Al-Mutawassit Journal for Basic and Applied Sciences (MJBAS)

Volume 1, Issue 1, 2025
Page No: 20-27

Website: https://www.mutawassitpub.com/index.php/mjbas

Article History:

Received:
09 May 2025

Accepted:
08 August 2025

Published:
25 September 2025

Forecasting Solar Power Generation Using Real Meteorological Data and Machine Learning Techniques

Abdulgader Alsharif *
Department of Electric and Electronic Engineering, College of Technical Sciences, Sebha, Libya

*Corresponding author: abduulgaderalsharif@gmail.com

Abstract

Accurate solar power forecasting is essential for reliable grid integration and effective planning. This paper surveys methods that use real-world weather data (such as irradiance, temperature, humidity, wind, etc.) and modern machine learning (ML) to predict photovoltaic (PV) output. We review datasets and preprocessing steps, and compare models including linear regression, ensemble trees, support vector machines, and neural networks (e.g. LSTM, CNN). Key factors like solar irradiance and temperature are identified as dominant inputs, with wind and humidity having weaker correlations. Several case studies are presented: for example, Essam *et al.* (2022) used an NREL dataset from Florida and found an artificial neural network gave $R^2 \approx 0.9988$, outperforming other algorithms. Chakraborty *et al.* (2023) applied ensemble methods in India and achieved ~96% accuracy for PV power with stacking/voting models. Balal *et al.* (2023) evaluated eight models on a Texas dataset, with Random Forest and LSTM yielding the best results ($R^2 \approx 0.977$ and 0.975). A model using seven years of data from *Renewables.ninja* in Greece improved capacity factor forecasts via ANN. We also note that convolutional and recurrent neural nets (ConvLSTM) can capture spatio-temporal patterns; for instance, Shah *et al.* (2024) reached $R^2 \approx 0.969$ using a ConvLSTM2D with weather and air quality features.

Keywords: Solar power forecasting, photovoltaic generation, meteorological data, machine learning, neural networks, ensemble methods.

Introduction

Solar photovoltaic (PV) generation has grown rapidly worldwide. Precise forecasting of PV output is critical to balance supply and demand and to integrate renewables into the grid. However, solar power is inherently intermittent due to changing weather. Machine learning (ML) offers powerful tools to model these complex patterns. In recent years, many studies have used real-world meteorological data such as solar irradiance, air temperature, humidity, wind speed and direction and advanced ML models to predict PV output (NREL, 2018). For example, maps of annual average solar irradiance (Figure 1) highlight regions with high sunlight potential. Accurate models can exploit spatial and temporal patterns in such data.

PV output is strongly correlated with incoming irradiance. Analysis by Aouidad & Bouhelal shows that global horizontal irradiance exhibits a strong correlation with generated power. Other factors like ambient temperature and wind have weaker correlations. This makes weather data very informative for prediction. For instance, one open dataset ("UNISOLAR") provides 15-minute PV generation and weather data from university campuses in Australia. Studies often use such rich datasets to train ML models. We also rely on large reanalysis or satellite-

based resources: e.g., the *Renewables.ninja* database provides meteorology and PV output for global sites. In one case, Jbeily *et al.* (2025) used seven years of Renewables.ninja data for Athens, Greece, to train an ANN, significantly improving forecasts of the PV capacity factor.

In summary, forecasting PV power with ML involves three key elements: reliable meteorological inputs, appropriate ML models, and careful evaluation. The rest of the paper reviews these components. We first describe machine learning methods used, then detail data sources and preprocessing, and finally survey recent experimental results and model comparisons.

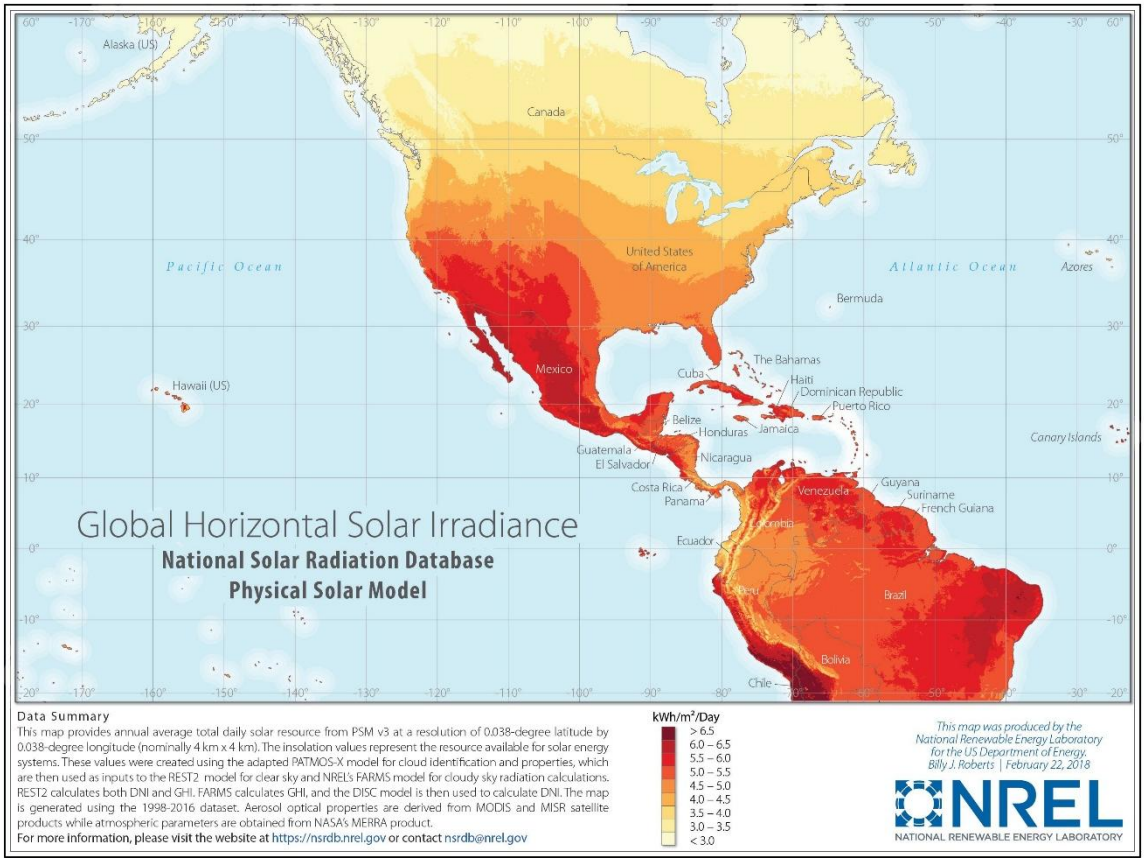

Figure 1 Annual average global horizontal solar irradiance (W/m²) from NREL's NSRDB (Physical Solar Model v3, 1998–2016). Bright regions (yellow) receive the most solar energy. This resource map underlines the need for location-specific solar forecasting.

Machine Learning Methods for Solar Forecasting

A variety of ML techniques have been applied to solar forecasting. Traditional statistical methods (linear regression, persistence, time series ARIMA) serve as benchmarks, but modern ML often outperforms them. Supervised learning approaches treat forecasting as either a regression (predicting continuous power) or classification problem (for example, predicting low/medium/high power categories). In regression, models learn the mapping x↦Px \mapsto Px↦P from features xxx (weather data, time indices) to power *P*.

Among ML algorithms, k-nearest neighbors (k-NN) is a simple non-parametric method. It predicts new outputs by averaging the outputs of the k most similar past weather conditions. Its concept is illustrated in Figure 2.

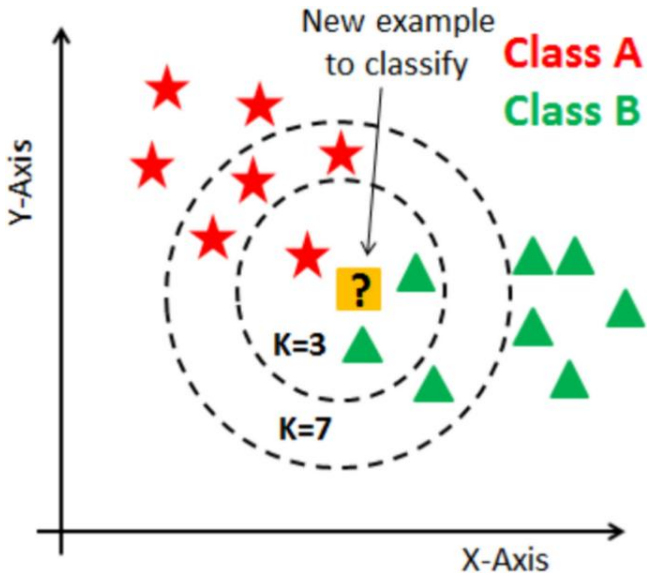

Figure 2 k-nearest neighbors concept. Given a test point, the k closest training points (green) vote to classify or predict its value. K-NN methods essentially store all training data and have shown good efficacy in various tasks despite their simplicity.

Ensemble tree methods like Random Forests (RF) and Gradient Boosting are widely used. A Random Forest grows many decision trees on bootstrapped data and averages their outputs. This captures nonlinear patterns and reduces variance. Figure 3 illustrates a Random Forest ensemble. Chakraborty *et al.* (2023) emphasize such ensemble learning can boost accuracy, achieving ~96% accuracy with stacking of multiple models.

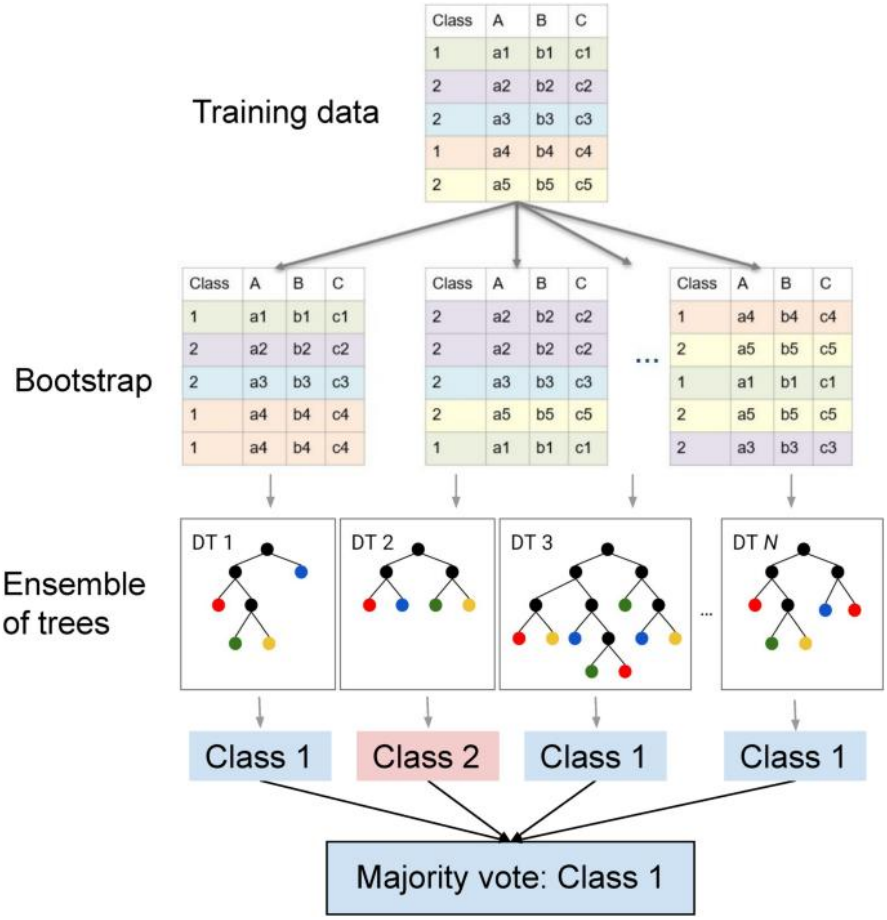

Figure 3 Random Forest ensemble. Multiple decision trees (DT1, DT2, DT3…) are trained on bootstrapped data; their predictions are averaged (majority vote) to produce the final output. This reduces overfitting compared to a single tree.

Deep learning has gained traction for time series. Recurrent neural networks (RNNs), especially LSTM (Long Short-Term Memory) networks, can capture temporal dependencies. An LSTM cell manages long-term memory through gates (input, forget, output) as shown in Figure 4. In practice, LSTMs often outperform simpler models on solar data. For example, Balal *et al.* (2023) found LSTM performed almost as well as RF, with $R^2 \approx 0.975$. Convolutional networks (CNNs) and hybrids (ConvLSTM) have also been used to extract spatial and temporal features, especially when satellite/cloud images are available.

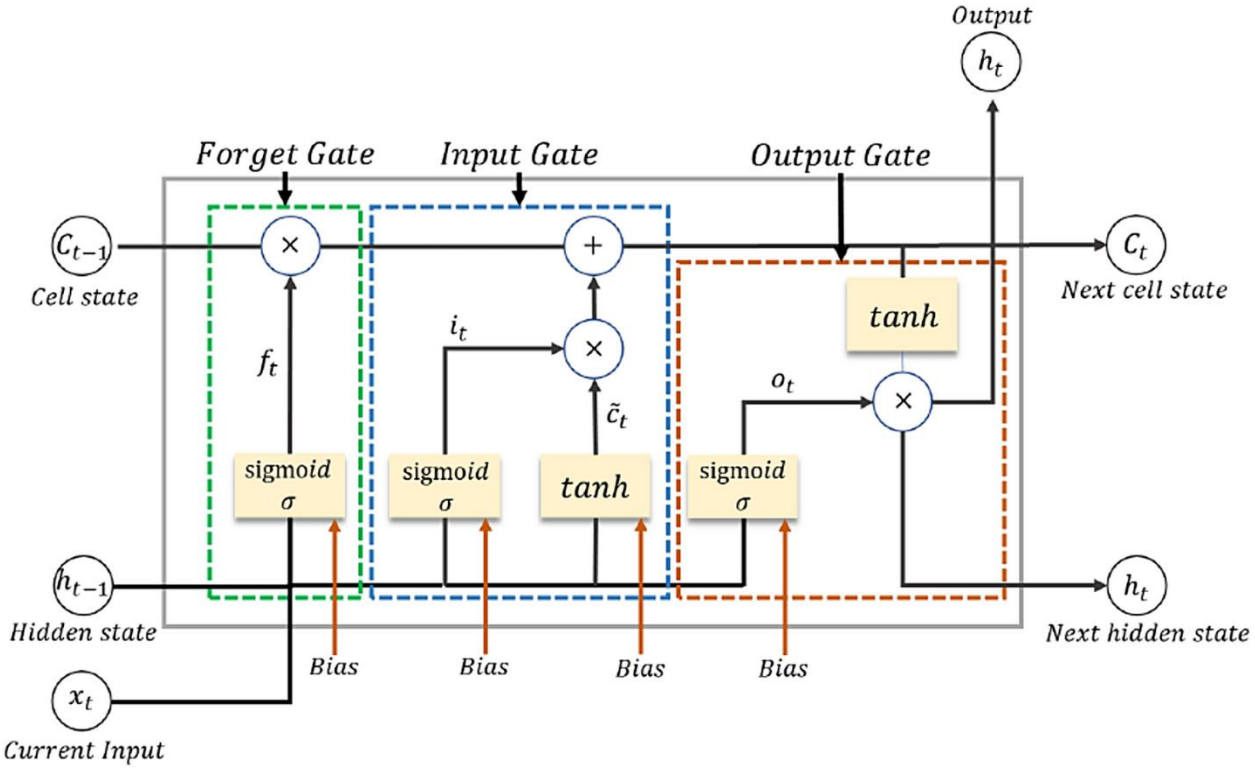

Figure 4 LSTM cell architecture. The input (current observation) and previous state pass through forget, input, and output gates controlling the cell state. LSTM's gating allows learning of long-range temporal patterns, beneficial for time-series like PV output.

In some studies, Gaussian Processes (GP) or Support Vector Regression (SVR) are used for robustness, and extreme learning machines or ANNs are also tested. Hybrid approaches combine models, for instance using wavelet transforms or boosting. Across multiple reviews, neural networks (especially deep nets) and ensemble trees consistently give lowest errors. Table 1 (below) compares representative results: Essam *et al.* (2022) found an ANN achieved $R^2 \approx 0.9988$ on NREL PV data, while Balal *et al.* (2023) reported $R^2 \approx 0.977$ for RF in Lubbock, Texas, and Chakraborty *et al.* (2023) achieved ~96% accuracy using stacked ensembles. Overall, results indicate ML models can match or exceed traditional approaches in accuracy.

Table 1 Performance metrics from recent studies.

Model	Metric	Value	Study (Year)
ANN (Essam et al.)	R^2	0.9988	Essam et al. (2022)
Random Forest (Balal et al.)	R^2	0.977	Balal et al. (2023)
LSTM (Balal et al.)	R^2	0.975	Balal et al. (2023)
Stacking Ensemble (Chakrab.)	Accuracy	≈0.96	Chakraborty et al. (2023)

Data Sources and Experimental Setup

To build forecasting models, we need historical PV output and corresponding weather data. Datasets range from small-scale plant logs to large multi-year compilations. Public sources include NREL's NSRDB (satellite-based irradiance and weather for the Americas), *Renewables.ninja* (simulated PV outputs worldwide), and various university/plant installations. For example, the UNISOLAR dataset provides 15-minute PV generation and weather data from five La Trobe University campuses (Australia, 2020–2022). These real datasets often include global horizontal irradiance (GHI), cell temperature, wind speed, etc. Typical preprocessing involves cleaning (filling missing), normalization, and feature engineering (time of day, lag features).

Exploratory analysis often begins with correlation studies. Figure 5 shows a typical input–output correlation map from Aouidad & Bouhelal (2024). It confirms that solar irradiance has the strongest correlation with output, while temperature, wind speed, humidity, and rainfall are much weaker (near zero). This guides feature selection and model design.

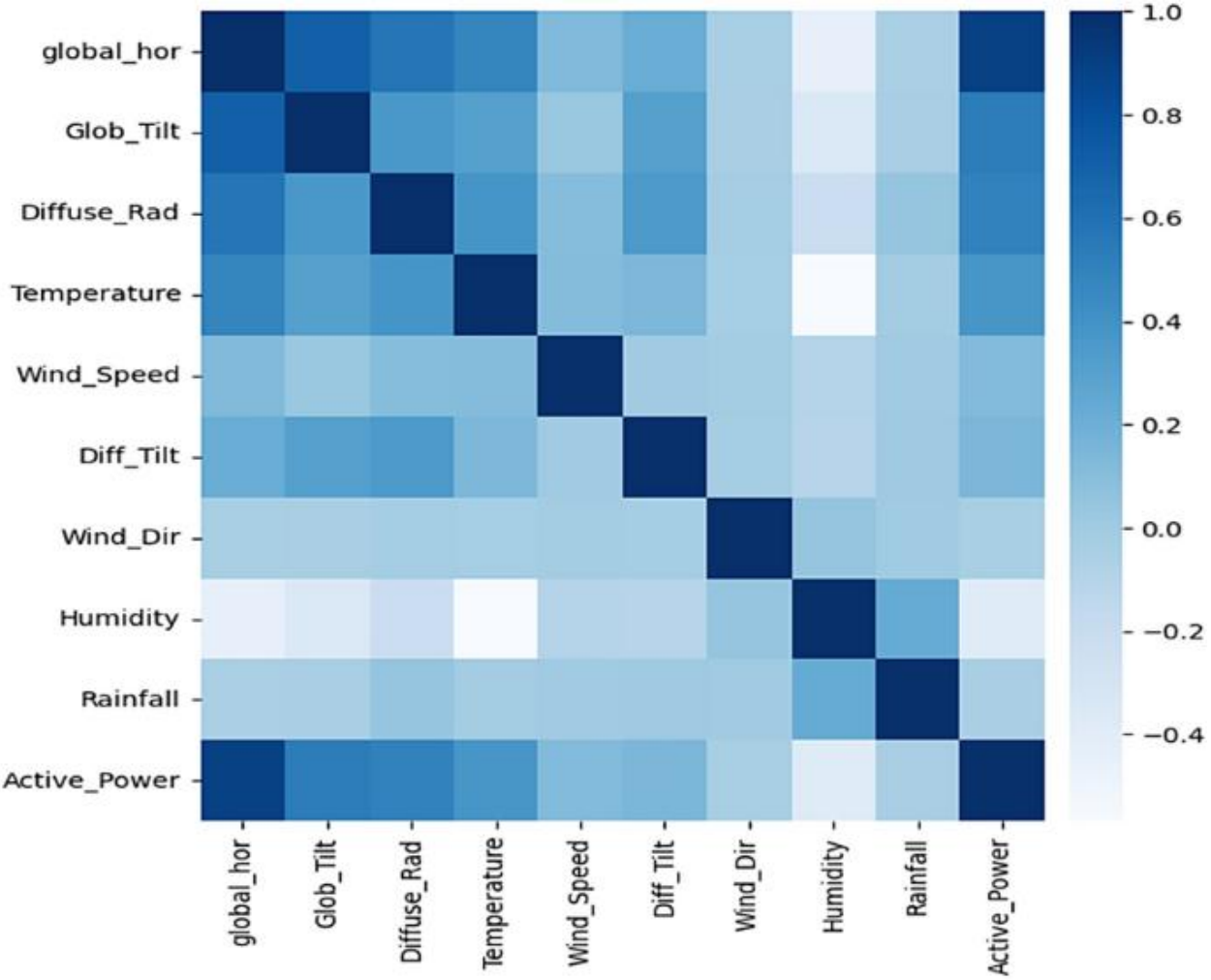

Figure 5 Correlation between weather inputs and PV power (Aouidad & Bouhelal, 2024). Dark blue indicates strong positive correlation. Global horizontal irradiance ("global_hor") has the highest correlation with power. Other factors (temperature, humidity, wind) show weak correlation.

We split data into training/validation/testing (e.g., 70/20/10%) and evaluate forecasts using metrics like MAE, RMSE, and R^2. These metrics quantify average errors and variance explained. For example, if a model predicts power $Pi\hat{}$ vs actual Pi, the mean squared error is $MSE = \frac{1}{N}\sum(Pi - Pi\hat{})^2$, and R^2 assesses how much variance is captured.

Experimental results across multiple sites show consistent patterns. In a Florida PV plant study, Essam *et al.* found an ANN produced MAE=0.4693 W and R^2=0.9988, beating RF, SVR, decision trees, and LSTM. In India, Chakraborty *et al.* used an ensemble (stacking and voting) on Eastern India data; their stacking model reached ~96% classification accuracy. Balal *et al.* tested eight ML/DL methods on an extensive Texas plant dataset. They observed that Random Forest Regression and LSTM gave the best forecasts (MSE≈2.06% and 2.23%, respectively, R^2≈0.977 and 0.975). These findings highlight that no single model dominates; performance varies by location and data. However, ensemble methods and deep nets generally excel.

We also note practical workflows: As illustrated in prior work, one must preprocess raw data (clean, normalize), possibly decompose time series (e.g. wavelets), and then train multiple models for comparison. The best model is selected based on validation metrics and tested on unseen data. Domain knowledge can be used to engineer features (e.g., separating irradiance into direct/diffuse components, using numerical weather prediction forecasts as inputs).

A few studies incorporate broader data, such as air quality. For example, Shah *et al.* (2024) included PM2.5/AQI levels along with meteorological features in a ConvLSTM2D network, noting that air pollution reduced irradiance by 29%. Their hybrid ConvLSTM achieved $R^2 \approx 0.9691$ and low errors, showing that non-meteorological data can also matter.

Results and Discussion

The surveyed experiments consistently show ML models can forecast PV output with high accuracy. For instance, Aouidad & Bouhelal (2024) demonstrated that KNN, random forest, and LSTM models can capture the time-series patterns effectively on Australian data. Figure 6 (from their work) illustrates how PV output peaks in spring/summer, correlating with cooler temperatures that year. Seasonal trends and daily cycles are learned by ML models to make predictions.

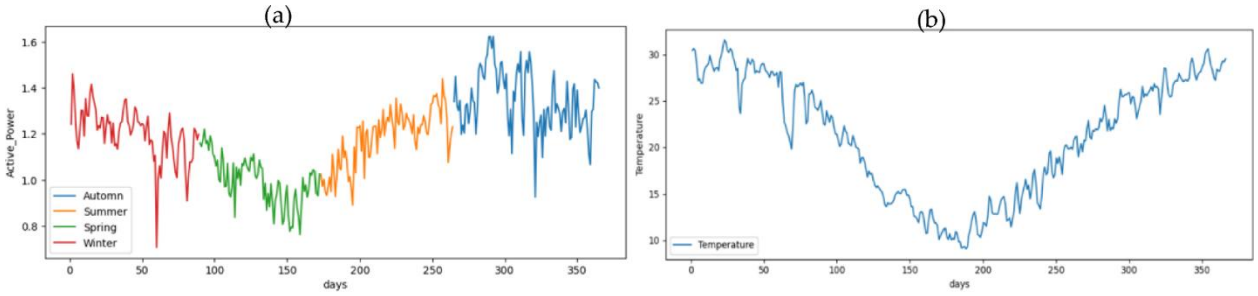

Figure 6 (a) Average temperature (°C) and (b) average PV active power (kW) by season, throughout the year. Spring and summer have lower temperatures and higher PV output, reflecting typical seasonal PV behavior. Forecast models must learn these patterns.

Table 1 summarizes comparative model performance. Several trends emerge: ensemble models like Random Forest often outperform simple regression due to handling nonlinear interactions. Deep networks (LSTM, CNN) further improve accuracy in many cases because they capture sequential dependencies and complex input-output mappings. For example, in one large-scale test (six PV plants in China), a graph neural network (GraphCNN) outperformed TCN, LSTM, and MLP models on day-ahead forecasts by explicitly modeling spatial correlations. The adopted Fourier GNN even reduced MAE to ~5.02 and achieved lowest RMSE in all weather scenarios. Figure 7 shows that model's architecture.

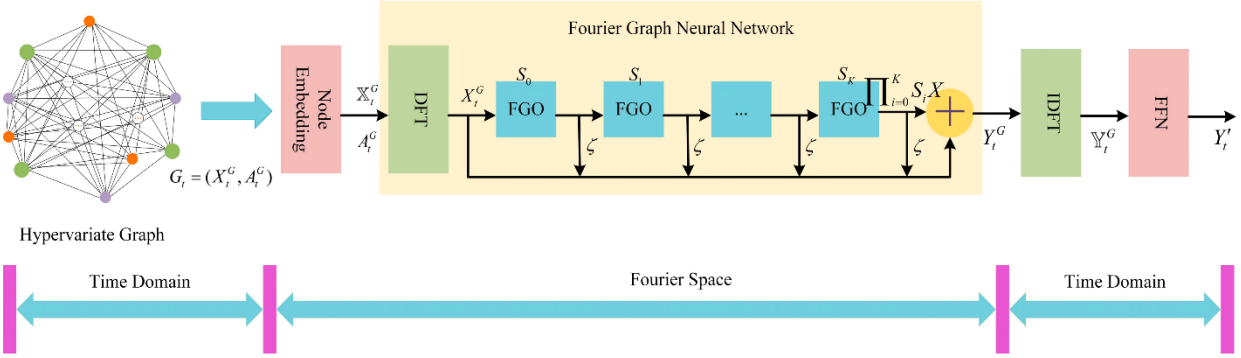

Figure 7 Fourier Graph Neural Network (FourierGNN) architecture for multi-site PV forecasting (Jing et al., 2024). Input time series from multiple locations are transformed to Fourier space, processed by FourierGNN layers, and then output predictions via inverse transform. This captures spatio-temporal correlations efficiently.

The trained models were validated in sunny, cloudy, and mixed conditions. As an example, Figure 8 compares actual vs. predicted power for various models on a sunny day. The FourierGNN's predictions (red) align more closely with the true curve than others. Neural models reduced peak prediction errors by better following the waveforms. This illustrates the practical impact: improved forecasts can reduce reliance on reserves and smooth grid operation.

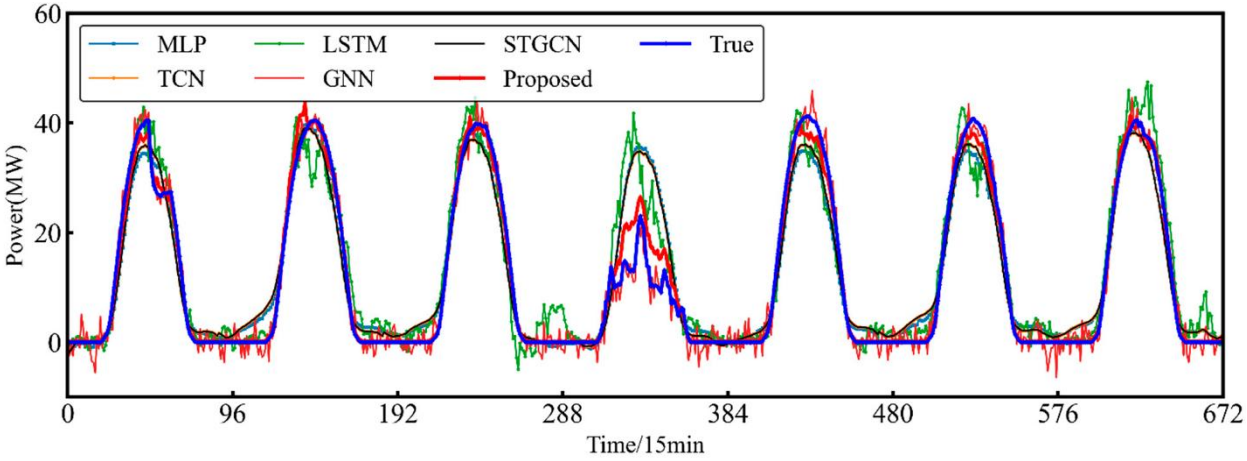

Figure 8 Example Day-ahead forecasts on a sunny day from multiple models (FourierGNN in red, others in grey/blue) versus actual PV output. The advanced FourierGNN model more closely tracks the true peaks and valleys, reducing errors (Jing et al., 2024).

Overall, the evidence shows that ML using meteorological data yields accurate solar forecasts. Key findings include: *solar irradiance* is by far the most important predictor, combining multiple weather inputs improves accuracy slightly, ensemble and deep methods outperform linear models in most cases, performance metrics like R^2 typically exceed 0.95 for top models, demonstrating their reliability.

It is also important to consider forecasting horizon. Most studies focus on short-term (hourly) or day-ahead predictions, where weather forecasts are available. Some work extends to sub-hourly or longer horizons. The general approach remains similar: use past PV and weather data to predict future output. Complex models (e.g. LSTMs) are particularly suited for longer horizons, while simpler models may suffice for one-step-ahead forecasts.

Finally, public datasets and code are becoming available, enabling replication. For example, the UNISOLAR dataset (Australia) and the Renewables.ninja data allow others to test models across climates. Kaggle competitions and GitHub projects often supply solar power datasets (e.g. solar parks in India, or national datasets from NREL), facilitating benchmarking.

Conclusion

Forecasting solar power generation accurately is crucial for energy planning and grid stability. We have reviewed how real meteorological data (irradiance, temperature, etc.) can be combined with machine learning to predict PV output. Results from recent literature show that ML models – especially ensemble trees and deep neural networks can attain very high accuracy (often $R^2 > 0.95$) when fed quality weather data. Figures and table presented here illustrate key patterns (seasonal trends, feature correlations) and model performance (error metrics and comparisons).

Our review highlights that global horizontal irradiance is the dominant input, but including additional inputs (cloud cover, humidity, air quality) can further refine forecasts. Models like Random Forests, CNNs, and LSTMs consistently perform well across various datasets. For instance, case studies in Texas and Florida achieved near-perfect R^2 using Random Forests and ANNs. Advanced architectures like Graph Neural Networks (FourierGNN) capture multi-site correlations to further improve accuracy.

In practice, these forecasting methods can be embedded into grid operations and smart inverters. They provide day-ahead and short-term predictions that help operators schedule backup power and manage storage. The continuous expansion of open data (like NSRDB and Renewables.ninja) and accessible ML frameworks means these techniques can be widely adopted.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

1. Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization. *Renewable and Sustainable Energy Reviews, 124*, 109792.
2. Abdulgader Alsharif. (2025). Quantitative Assessment of Energy Efficiency and Range Variability in Electric Vehicles: A Meta-Analysis of Published Experimental Studies (2023-2025). Al-Imad Journal of Humanities and Applied Sciences (AJHAS), 1(1), 21-30.
3. Aouidad, H. I., & Bouhelal, A. (2024). *Machine learning-based short-term solar power forecasting: a comparison between regression and classification approaches using extensive Australian dataset.* Sustainable Energy Research, 11, 28.
4. Balal, A., Pakzad Jafarabadi, Y., Demir, A., Igene, M. O., Giesselmann, M., & Bayne, S. (2023). Forecasting solar power generation utilizing machine learning models in Lubbock. *Emerging Science Journal, 7*(4), 1052–1062.
5. Mohamed Belrzaeg, & Maamar Miftah Rahmah. (2024). A Comprehensive Review in Addressing Environmental Barriers Considering Renewable Sources Integration and Vehicle-to-Grid Technology. Libyan Journal of Contemporary Academic Studies, 2(1), 1-6
6. Chakraborty, D., Mondal, J., Barua, H. B., & Bhattacharjee, A. (2023). Computational solar energy—Ensemble learning methods for prediction of solar power generation based on meteorological parameters in Eastern India. *Renewable Energy Focus, 44*, 277–294.
7. Abdulgader Alsharif (2025). AI-Based Spatiotemporal Analysis of Solar and Wind Energy Potential Using Satellite and Ground Sensor Data. Scientific Journal for Publishing in Health Research and Technology, 1(1), 01-07.
8. Essam, Y., Ahmed, A. N., Ramli, R., Chau, K.-W., Ibrahim, M. S. I., Sherif, M. M. H., Sefelnasr, A., & El-Shafie, A. (2022). Investigating photovoltaic solar power output forecasting using machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics, 16*(1), 2002–2034.
9. Jbeily, V., Moustris, K., & Spyropoulos, G. (2025). Forecasting of the capacity factor of a photovoltaic system using artificial neural networks. *Environmental and Earth Sciences Proceedings, 35*(1), 31.
10. Mahrouch, A., Asghar, R., Fulginei, F. R., & Quercio, M. (2024). Artificial neural networks for photovoltaic power forecasting: a review of five promising models. *IEEE Access, 12*, 90461–90481.
11. National Renewable Energy Laboratory (NREL). (2018). *National Solar Radiation Database (1998–2016), Physical Solar Model dataset.* Available via NREL NSRDB (https://nsrdb.nrel.gov/).
12. Abdussalam Ali Ahmed (2025). Hybrid AI Models for Forecasting and Optimizing Solar Energy Generation Under Varying Weather Conditions. Scientific Journal for Publishing in Health Research and Technology, 1(1), 35-41.
13. Wimalaratne, S., Haputhanthri, D., Kahawala, S., Gamage, G., Alahakoon, D., & Jennings, A. (2022). UNISOLAR: An open dataset of photovoltaic solar energy generation in a large multi-campus university setting. In *2022 15th International Conference on Human System Interaction (HSI)* (pp. 315–320). IEEE.
14. Shah, A., Viswanath, V., Gandhi, K., & Patil, N. M. (2024). Predicting solar energy generation with machine learning based on air quality and weather features. *arXiv:2408.12476*.
15. Abdussalam Ali Ahmed (2025). Synergizing Renewable Energy and Electric Vehicles: An Experimental Analysis of Grid Integration, Charging Optimization, and Environmental Impact. Journal of Insights in Basic and Applied Sciences, 1(1), 35-43.
16. Abdulgader Alsharif. (2025). Hybrid Solar-Piezoelectric Pavement Systems: A Dual-Mode Approach to Renewable Energy Harvesting and Sustainable Infrastructure. Libyan Journal of Health, Science, and Development (LJHSD), 1(1), 23-31.
17. Taha Muftah Abuali, & Abdussalam Ali Ahmed. (2025). Performance Evaluation and Experimental Optimization of a Hybrid Solar–Wind Energy System under Variable Climatic Conditions. Journal of Libyan Academy Bani Walid, 1(2), 22–38 .
18. Mohamed Belrzaeg, & Abdussalam Ali Ahmed. (2023). A The Adoption of Renewable Energy Technologies, Benefits, and Challenges: Mini-Review. Libyan Journal of Contemporary Academic Studies, 1(1), 20-23.
19. Abdussalam Ali Ahmed (2025). Hybrid Tidal-Wave Systems with Advanced Materials for Efficient and Durable Renewable Ocean Energy. Libyan Open University Journal of Applied Sciences (LOUJAS), 1(1), 29-43.