내고, 다시 제곱근을 취하는 방식으로 계산된다.

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\Sigma_j^M (\hat{y}_j - y_j)^2}$$

여기서 M은 샘플 수를 의미한다. RMSE는 원시 자료와 동일한 단위를 가지므로, 오차의 물리적 크기를 쉽게 해석할 수 있는 장점이 있다. 반면 nMAE는 발전량의 규모와 무관하게 모델의 상대적 오차를 평가하기 위해 활용하였다. nMAE는 설비이용률(CF) 오차의 절댓값을 평균한 값으로 다음과 같이 계산된다.

$$\mathrm{nMAE} = \left(\frac{1}{M}\right)\Sigma\left|CF_{estimated} - CF_{actual}\right|$$

한국전력거래소 기준을 따라, nMAE는 설비이용률이 10% 이상인 시간대만을 대상으로 계산하였다. 자료 전처리 단계에서 노이즈 제거를 위해 사용된 1% 임계값과 달리, 평가에 적용된 10% 임계값은 의미 있는 발전 조건에서의 상대적 성능을 분석하기 위함이다. 단일 발전소 분석에서도 절대 발전량(MWh) 기반 MAE는 기간 중 설비 용량 변화나 출력 제한, 저발전 시간대의 영향에 따라 지표 해석이 달라질 수 있는 반면, CF 기반 지표는 발전 조건이 유사한 상황에서의 성능을 일관되게 비교할 수 있으므로 nMAE를 주요 평가지표로 선정하였다. 또한 CF는 실무에서 발전소 성능을 판단할 때 널리 사용하는 정규화 지표이므로, 예측 오차를 운영 관점에서 직관적으로 해석하는 데 유리하다.

4. 결 과

Figure 2는 전체 기간동안 영암 F1 태양광 발전소의 발전량 특징을 보여준다. 시간별 발전량의 시계열은 뚜렷한 계절성을 보인다(Fig. 2a). 전반적으로 4, 5월에 4 MWh 정도로 가장 높은 발전량을 기록하고 여름철에 조금 낮아졌다가 겨울철인 12월에 2 MWh 정도로 가장 낮은 발전량이 나타난다(Fig. 2b). 4, 5월이 여름철보다 발전량이 높은 것은 일사량은 충분하면서도 패널 온도가 적당하여 발전 효율이 최적화되기 때문으로 분석된다. 또한 여름철은 높은 기온으로 패널의 발전 성능이 떨어지고, 장마와 태풍 등 악기상으로 인해 발전량이 낮은 것으로 분석된다. 동일한 계절성이 재분석 자료로 추정한 한반도 및 동아시아 태양광 발전의 잠재력에서도 나타난다(Choi et al., 2024, 2025). 시간대별 평균 발전량은 오전 7시 이후 급격히 증가하고 13시 부근에서 7 MWh 이상으로 최대치 도달 후 다시 19시까지 감소하는 종모양 분포로, 전형적인 일사량 곡선을 따른다(Fig. 2c). 뚜렷한 일

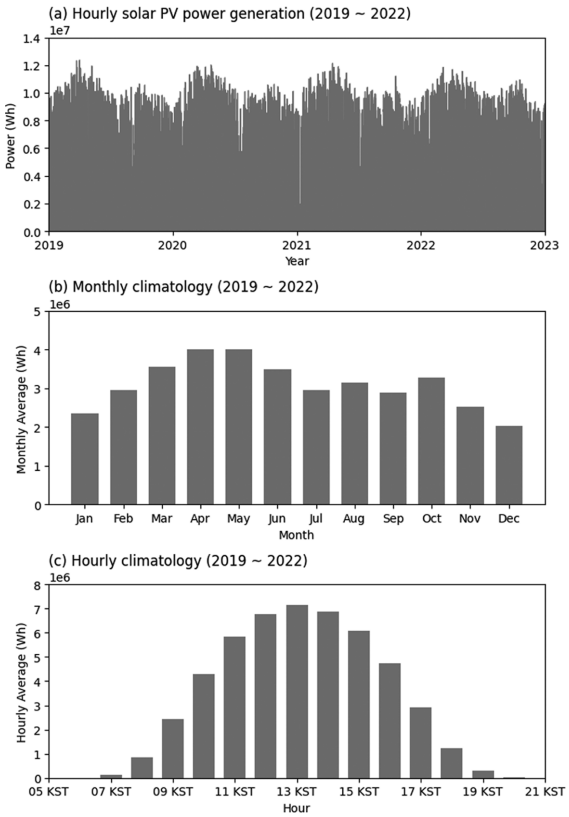

Fig. 2. (a) Time series of solar PV power generation at the Yeongam F1 solar power plant from 2019 to 2022, (b) its monthly climatology, and (c) hourly climatology.

주기 및 계절 주기와 달리, 장기 경향성은 뚜렷하지 않다. 이는 분석 기간 동안 설비 변화, 패널 오염, 패널 노후화 등이 크지 않았음을 의미한다.

태양광 발전량 추정을 위해 기상 변수와의 상관성을 먼저 평가하였다. 사용한 기상 변수는 기온(°C), 강수량(mm), 풍속(m s^{-1}), 습도(%), 일사량(MJ m^{-2}), 강설량(cm), 운량(10분위)의 7가지 변수이다. 이들 변수는 태양광 패널에 도달하는 유효 일사량 및 패널 온도에 영향을 주어 발전 효율을 변화시키는 요인들이다. 예를 들어 풍속이 증가하면 대류 열전달이 강화되어 패널의 냉각이 촉진되고, 패널 온도 하강을 통해 발전 효율이 개선될 수 있다. 반대로 습도가 높을수 록 수증기에 의한 태양광의 산란, 흡수가 증가하고 패널에 도달하는 일사량을 감소시킬 수 있다. 여기에 시간 정보인 cos_day와 cos_hour를 추가하였다. cos_day는 그 날짜가 일 년 중 몇 번째 날인지 계산하고 연중 날짜의 순환적 특성을 반영하기 위하여 cos 함수에 대입하였다. 마찬가지로 시간 또한 cos 함수에 대입하여 24시간 주기의 순환성을 보존하였다(Ki et al., 2023). Table 3은 9개 변수와 발전량 사이의 상관계수를 나타낸다. 일사량, cos_hour, 습도의 3개 변수는 상관계수의 절댓값이 0.5를 넘는 강한 상관관계를 가졌다. 운량, 기온, 풍속의 3개 변수는 상관계수의 절댓값이 0.2~0.3 정도로 다소 낮은 상관관계를 가졌으며, cos_day, 강수량, 강설량의 3개 변수는 상관계수 절댓값이 0.1 미만으로 약한 상관관계를 가졌다. 이 연구에서는 9개의 모든 변수를 사용한 모델과 강한 상관관계와 중간 상관관계를 가진 6개를 변수를 사용한 모델, 강한 상관관계를 가진 3개의 변수만 사용한 모델로 나누어 변수의 개수가 모델의 성능에 미치는 영향을 확인하였다.

상관계수에 기반한 변수 선택은 태양광 발전량 예측에서 효율적으로 활용되고 있다(Thipwangmek et al., 2024). 그러나 선형 관계에만 의존하므로 비선형 효과

Table 3. Hourly correlation coefficients between solar PV power generation at the Yeongam F1 solar power plant and meteorological variables observed at the Mokpo weather station for the period 2019~2021.

High correlation variables	Solar radiation	0.97
	cos_hour	−0.74
	Humidity	−0.56
Moderate correlation variables	Cloud cover	−0.27
	Windspeed	0.22
	Temperature	0.21
Low correlation variables	cos_day	−0.07
	Precipitation	−0.07
	Snowfall	−0.06

혹은 변수간 상호작용을 충분히 반영하지 못하는 한계가 있다. 그럼에도 불구하고, 본 연구에서는 1차적인 전처리 도구로 상관계수에 기반한 변수 선택을 활용하였다.

4.1 태양광 발전량 추정

먼저 목포 ASOS 관측 자료를 이용해 영암 F1 태양광 발전소의 발전량을 추정하였다. 전반적인 성능을 검증하기 위해 2019년부터 2022년까지 4개년의 ASOS 자료를 활용하여 연 단위 4-겹 교차 검증을 수행하였다. Table 4는 임의로 선택된 3년 자료로 모델을 훈련한 다음, 남은 1년 자료(Fold)로 성능을 검증한 결과를 보여준다. 기준 모델은 2019~2021년 자료를 훈련에 활용한 모델(Fold 1)로, 이후 분석에 활용된 모델이다. 전반적으로 기계학습 모델들은 선형 모델보다 우수한 성능을 보였으며, 모든 모델에서 입력변수의 수가 증가할수록 오차가 줄어 들었다(가장 오른쪽 열). 특히 기계학습 모델에서는 3개의 변수만을 사용하여도(RF9, LGB9, XGB9), 9개의 변수를 모두 사용한 선형모델(MLR9) 보다 더 나은 결과를 달성했

Table 4. nMAE (%) of solar PV power generation estimated from meteorological measurements at the Mokpo weather station using 4-fold cross-validation (2019~2022). The number following the model abbreviation indicates the number of input variables. The values in parenthesis in the last column indicate the standard deviation. Here, each model was trained with three years by excluding the 'Fold' year, and evaluated against the 'Fold' year. The reference model used in Figs. 4, 5 corresponds to 'Fold 1'.

Model	Fold 1 (2022) Reference	Fold 2 (2021)	Fold 3 (2020)	Fold 4 (2019)	Mean (Std)
MLR3	6.68	6.75	6.06	6.92	6.63 (0.33)
MLR6	6.27	6.22	5.66	6.40	6.14 (0.28)
MLR9	5.96	5.89	5.45	6.06	5.84 (0.23)
RF3	5.45	5.71	5.24	5.59	5.50 (0.17)
RF6	4.76	4.84	4.66	5.07	4.83 (0.15)
RF9	4.54	4.68	4.50	4.89	4.65 (0.15)
LGB3	5.45	5.57	5.23	5.51	5.44 (0.13)
LGB6	4.31	4.38	4.32	4.46	4.37 (0.06)
LGB9	4.23	4.32	4.28	4.42	4.31 (0.07)
XGB3	5.51	5.69	5.36	5.62	5.54 (0.13)
XGB6	4.27	4.34	4.31	4.42	4.33 (0.06)
XGB9	4.16	4.26	4.23	4.36	4.25 (0.07)

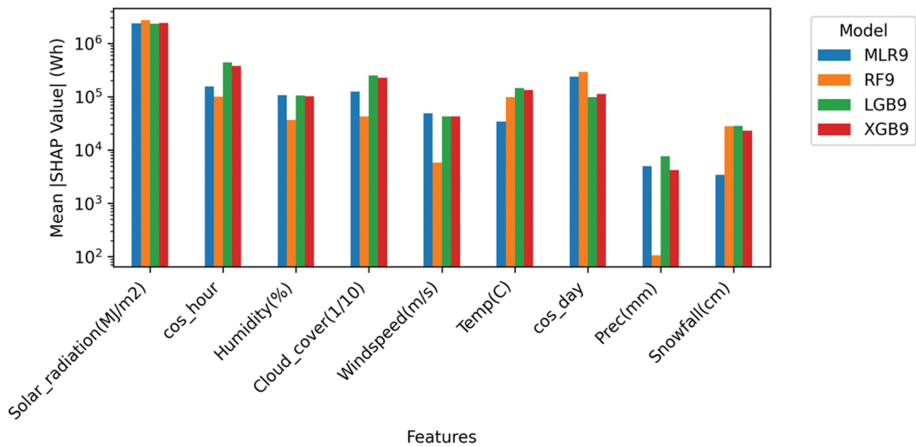

Fig. 3. Feature importance analysis using mean absolute SHAP values for MLR9 (blue), RF9 (orange), LGB9 (green), and XGB9 (red) models for solar PV power estimation in 2022. The y-axis represents the mean absolute SHAP value in log scale, indicating the relative contribution of each feature to the solar PV power estimation.

다. 가장 성능이 우수한 XGB9을 동일 조건의 기준 모델 MLR9와 비교했을 때, 평균 nMAE가 5.84%에서 4.25%으로 약 27% 감소하였다. 특히 XGB9에서 표준편차가 0.07%로 매우 낮게 나타나 연도별 기상 변동성에 관계없이 안정적인 예측 성능을 유지함을 확인하였다. 운량과 강수량 등 상관관계가 낮은 변수 또한 성능 개선에 기여한 것은 이들 변수가 일사량의 감쇠 및 패널 온도 변화를 반영하여 다른 변수들과 비선형적으로 상호작용하기 때문이다. 따라서 9개 변수를 모두 사용할 경우, 모델이 실제 유효 일사량과 패널 온도의 변화를 성공적으로 포착하여 성능이 개선된 것으로 해석된다. 한편, 가장 상관관계가 큰 3개의 변수만 사용했을 때, RF3가 LGB3와 XGB3와 성능이 비슷하게 나타났다. Gradient Boosting은 Random Forest보다 학습 방법이 복잡해서, 비교적 정보가 부족한 상황(변수 3개)에서는 성능이 떨어지는 것으로 보인다.

개별 변수의 중요성을 확인하기 위해, 임의의 발전량 추정치 1000개를 활용하여 Shapley Additive explanation (SHAP) 분석을 수행하였다(Fig. 3). 분석결과 상관계수(Table 3)와 특징 중요도가 대체로 일치하였지만 일부 예외가 있었다. 풍속의 경우 상관계수에 비해 특징 중요도가 낮게 산출되었으며 cos_day의 경우 상관계수에 비해 특징 중요도가 높게 산출되었다. 이는 상관계수가 개별 변수와 목표값 간의 2변수 관계만을 평가하는 반면, SHAP은 다른 모든 변수가 포함된 다변량 환경에서 각 변수의 한계 기여도를 측정하기 때문이다. 풍속은 일정한 상관성을 보이지만 다른 기상변수들이 이미 발전량 변동을 충분히 설명하고 있어 추가적인 기여도가 제한적이다. 반면 cos_day는 하루 동안 값이 일정하기 때문에 시간별 발전량과 직접적인 상관성이 낮지만 계절적인 패턴 학습에서 중요하게 쓰여 기여도는 높게 나타난다. 따라서 보다 효율적이고 정확한 태양광 발전량 예측을 위해서는 단순 상관성 분석을 넘어선 다각적인 변수 중요도 평가가 필요하다.

Figure 4는 특정한 날짜의 발전량을 추정한 예시를 보여준다. 각각은 맑은 날(Fig. 4a), 흐린 날(Fig. 4b), 그리고 비 온 날(Fig. 4c)에 대해 태양광 발전량을 추정한 결과이다. 비 온 날은 일강수량 45.4 mm로 비교적 많은 비가 내린 날을 선택하였다. 구름 없는 맑은 기상조건일 경우, 기준 모델인 MLR9(파란색 점선)은 발전량을 과대 추정하는 경향이 있는 반면, 기계학습 모델들은 발전량을 보다 정확히 추정하였다. 흐린 날의 경우에도 기계학습 모델은 비교적 정확히 발전량을 추정하지만, 구름이 끼기 시작한 13시 이후로는 MLR9도 높은 정확도를 보였다. MLR9의 경우 운량이 적은 경우에 상대적으로 높은 오차를 보였는데

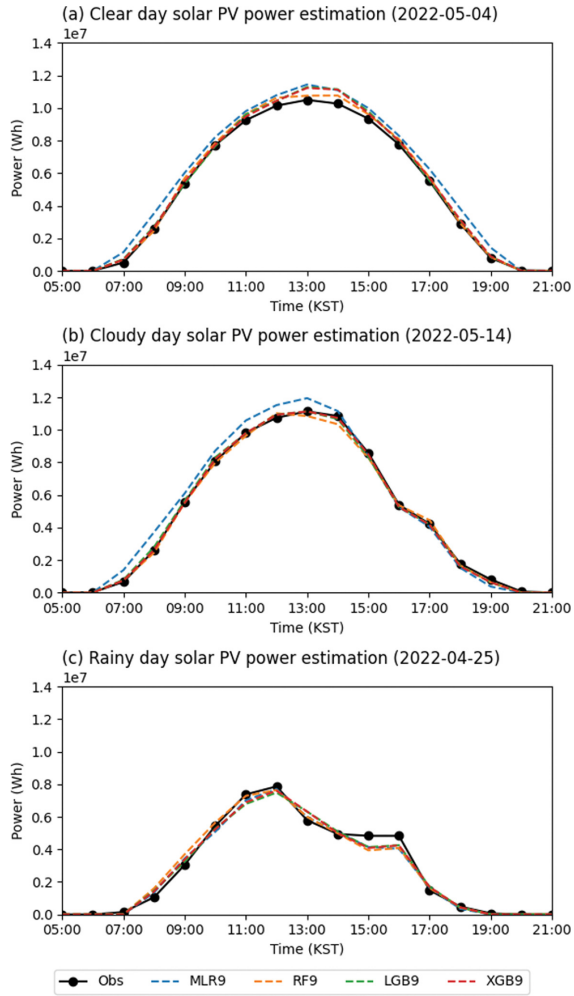

Fig. 4. Solar PV power estimation from meteorological measurements at the Mokpo weather station for (a) clear day, (b) cloudy day, and (c) rainy day. Each line denotes actual solar PV power generation (black line), MLR9 estimation (blue dashed line), RF9 estimation (orange dashed line), LGB9 estimation (green dashed line), and XGB9 estimation (red dashed line).

(Figs. 4a, b 13시 이전), 이는 선형 모델이 운량이 0인 경우 발전 특성을 충분히 반영하지 못했기 때문으로 추정된다. 비 온 날은 가장 복잡한 기상조건으로 인해 예측의 난이도가 높아 모든 모델에서 과대 추정 혹은 과소 추정의 불규칙한 패턴이 나타났다.

검증 기간 전체(2022년)에 대해 시간별 및 월별 RMSE와 nMAE 분포를 Fig. 5에 나타냈다. 시간별 오차는 13시 부근에서 최대값을 보였으며 월별 오차는 여름철과 겨울철에 상대적으로 높게 나타났다. 전반

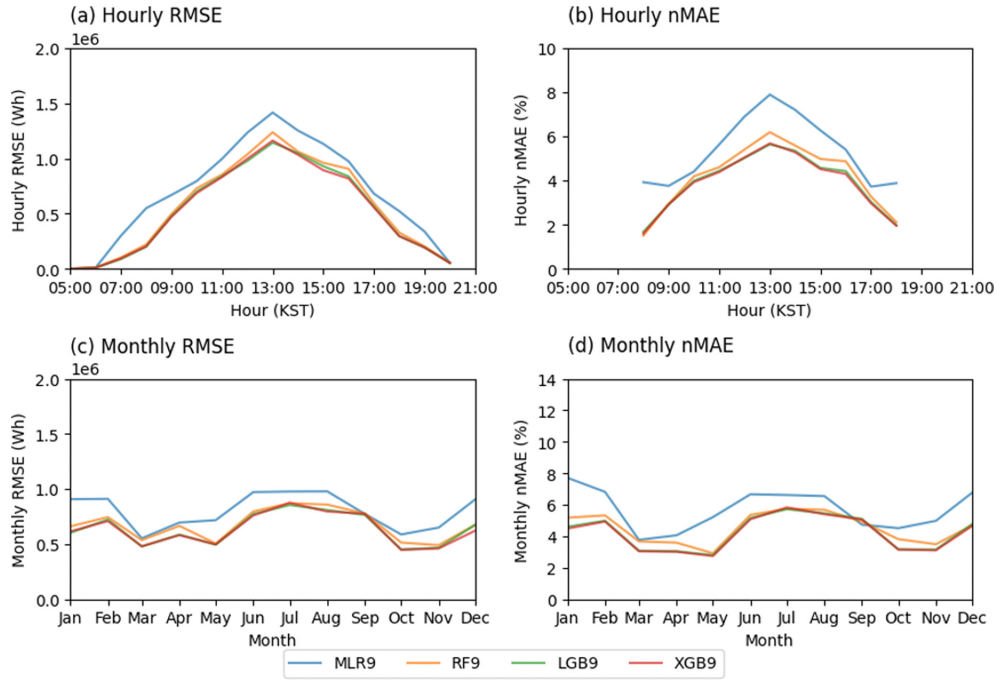

Fig. 5. (a, b) Hourly RMSE and nMAE of MLR9 (blue line), RF9 (orange line), LGB9 (green line), and XGB9 (red line) for solar PV power estimation in 2022. (c, d) Same as (a, b) but for the monthly average of RMSE and nMAE.

적인 오차의 크기는 MLR9에서 가장 크고, RF9는 중간, LGB9와 XGB9은 가장 작은 값으로 유사하게 나타났다. 여름철의 높은 오차는 장마, 태풍, 국지성 호우 등 다양한 기상현상이 복합적으로 작용하여 난이도가 증가한 것으로 판단된다. 겨울철의 높은 오차는 적설의 영향을 일부 받은 것으로 분석된다. 태양광 패널 표면의 적설은 일사량 차단을 통해 발전량을 직접적으로 감소시키며 적설 일수의 상대적 희소성으로 인한 학습 자료 부족도 정확도 저하에 기여한 것으로 판단된다.

4.2 태양광 발전량 예측

실제 예측 상황에서의 성능 비교를 위해 LDAPS 자료를 이용해 예측을 진행하였다(Figs. 6, 7). 2022년에 대해 예측을 수행하였으며, 2019~2021년 자료를 훈련한 발전량 추정 모델과 동일한 모델을 활용하였다. 다만, 발전량 추정(Figs. 4, 5)과 달리 발전량 예측(Figs. 6, 7)은 ASOS 관측값 대신 오차 보정한 LDAPS 예측 자료를 입력값으로 처방하였다. 전날 2100 KST에 초기화된 LDAPS 48시간 예측 자료를 활용하였다.

Figure 4와 동일 사례에 대해 전날 2100 KST에 초기화된 LDAPS 예측 자료를 활용한 태양광 발전량 예측 결과를 Fig. 6에 나타냈다. 맑은 날과 흐린 날의 경우는 ASOS를 이용한 추정 결과와 유사하게 발전량을 예측하였다(Figs. 6a, b). 그러나 강수가 있는 날에는 전반적으로 발전량을 과대 예측하는 경향이 있었다(Fig. 6c). 특히 발전량이 급속하게 감소한 12시부터 16시 사이에는 다른 시간대에 비해 오차가 크게 나타났다. ASOS 자료를 사용했을 때 발전량이 잘 추정된 것을 감안하면(Fig. 4c), 강수에 의한 일사량의 감소를 LDAPS가 과소 모의했기 때문에 오차가 발생한 것으로 판단된다. 실제로 이 시간대에 ASOS는 일사량을 1.18~1.8 MJ m^{-2} 정도로 관측했지만 LDAPS는 2.34~5.05 MJ m^{-2} 정도로 예측하였다.

발전량 예측의 시간별, 월별 오차를 Fig. 7에 제시하였다. ASOS 자료에 기반한 발전량 추정(Fig. 5)에서는 13시에 가장 큰 RMSE과 nMAE가 나타난 것과 달리, LDAPS를 이용한 예측의 경우 RMSE는 12~14시 근처에서, nMAE는 14시에 가장 크게 나타났다. 월별 오차의 경우, 여름과 겨울에 오차가 커지는 전체적인 개형은 ASOS를 활용한 발전량 추정 오차와 유사하지만, 선형 모델과 기계학습 모델 간의 차이는 감소한 것으로 나타났다.

전체 사례에 대한 발전량 예측 오차를 Table 5에 제시하였다. 전체적으로 입력 변수 증가에 따른 예측 성능 향상, 선형 모델에 비해 우수한 기계학습 모델 등 오차의 경향은 ASOS 자료를 이용하여 추정한 결과(Table 4의 Fold 1)와 유사하다. 선형 모델보다는 트

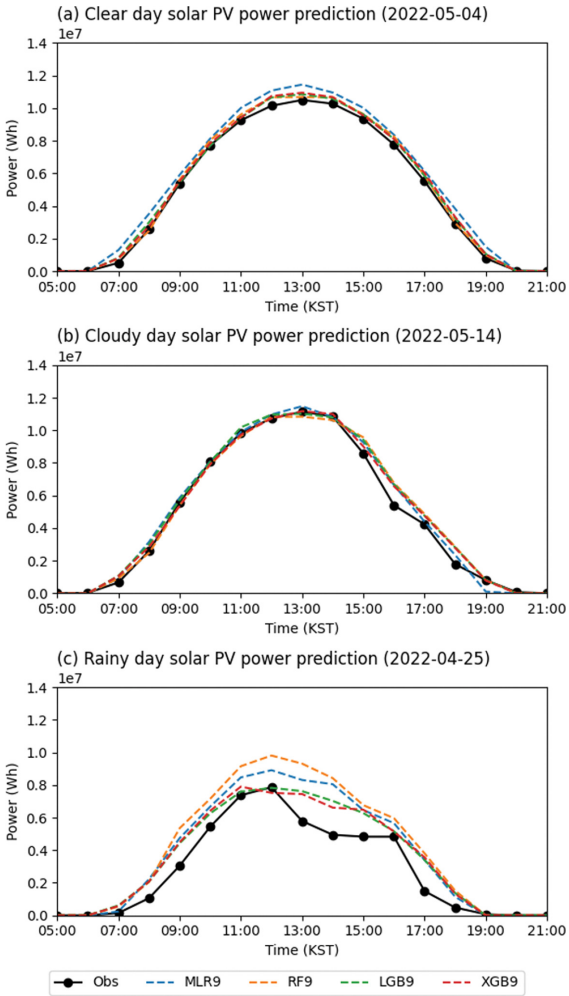

Fig. 6. Same as Fig. 4 but for the prediction using LDAPS forecasts initialized every 2100 KST in 2022.

리기반의 기계학습 모델들이 예측에 있어서 더 좋은 성능을 보였고 기계학습 모델 중에서도 Gradient Boosting 기반의 LGB와 XGB의 성능이 더 우수하였다. 입력된 변수 개수가 증가할수록 nMAE가 감소한다는 경향도 일치했다. 이러한 결과는 제안된 기계학습 접근법의 안정성을 의미한다. 특히 RF가 3개 변수에서 최고 성능을 보이는 패턴이 반복된 것은 모델별 특성이 입력 자료와 관계없이 일관되게 유지됨을 의미한다. 그러나 절대적인 성능 지표 값은 ASOS를 이용한 추정에 비해 LDAPS을 이용한 예측에서 약 3.0~3.7%(입력 변수 9개 모델 기준) 증가하였다. ASOS에서 가장 성능이 우수한 XGB9 모델의 경우에도 nMAE가 4.16%에서 7.89%로 약 1.9배 증가하였다(Table 4 Fold 1과 Table 5 비교). 이러한 오차 증가는 수치예보 모델의 고유한 불확실성이 작용한 결과로 분석된다.

Table 5. nMAE (%) of solar PV power generation prediction with LDAPS forecasts initialized every 2100 KST in 2022. The number following the model abbreviation indicates the number of input variables.

Model	nMAE (%)
MLR3	8.94
MLR6	8.83
MLR9	8.74
RF3	8.48
RF6	8.15
RF9	8.10
LGB3	8.61
LGB6	7.93
LGB9	7.91
XGB3	8.72
XGB6	7.92
XGB9	7.89

초기화 시간에 따른 예측 오차의 민감도를 Table 6에 나타냈다. 초기장에 따른 오차 변화가 크게 나타나진 않지만, 선형모델 MLR9의 경우 모델의 예측 선행시간(lead time)이 짧아 질수록 nMAE가 감소하는 경향이 나타났다. 그 결과 초기장이 18시간 차이(예측 전날 0900 KST 초기화와 예측 당일 0300 KST 초

Table 6. Sensitivity of solar PV power generation prediction to model initialization time in nMAE (%). The values in parentheses for the 2100 KST initialization denote the nMAEs when bias corrections are not applied.

Model	Initialization			
	0900 KST Previous Day	1500 KST Previous Day	2100 KST Previous Day	0300 KST
MLR9	9.08	9.03	8.74 (8.83)	8.75
RF9	8.38	8.38	8.10 (8.31)	8.23
LGB9	8.22	8.23	7.91 (8.04)	7.96
XGB9	8.21	8.19	7.89 (7.97)	7.96

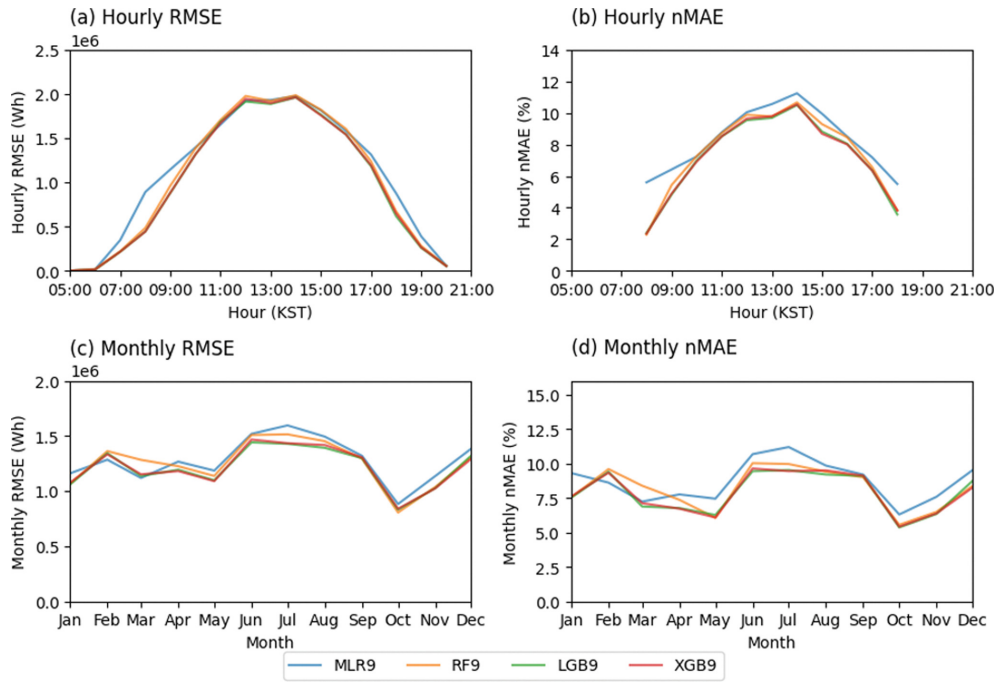

Fig. 7. Same as Fig. 5 but for the prediction using LDAPS forecasts initialized every 2100 KST in 2022.

기화) 날 때 예측오차는 약 0.25% 감소했다. 그러나 기계학습 모델들의 경우, 선행시간 단축에 따른 예측 오차 감소는 예측 전날 2100 KST까지만 나타나고 예측 당일 0300 KST 초기화에서는 오히려 근소하게 증가하였다. 이러한 특징이 기계학습 모델에서 두드러진 원인에 대해서는 추가적인 분석이 요구된다. 주목할 점은 예측 전날 0900 KST 초기화한 LDAPS 자료를 활용한 기계학습모델이 예측 당일 0300 KST 초기화한 LDAPS 자료를 활용한 선형 모델 보다 더 우수한 성능을 보인다는 점이다(MLR9 nMAE = 8.75 대비 XGB9 nMAE = 8.21). 이는 단기 태양광 발전량 예측에서 수치모델과 기계학습을 결합한 하이브리드 예측 체계가 선형 모델 기반보다 더 정확함을 의미한다.

수치모델 자료의 오차 보정 효과를 추가적으로 살펴보았다. Table 6의 2100 KST 괄호 안에 오차 보정하지 않은 LDAPS 예측 자료를 활용한 결과를 제시하였다. 오차 보정을 통해 모든 모델에서 nMAE가 0.08~0.21% 감소하였다. 이는 수치모델의 계통적인 오차를 제거하는 것이 발전량 예측에 있어서 필수적임을 시사한다.

5. 요약 및 토의

본 연구에서는 영암 F1 태양광 발전소와 목포 관측소의 4년(2019년~2022년)간 자료를 활용하여 태양광 발전량을 추정하는 모델을 개발하였다. 선형 모델(MLR)과 3가지 기계학습 모델(RF, XGB, LGB)을 고려하였으며, 연도별 4-겹 교차 검증을 통해 일반화 성능을 검증하였다. 뿐만 아니라 2019년부터 2021년 자료로 학습된 모델을 2022년 기상청 LDAPS 예측 자료에 적용하여 현업에서 태양광 발전량 예측 가능성을 평가하였다.

목포 관측소 ASOS 자료를 활용해 영암 F1 태양광 발전소의 발전량을 추정한 결과, 개별 모델의 성능은 입력 변수가 늘어날수록 향상되었다. 일사량 등 9개의 변수를 모두 활용한 경우, LGB와 XGB의 nMAE가 각각 4.31%와 4.25%로 매우 우수한 성능을 보였다. 예측 전날 2100 KST에 초기화된 LDAPS 예측에 발전량 추정 모델을 적용한 경우에도, LGB와 XGB가 nMAE 7.91%와 7.89%로 우수한 성능을 보였다. RF는 적은 수의 변수에서 우수한 성능을 보여 계산 효율성 측면에서 장점을 보였다. 반면 선형 모델인 MLR은 전반적으로 성능이 낮았다. 특히 주목할 점은 변수를 3개만 사용한 기계학습 모델들이 변수 9개를 사용한 선형 모델 보다 성능이 뛰어나다는 점이다. 이는 태양광 발전 예측에 있어 단순히 사용하는 변수의 개수보다는 변수 간에 존재하는 비선형 상호작용을 얼마나 효과적으로 학습하는지가 예측 정확도를 결정

하는 요인임을 시사한다.

태양광 발전량 예측 현업 적용성을 심화 분석한 결과, LDAPS의 초기화 시간에 따른 예측 정확도의 차이를 확인할 수 있었다. 전반적으로 예측 선행시간이 짧을수록 예측 오차가 감소하는 경향을 보였으나, 오차 감소비율은 비선형적으로 나타났다. 특히 기계학습 모델에서는 예측 전날 2100 KST에 초기화된 예보 자료를 활용했을 때, 선행시간이 더 짧은 예측 당일 0300 KST 초기화된 자료를 활용했을 때보다 더 높은 정확도를 보였다. 이 결과는 예측 시점의 신속성과 정확성을 고려하여 모델을 효율적으로 운용해야 함을 시사한다. 예를 들어, 예측 전날 0900 KST 초기화 예보 자료는 전반적인 발전 동향을 파악하는 조기 예측에, 2100 KST 초기화 예보 자료는 다음날 전력 거래 및 계통 운영에 활용할 필요가 있다. 이러한 체계는 발전소 운영의 유연성과 안정성을 동시에 확보하는 효과적인 방안이 될 것이다.

본 연구는 기계학습 알고리즘과 단기 수치예보를 이용해 태양광 발전량을 예측하는 방법론을 제시하였다. 그러나 다음과 같은 한계점이 존재한다. 먼저, 본 연구는 영암 F1 태양광 발전소에 국한된 것으로 일반화가 어렵다. 향후 전국에 걸쳐 다양한 발전소 자료를 활용하여 모델의 범용성을 검증하는 과정이 필수적이다. 또한 LDAPS 예측 자료의 계통적 오차 보정을 위해, 보다 장기간 자료를 활용한 오차 보정과 검증이 요구된다. 무엇보다도 예측이 어려운 악기상 상황에서 발전량 예측 성능 저하를 개선할 수 있는 방안 모색이 필요하다.

감사의 글

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[NO.RS-2021-II211343, 인공지능대학원 지원(서울대학교)].

REFERENCES

Ahmed, R., V. Sreeram, Y. Mishra, and M. D. Arif, 2020: A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization. *Renew. Sustain. Energy Rev.*, **124**, 109792, doi:10.1016/j.rser.2020.109792.

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama, 2019: Optuna: A Next-Generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), New York, 2623-2631, doi:10.1145/3292500.3330701.

Alkabbani, H., A. Ahmadian, Q. Zhu, and A. Elkamel, 2021: Machine learning and metaheuristic methods for renewable power forecasting: a recent review. *Front. Chem. Eng.*, **3**, 665415, doi:10.3389/fceng.2021.665415.

Antonanzas, J., N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, 2016: Review of photovoltaic power forecasting. *Sol. Energy*, **136**, 78-111, doi:10.1016/j.solener.2016.06.069.

Breiman, L., 2001: Random Forests. *Machine Learning*, **45**, 5-32, doi:10.1023/A:1010933404324.

Chen, T., and C. Guestrin, 2016: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), New York, NY, USA, 785-794, doi:10.1145/2939672.2939785.

Choi, J., S.-W. Son, S. Lee, and S. Park, 2024: Advancing global solar photovoltaic power forecasting with subseasonal climate outlooks. *Renew. Energy*, **237**, 121803, doi:10.1016/j.renene.2024.121803.

_____, _____, and S.-Y. Jun, 2025, Climatic characteristics and complementarity of solar and wind energy in Korea: potential benefits of hybrid power generation for mitigating energy droughts, *Atmosphere*, **35**, 315-329, doi:10.14191/Atmos.2025.35.3.315.

Gaboitaolelwe, J., A. M. Zungeru, A. Yahya, C. K. Lebekwe, D. N. Vinod, and A. O. Salau, 2023: Machine learning based solar photovoltaic power forecasting: a review and comparison. *IEEE Access*, **11**, 40820-40845, doi:10.1109/ ACCESS.2023.3270041.

Gottwald, D., M. Parmar, and A. Zureck, 2024: Forecasting solar power generation: A comparative analysis of machine learning models. 2024 International Conference on Renewable Energies and Smart Technologies, 1-5, doi:10.1109/REST59987.2024.10645372.

Gudmundsson, L., J. B. Bremnes, J. E. Haugen, and T. Engen-Skaugen, 2012: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods. *Hydrol. Earth Syst. Sci.*, **16**, 3383-3390, doi:10.5194/hess-16-3383-2012.

Iheanetu, K. J., 2022: Solar Photovoltaic Power Forecasting: A Review. *Sustainability*, **14**, 17005, doi:10.3390/su142417005.

KEA, 2024: Summary of 2023 New & Renewable Energy Supply Statistics (2024 Official Announcement). Korea Energy Agency Statistics Analysis Office, 1-4 [Available onlime at: https://www.knrec.or.kr/biz/pds/statis-

tic/view.do?no=390].

Khaire, M. V., A. G. Thosar, and V. N. Pande, 2023: Prediction of Solar Power Generation Using NWP and Machine Learning. *2023 3rd Asian Conference on Innovation in Technology*, 1-6, doi:10.1109/ASIANCON58793.2023.10270030.

Ki, J. H., S.-J. Baek, J.-Y. So, H.-G. Eom, and J.-H. Shin, 2023: Solar power generation prediction using machine learning and study of power generation using solar tracking panels. Trans. Korean Soc. Mech. Eng. B, **47**, 55-62, doi:10.3795/KSME-B.2023.47.1.055.

Kim, C. K., H.-G. Kim, Y.-H. Kang, and C.-Y. Yun, 2019: Evaluation of UM-LDAPS prediction model for daily ahead forecast of solar power generation. *J. Korean Solar Energy Soc.*, **39**, 71-80, doi:10.7836/kses.2019.39.2.071.

KMA, 2025: Local Data Assimilation and Prediction System (LDAPS) information. Korea Meteorological Administration, accessed August 14, 2025 [Available online at https://data.kma.go.kr/data/rmt/rmtList.do?code=340&pgmNo=65].

Lee, Y., D. Kim, W. Sin, C. Kim, H. Kim, and S. W. Han 2021: A comparison of machine learning models in photovoltaic power generation forecasting. *J. Korean Inst. Ind. Eng.*, **47**, 444-458, doi:10.7232/JKIIE.2021.47.5.444.

Markovics, D., and M. J. Mayer, 2022: Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. *Renew. Sustain. Energy Rev.*, **161**, 112364, doi:10.1016/j.rser.2022.112364.

MOTIE, 2025: The 11th basic electricity supply and demand plan. Ministry of Trade Industry and Resources, 120 pp [Available online at https://www.motie.go.kr/kor/article/ATCLc01b2801b/70152/view#].

Nguyen, H. N., Q. T. Tran, C. T. Ngo, D. D. Nguyen, and V. Q. Tran, 2025: Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms. *PLOS ONE*, **20**, e0315955, doi:10.1371/journal.pone.0315955.

Oh, S.-G., S.-W. Son, Y.-H. Kim, C. Park, J. Ko, K. Shin, J.-H. Ha, and H. Lee, 2024: Deep learning model for heavy rainfall nowcasting in South Korea. *Weather Clim. Extremes*, **44**, 100652, doi:10.1016/j.wace.2024.100652.

Shim, C.-Y., G.-M. Baek, H.-S. Park, and J.-Y. Park, 2024: Comparison of solar power generation forecasting performance in Daejeon and Busan based on preprocessing methods and artificial intelligence techniques: using meteorological observation and forecast data. *Atmosphere*, **34**, 177-185, doi:10.14191/ATMOS.2024.34.2.177.

Sobri, S., S. Koohi-Kamali, and N. Abd Rahim, 2018: Solar photovoltaic generation forecasting methods: a review. *Energy Convers. Manag.*, **156**, 459-497, doi:10.1016/j.enconman.2017.11.019.

Thipwangmek, N., K. Woradit, N. Suetrong, and N. Promsuk, 2024: Feature selection approaches for short-term solar photovoltaic power forecasting. *2024 13th International Conference on Renewable Energy Research and Applications*, Nagasaki, Japan, 252-257, doi:10.1109/ICRERA62673.2024.10815326.