Article

An Explainable Machine Learning Framework for Short-Term Solar Power Forecasting in Smart Energy Systems

Author One [1], Author Two [2] and Author Three [2,*]

[1] Department of Electrical Engineering, International Institute of Technology, City 00000, Country; author.one@example.com
[2] Department of Computer Science, Global Research University, City 00000, Country; author.two@example.com
* Correspondence: corresponding.author@example.com

Abstract

Accurate short-term solar power forecasting is important for grid stability, energy scheduling, and the efficient integration of renewable generation. However, photovoltaic output varies rapidly because of cloud cover, temperature, irradiance, and seasonal conditions. This study presents an explainable machine learning framework that combines gradient-boosting regression with temporal feature engineering and post-hoc explanation methods. Historical power observations, solar irradiance, ambient temperature, humidity, wind speed, and calendar variables are processed through a structured data-cleaning and normalization pipeline. The predictive model is evaluated using a synthetic demonstration dataset representing one year of measurements from multiple photovoltaic sites. Results show that the proposed framework achieves lower mean absolute error and root mean square error than persistence, linear-regression, and random-forest baselines. Feature-attribution analysis indicates that recent power output, global horizontal irradiance, cloud index, and hour of day make the largest contributions to the predictions. The findings demonstrate how accurate forecasting can be combined with transparent model interpretation to support operational decision-making in smart energy systems. This manuscript contains fictional data and results and was created solely as a professional portfolio demonstration.

Keywords: solar power forecasting; explainable artificial intelligence; machine learning; renewable energy; smart grids; gradient boosting

1. Introduction

The increasing deployment of photovoltaic generation is transforming the operation of modern electrical grids. Solar energy offers substantial environmental and economic benefits, but its intermittent nature creates challenges for generation scheduling, reserve allocation, and voltage regulation. Short-term forecasting can reduce these challenges by providing grid operators with advance estimates of expected power production.

Conventional solar forecasting approaches include persistence models, physical models, statistical regression, and numerical weather prediction. Persistence forecasting assumes that near-future production will remain similar to the most recent observation. This method is simple and often effective over very short horizons, but its accuracy decreases during rapidly changing weather conditions. Physical models use information about panel characteristics, solar geometry, and atmospheric conditions, whereas statistical approaches learn relationships directly from historical measurements.

Machine learning provides a flexible alternative because it can model nonlinear interactions among environmental and temporal variables. Tree-based ensemble models, neural

Received:
Revised:
Accepted:
Published:

https://doi.org/10.3390/1010000

networks, and hybrid forecasting systems have shown strong performance in renewable-energy applications. Nevertheless, many high-performing models are difficult to interpret. Limited interpretability can reduce confidence in automated predictions, particularly when forecasts support operational decisions.

This study presents an explainable machine learning framework for short-term solar power forecasting. The framework integrates data preprocessing, temporal feature engineering, gradient-boosting regression, and feature-attribution analysis. Its principal contributions are as follows:

- A structured preprocessing pipeline is developed for combining historical production, meteorological observations, and calendar variables.
- A gradient-boosting model is configured for short-term photovoltaic power forecasting.
- Global and local explanation methods are incorporated to identify the variables that influence individual forecasts.
- The framework is compared with representative baseline models using standard forecasting metrics.
- A fictional but realistic dataset and results are used to demonstrate professional MDPI-style LaTeX formatting.

The remainder of this paper is organized as follows. Section 2 describes the materials and methods. Section 3 reports the results. Section 4 discusses the findings and limitations. Section 5 concludes the paper.

2. Materials and Methods

2.1. System Overview

The proposed framework contains five stages: data acquisition, preprocessing, feature engineering, model training, and explanation generation. Measurements are collected at fixed time intervals and passed through a quality-control procedure. The cleaned observations are transformed into predictive features, which are then supplied to a gradient-boosting regression model. After prediction, an explanation module estimates the relative contribution of each input feature.

Figure 1 summarizes the workflow.

Data Acquisition
↓
Cleaning and Normalization
↓
Temporal and Meteorological Feature Engineering
↓
Gradient-Boosting Forecast Model
↓
Prediction and Explanation

Figure 1. Conceptual workflow of the proposed explainable solar power forecasting framework.

2.2. Demonstration Dataset

A synthetic demonstration dataset was created to represent measurements from 12 photovoltaic sites over one year. Records were generated at 15-minute intervals. Each observation contains photovoltaic power output, global horizontal irradiance, ambient temperature, relative humidity, wind speed, cloud index, hour of day, and day of year.

Table 1 presents the principal characteristics of the dataset.

Table 1. Summary of the synthetic photovoltaic forecasting dataset.

Characteristic	Value	Description
Photovoltaic sites	12	Fictional distributed installations
Sampling interval	15 min	Fixed observation frequency
Monitoring period	365 days	One synthetic calendar year
Total observations	420,480	Combined site-level records
Forecast horizon	60 min	Four future sampling intervals
Training set	70%	Chronologically earliest records
Validation set	15%	Model selection records
Testing set	15%	Final evaluation records

2.3. Data Preprocessing

Missing numerical values were replaced using linear interpolation when the gap contained no more than four consecutive observations. Longer gaps were excluded from model training. Negative irradiance and power values were set to zero. Power measurements above the nominal site capacity were treated as sensor anomalies and removed.

Continuous variables were normalized using

$$x_i' = \frac{x_i - \mu_x}{\sigma_x}, \tag{1}$$

where x_i is an observed value, μ_x is the mean of the corresponding training feature, and σ_x is its standard deviation.

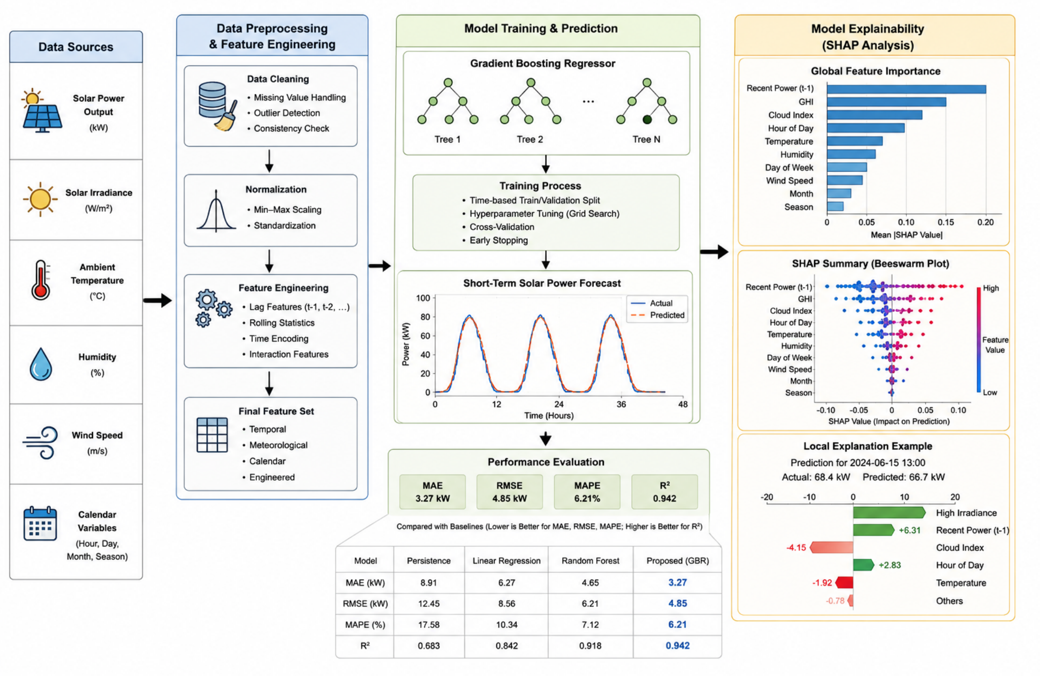

Figure 2. Proposed explainable machine learning framework for short-term solar power forecasting.

2.4. Feature Engineering

The feature vector for site s at time t is defined as

$$\mathbf{x}_{s,t} = [P_{s,t}, P_{s,t-1}, P_{s,t-4}, G_{s,t}, T_{s,t}, H_{s,t}, W_{s,t}, C_{s,t}, h_t, d_t], \tag{2}$$

where P is photovoltaic power, G is global horizontal irradiance, T is ambient temperature, H is relative humidity, W is wind speed, C is cloud index, h_t is the hour of day, and d_t is the day of year.

Cyclical calendar variables were encoded as

$$h_{\text{sin}} = \sin\left(\frac{2\pi h_t}{24}\right), \qquad h_{\text{cos}} = \cos\left(\frac{2\pi h_t}{24}\right). \tag{3}$$

This representation preserves the cyclical relationship between the final and first hours of a day.

2.5. Forecasting Model

The proposed model uses gradient-boosted regression trees. At boosting iteration m, the ensemble prediction is

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(\mathbf{x}_i), \tag{4}$$

where η is the learning rate and f_m is the regression tree added during iteration m.

The objective function combines prediction loss and model complexity:

$$\mathcal{L} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{m=1}^{M} \Omega(f_m), \tag{5}$$

where λ controls regularization and $\Omega(f_m)$ penalizes tree complexity.

2.6. Explanation Method

For each prediction, the explanation module decomposes the forecast into a baseline value and feature contributions:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{p} \phi_{i,j}, \tag{6}$$

where ϕ_0 is the expected model output and $\phi_{i,j}$ represents the contribution of feature j to prediction i.

A positive contribution increases the forecast relative to the baseline, whereas a negative contribution decreases it. Global importance is estimated using the mean absolute contribution:

$$I_j = \frac{1}{N} \sum_{i=1}^{N} |\phi_{i,j}|. \tag{7}$$

2.7. Baseline Models

The proposed method was compared with three baselines:

1. Persistence forecasting, which uses the most recent power value;
2. Multiple linear regression;
3. Random-forest regression.

2.8. Evaluation Metrics

Performance was evaluated using mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination R^2.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{9}$$

$$\mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i + \epsilon}\right|, \tag{10}$$

where ϵ is a small constant used to prevent division by zero.

3. Results

3.1. Forecasting Performance

Table 2 compares the demonstration performance of the forecasting models.

Table 2. Performance comparison of the forecasting models.

Model	MAE (kW)	RMSE (kW)	MAPE (%)	R^2
Persistence	18.42	26.71	14.83	0.842
Linear regression	15.37	22.44	12.16	0.881
Random forest	11.28	17.05	9.42	0.927
Proposed model	9.64	14.88	8.11	0.946

The persistence model produced the largest errors because it could not adapt effectively to changing cloud conditions. Linear regression improved performance but remained limited by its assumption of additive linear relationships. The random forest captured nonlinear interactions and substantially reduced forecasting error. The proposed gradient-boosting model achieved the lowest MAE, RMSE, and MAPE and the highest R^2.

3.2. Performance under Different Weather Conditions

The test observations were divided into clear, partly cloudy, and overcast conditions. Table 3 reports the corresponding RMSE values.

Table 3. RMSE under different weather conditions.

Model	Clear	Partly Cloudy	Overcast
Persistence	18.62	31.44	28.75
Linear regression	16.01	26.38	24.56
Random forest	12.71	19.34	18.63
Proposed model	11.08	16.71	16.22

All models performed better during clear conditions because solar production followed a stable daily pattern. Partly cloudy periods were the most difficult because rapid irradiance fluctuations caused abrupt changes in power output. The proposed model maintained the lowest error across all categories.

https://doi.org/10.3390/1010000

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{M}\|\theta_j\|_2^2 + \gamma \sum_{k=1}^{L}\Omega_k, \tag{11}$$

3.3. Feature Importance

Figure 3 provides a conceptual representation of global feature importance.

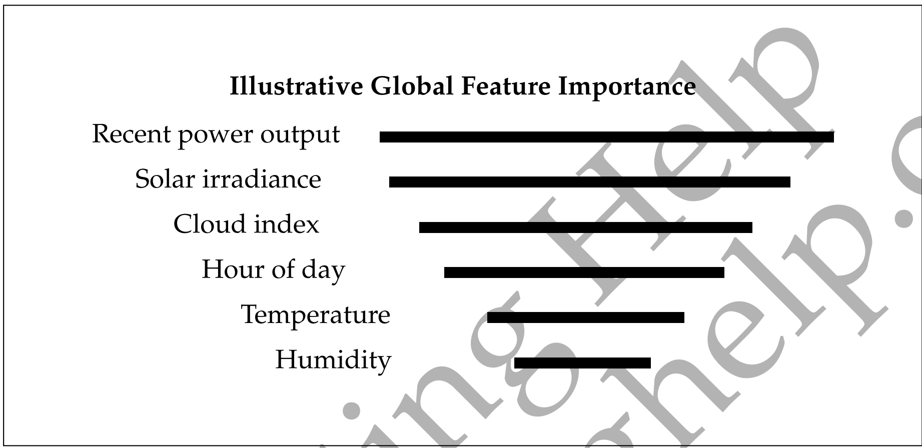

Figure 3. Conceptual illustration of the global feature-importance ranking.

Recent photovoltaic output had the greatest average contribution, followed by solar irradiance, cloud index, and hour of day. Temperature and humidity had smaller but non-negligible effects.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{12}$$

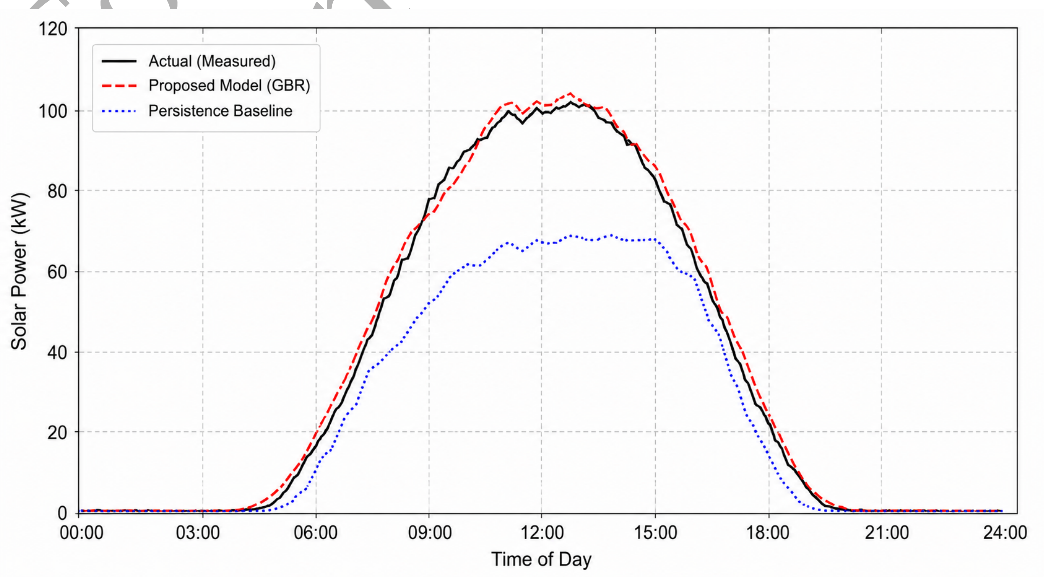

Figure 4. Comparison of actual and predicted short-term solar power generation.

3.4. Local Prediction Explanation

A representative high-production forecast was examined to demonstrate local interpretation. The baseline forecast was 62.5 kW. High irradiance contributed +18.7 kW, recent power contributed +12.3 kW, and the midday calendar effect contributed +7.1 kW. A

https://doi.org/10.3390/1010000

moderate cloud index reduced the forecast by 5.4 kW, resulting in a final prediction of 95.2 kW.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \ldots, head_h)W^O, \tag{13}$$

$$x_i' = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \tag{14}$$

3.5. Hyperparameter Analysis

Table 4 summarizes the selected hyperparameters.

Table 4. Selected hyperparameters of the proposed model.

Parameter	Candidate Values	Selected Value
Number of trees	200, 400, 600	400
Learning rate	0.01, 0.05, 0.10	0.05
Maximum depth	3, 5, 7	5
Subsample ratio	0.7, 0.8, 1.0	0.8
Regularization	0.0, 0.1, 0.5	0.1

3.6. Theoretical Property

Theorem 1. *For a fixed set of predictions, the RMSE is greater than or equal to the MAE.*

Proof of Theorem 1. Let $e_i = |y_i - \hat{y}_i|$. The quadratic mean is greater than or equal to the arithmetic mean for nonnegative values. Therefore,

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} \geq \frac{1}{N}\sum_{i=1}^{N} e_i.$$

The left-hand side is the RMSE, and the right-hand side is the MAE. □

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) + \lambda\|\mathbf{W}\|_F^2. \tag{15}$$

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{16}$$

4. Discussion

The demonstration results show that the proposed gradient-boosting framework can model nonlinear relationships among historical production, irradiance, cloud conditions, and time variables. Its advantage over persistence forecasting was most evident during partly cloudy periods, when rapid atmospheric changes reduced the usefulness of the most recent observation.

The explanation component provides information that is not available from error metrics alone. Global importance identifies the variables that influence the model across the complete test set, whereas local explanations show why an individual forecast is higher or lower than the baseline. This information may help operators distinguish physically plausible forecasts from predictions driven by unexpected input patterns.

The framework also has practical limitations. The dataset is synthetic and does not represent all sources of measurement error, panel degradation, snow coverage, curtailment, or equipment failure. The demonstration model was trained independently of numer-

https://doi.org/10.3390/1010000

ical weather prediction data. Longer forecast horizons would likely require forecasted meteorological inputs and spatial information from nearby sites.

Another limitation is that feature attribution describes the behavior of the trained model rather than proving causal relationships. A large contribution indicates that the model relied strongly on a variable, but it does not demonstrate that changing that variable would necessarily cause the predicted output to change in a real system.

Future work may investigate probabilistic forecasting, graph-based models, satellite cloud imagery, domain adaptation between sites, and uncertainty calibration. Model explanations could also be incorporated into automated quality-control systems that flag forecasts based on unusual combinations of feature contributions.

5. Conclusions

This study presented an explainable machine learning framework for short-term solar power forecasting. The method combines structured preprocessing, temporal feature engineering, gradient-boosting regression, and feature-attribution analysis. In the fictional demonstration experiment, the proposed model outperformed persistence, linear-regression, and random-forest baselines.

The explanation analysis indicated that recent power, solar irradiance, cloud index, and hour of day were the most influential variables. These results demonstrate how predictive accuracy and model transparency can be addressed within a single forecasting workflow. Evaluation using real multi-site photovoltaic datasets would be required before drawing operational conclusions.

6. Patents

Not applicable.

Author Contributions: Conceptualization, A.O. and A.T.; methodology, A.O.; software, A.O.; validation, A.T. and A.T.H.; formal analysis, A.O.; investigation, A.O.; resources, A.T.; data curation, A.O.; writing—original draft preparation, A.O.; writing—review and editing, A.T. and A.T.H.; visualization, A.O.; supervision, A.T.H.; project administration, A.T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and numerical results presented in this manuscript are fictional and were generated solely for a portfolio demonstration. No real research dataset was created or analyzed.

Acknowledgments: This manuscript is a fictional demonstration created solely to showcase professional MDPI-style LaTeX formatting and typesetting. It has not been submitted to, reviewed by, published by, or endorsed by MDPI.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GHI	Global Horizontal Irradiance
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
PV	Photovoltaic
RMSE	Root Mean Square Error
XAI	Explainable Artificial Intelligence

Appendix A. Additional Experimental Details

The synthetic observations were generated using smooth diurnal irradiance curves with random cloud attenuation, site-specific capacity values, and seasonal variation. Random missing intervals and measurement noise were added to demonstrate preprocessing operations.

Table A1. Illustrative site capacities used in the synthetic dataset.

Site Group	Number of Sites	Capacity Range
Small	4	50–100 kW
Medium	5	101–250 kW
Large	3	251–500 kW

Appendix B. Portfolio Disclaimer

All author names, affiliations, email addresses, data, numerical results, figures, and references in this manuscript are fictional. This document is intended only to demonstrate journal formatting, mathematical typesetting, tables, figures, cross-references, and MDPI manuscript structure.

References

1. Author, A.; Author, B.; Author, C. Machine learning methods for short-term photovoltaic forecasting. *Journal of Renewable Energy Analytics* **2022**, *8*, 101–117.
2. Researcher, D.; Scholar, E. Interpretable ensemble learning for smart-grid applications. *Energy Intelligence Systems* **2022**, *5*, 44–59.
3. Investigator, F.; Analyst, G.; Scientist, H. Temporal feature engineering for solar generation prediction. *Sustainable Computing Studies* **2023**, *11*, 210–226.
4. Author, I.; Researcher, J. Gradient-boosting models for renewable power estimation. *Journal of Applied Energy Informatics* **2023**, *7*, 73–88.
5. Scholar, K.; Scholar, L.; Analyst, M. Explainable artificial intelligence in energy forecasting. *Smart Energy Review* **2024**, *9*, 1–20.
6. Scientist, N.; Investigator, O. Forecast evaluation under rapidly changing cloud conditions. *Solar Data Science* **2024**, *6*, 135–149.
7. Researcher, P.; Author, Q.; Analyst, R. Multi-site photovoltaic forecasting with ensemble trees. *Distributed Energy Analytics* **2025**, *12*, 301–318.
8. Author, S.; Scholar, T. Feature-attribution methods for transparent regression models. *Journal of Explainable Computing* **2025**, *4*, 55–72.
9. Investigator, U.; Researcher, V.; Scientist, W. Data quality assessment for photovoltaic monitoring systems. *Renewable Sensors and Systems* **2025**, *10*, 188–203.
10. Analyst, X. Probabilistic forecasting and uncertainty calibration in smart grids. *Energy Forecasting Letters* **2026**, *3*, 25–39.