# PV Forecasting with Constrained Ensemble Learning and Context-Aware Stacking

Leïla Sakli and Seifeddine Ben El Ghali
*Laboratoire d'Informatique et Systemes UMR 7020, Aix-Marseille Université, CNRS*

***Abstract*—We propose a hybrid ensemble learning architecture for short-term photovoltaic (PV) power forecasting. The system integrates a physics-based model, a Bi-LSTM network, and an XGBoost regressor within a two-layer prediction pipeline. A constrained Ridge regression assigns robust weights to the base models, while a context-aware stacking layer dynamically refines predictions based on temporal and meteorological interaction features. Post-processing filters ensure physical consistency and enhance resilience to anomalies. Evaluated on real-world PV datasets, the proposed architecture consistently outperforms individual models and baseline ensembles, demonstrating superior accuracy, robustness, and readiness for deployment in intelligent energy management platforms.**

## 1. Introduction

The domain of Artificial Intelligence (AI) has emerged as a fundamental component in the advancement of contemporary energy infrastructures [1]. With the rapid deployment of renewable energy sources [2], especially photovoltaic (PV) systems, AI is now essential to enable intelligent grids, optimize demand-response strategies, and ensure fine-grained forecasting at scale [1]. In the context of smart grids and virtual power plants, accurate and reliable short-term PV forecasting plays a critical role in grid balancing, storage scheduling, and operational flexibility [3].

Despite substantial advances in both physical modeling and machine learning for PV forecasting, these approaches often suffer from structural limitations when applied independently [4] [5]. Physics-based models, such as those embedded in PVLib, provide interpretable predictions grounded in solar geometry and system parameters, but they are vulnerable to uncertain meteorological inputs [6]. In contrast, data-driven models such as LSTM and gradient-boosted trees (e.g. XGBoost) are highly flexible but tend to overfit noisy signals or violate physical constraints under edge-case conditions [7].

In this paper, we propose a robust ensemble learning framework that combines the strengths of physical knowledge and AI to enhance both the accuracy and robustness of PV production forecasting. Our approach dynamically weights heterogeneous predictors based on time, season, and weather conditions; integrates physics-aware constraints to ensure realistic outputs; and mitigates prediction failures caused by outliers or partial data unavailability. The proposed architecture follows a layered design, featuring a constrained Ridge regression layer that aggregates base predictions while enforcing regularization and preserving physical plausibility, providing a stable mechanism for reconciling diverse predictive signals even under degraded data conditions. Complementing this, a context-aware stacking layer based on XGBoost is employed, enriched with engineered interaction features that allow the model to learn contextual behaviors dynamically and implicitly, avoiding the limitations of manual data segmentation often used in time-series energy forecasting. Ultimately, this research pushes the boundaries of PV power forecasting by putting data quality front and center, paving the way for more reliable and accurate predictions in the field. This work contributes to the field of AI for energy systems by proposing a generalizable framework that improves predictive resilience and decision support in uncertain environments. The proposed method is evaluated on real-world PV datasets and benchmarked against individual predictors, naïve ensembles, and classic stacking models. The remainder of this paper is organized as follows: Section 2 presents related work on forecasting in energy systems. Section 3 describes the system architecture. Section 4 details the core algorithms. Section 5 outlines the experimental setup and discusses results, and Section 6 concludes.

## 2. Related Works

Real-world photovoltaic (PV) data often suffer from missing values due to sensor outages or communication failures [8], outliers resulting from hardware faults or atmospheric anomalies, and physically implausible readings such as nonzero generation during night or outputs exceeding system capacity [9]. These data issues underscore the need for forecasting systems that not only aim for accuracy under ideal inputs but also exhibit robustness to noise, adaptability to degraded or incomplete data, and consistency with physical constraints [10] [11]. Hybrid modeling approaches, which combine explicit physical representations with data-driven techniques, have demonstrated strong generalization performance across diverse weather conditions. For instance, models that integrate PVLib-based irradiance estimations with machine learning components can correct for local deviations while retaining physical interpretability [12]. In the context of solar forecasting, the fusion of physical models with Bi-LSTM and XGBoost architectures has shown empirical promise by effectively capturing temporal dependencies and non-linear interactions [12]. Ensemble learning, particularly stacking methods with learned weighting mechanisms, has been shown to enhance predictive accuracy and model stability [13]. Constrained stacking approaches, such as those employing Ridge regression to regulate model contributions, improve interpretability and reduce overfitting by preventing dominance of a single predictor. Similarly, multi-stage ensembles combining Ridge-based weighting and gradient boosting as a meta-learner have yielded reliable and explainable forecasting architectures [12].

To ensure operational safety and domain consistency, post-prediction filtering layers have been introduced in recent energy forecasting frameworks. These layers apply statistical bounds on model outputs, enforce upper limits based on system capacity, and zero-out predictions during hours without sunlight [14]. Such filters enable safe deployment of forecast models by correcting anomalies while keeping them lightweight and interpretable.

The design principles underlying recent photovoltaic forecasting approaches, such as hybrid pipelines, ensemble learning, nonlinear features, and domain-aware integration, reflect a broader shift toward models that are not only accurate but also resilient and sustainable. For instance, combining convolutional and recurrent neural architectures with environmental models enables a finer understanding of the climate-energy interaction at multiple scales [15]. The MIC-XLL model combines the maximum information coefficient (MIC) with LightGBM for feature selection and leverages a stacking ensemble of XGBoost [16], LightGBM, and LSTM to handle the complexity of PV power prediction [17]. Experimental results across two datasets confirm the performance advantages of this approach over traditional models, highlighting the potential of ensemble-based strategies in energy forecasting applications.

## 3. System Architecture

Figure 1 depicts the proposed system architecture, designed as a modular and robust prediction framework that synergistically combines domain expertise with machine learning techniques. It integrates three base predictors: (i) a physics-based model derived from PVLib, (ii) a bidirectional Long Short-Term Memory (Bi-LSTM) neural network, and (iii) an XGBoost regressor—each capturing distinct aspects of the photovoltaic generation process. These models are orchestrated through a two-level ensemble strategy.

At the first level, a Ridge regression with inequality constraints learns a weighted average of the base model predictions. This layer ensures stability and interpretability by restricting each model's contribution through convex constraints ( $w_m \geq -0.5$ for all $m$ and $\sum_m w_m > 0$, allowing small negative contributions while keeping the overall ensemble strictly positive). It guarantees that no model can dominate disproportionately or destabilize the ensemble due to erratic outputs.

At the second level, a meta-learner based on gradient boosting (XGBoost) performs conditional stacking.

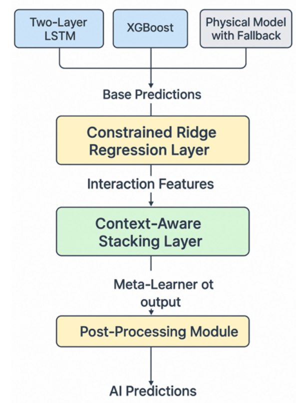


Figure 1. Prediction Framwork.

It receives the base predictions enriched with engineered context-aware features, including temporal encodings (hour, day-of-year), meteorological indicators (e.g., cloud cover, solar angle), and interaction terms (e.g., predicted power × hour_sin). This allows the meta-model to adjust prediction dynamics depending on the contextual environment.

To ensure physical plausibility and robustness, the system includes post-prediction filters: a statistical bounding mechanism that discards individual model outliers (values beyond $\mu + 3\sigma$), a cap on predicted output based on the nominal PV capacity ($P_{DC0}$), and a day/night logic that zeroes predictions when the sun is below the horizon.

This architecture aims to be generalizable, interpretable, and deployable across various PV sites with minimal retraining, while retaining robustness under uncertain weather conditions and input disturbances.

## 4. Core Algorithms

The proposed system relies on a set of learning and filtering mechanisms designed to enhance both accuracy and resilience. We propose a layered prediction architecture using an ensemble combining Bi-LSTM, XGBoost, and a physical model as base predictors. This section describes the key algorithmic components.

### 4.1. Base Predictors

The system integrates three base models, each capturing different aspects of PV behavior:

- **Physical Model:** Implemented using PVLib, it estimates the irradiance components (GHI, DNI, DHI) via the Erbs model and projects them onto the PV plane using Hay-Davies transposition. It incorporates a temperature correction factor (SAPM) and loss modeling (soiling, mismatch, cabling, etc.) to compute DC and AC outputs. A fallback model ensures continuity in case of missing inputs.
- **Bi-LSTM Network:** A two-layer architecture trained on 24-hour sequences of meteorological and

temporal features, including cyclic temporal encodings (e.g., `hour_sin`, `day_cos`), lagged PV power values, and derived meteorological indicators such as dew point and wind components. The model is optimized using a custom loss function with sample weighting to reduce over-penalization during low-irradiance periods.

- **XGBoost Regressor:** This model captures nonlinear feature interactions and is trained on a wide set of static and dynamic features: time encodings, smoothed cloud cover, logarithmic irradiance, daylight flag, and interaction features such as $temperature \times (1 - cloud\_cover)$. Delayed power outputs (lags up to 24h) are also injected to capture autocorrelation. It is particularly effective in exploiting tabular data and short-term meteorological patterns.

Each predictor is trained independently and contributes to the ensemble output via learned weights.

## 4.2. Ridge-Based Weighted Averaging

To combine base model outputs in a stable and interpretable way, a first-level ensemble is built using a Ridge regression with constraints:

$$\min_{w} \quad \|Pw - y\|^2 + \alpha\|w\|^2 \tag{1}$$

$$\text{s.t.} \quad w_m \geq -0.5 \quad \forall m, \quad \sum_{m} w_m > 0 \tag{2}$$

Here, $P$ contains the predictions from the base models, $\mathbf{y}$ is the target output, and $\alpha$ is a regularization parameter promoting smoother weight distributions and mitigating overfitting. The inequality constraints enforce numerical stability: they guarantee that the overall sum of weights remains strictly positive, while each individual weight is bounded below by $-0.5$. This lower bound was chosen empirically as a compromise, allowing for small negative contributions that help correct systematic biases of certain predictors without destabilizing the ensemble.

## 4.3. Context-Aware Meta-Learner (Stacking)

A second-level model refines the ensemble via supervised stacking using gradient boosting. The input feature vector is defined as:

$$X_{\text{stack}} = [p_{\text{physical}}, p_{\text{xgb}}, p_{\text{lstm}}, f_{\text{temporal}}, f_{\text{weather}}, f_{\text{interaction}}] \tag{3}$$

Where:

- $p_{\text{physical}}, p_{\text{xgb}}, p_{\text{lstm}}$ are the predictions from the base models.
- $f_{\text{temporal}}, f_{\text{weather}}, f_{\text{interaction}}$ are context-aware augmented features, including interactions with variables such as solar position, cloud cover, etc.

The $f$interaction features capture non-linear synergies between model predictions and contextual variables.

$$f^{(m)}_{\text{interaction}} = \left\{ p_m \cdot h_{\sin},\ p_m \cdot d_{\sin},\ p_m \cdot \frac{100 - CC}{100} \right\} \tag{4}$$

## 4.4. Post-Prediction Filtering

While ensemble learning, and particularly stacking, can enhance predictive accuracy by capturing complementary model behaviors, it does not inherently guarantee that predictions remain physically valid or robust to outliers. This is particularly critical in the context of photovoltaic forecasting, where predictions must respect real-world constraints such as system capacity, daylight availability, and the bounded nature of power generation.

Unlike most stacking implementations in the literature—which typically treat the output of the meta-learner as the final prediction—we introduce an essential *post-prediction filtering layer*. This choice is motivated by both domain-specific considerations and practical deployment concerns:

- **Physical Output Clipping:** The ensemble's final output $\hat{y}$ is bounded by a scaled multiple of the nominal DC system capacity $P_{DC0}$. This constraint ensures that the system does not produce physically implausible outputs that exceed the design limits of the PV plant. It acts as a hard upper limit, particularly important under uncertain weather conditions or model misalignment. A safety factor (e.g., 1.5) provides tolerance for transient overshoots caused by forecast errors or irradiance spikes.
- **Model-Level Outlier Control:** Each individual model prediction $p_m$ is examined in relation to its own historical distribution. Values exceeding $\mu_m + 3\sigma_m$ are truncated:

$$p_m = \min(p_m, \mu_m + 3\sigma_m) \tag{5}$$

This rule acts as a form of localized anomaly detection, protecting the ensemble from being skewed by instabilities in a single model, while preserving diversity among base predictors.

$$\hat{y}_{\text{final}} = \min(\hat{y},\ 1.5 \cdot P_{\text{DC0}}) \tag{6}$$

- **Daylight Enforcement:** Predictions are zeroed when the solar elevation is below the horizon, reflecting physical inactivity of the PV system during nighttime.

$$\hat{y}(t) = \begin{cases} 0 & \text{si } elevation_{\text{sun}}(t) \leq 0 \\ \hat{y}(t) & \text{sinon} \end{cases}$$

These rules are, non-parametric, interpretable, and computationally efficient.

# 5. Experiments and Results

## 5.1. Experimental Setting

To evaluate the effectiveness of the proposed layered ensemble system, we conducted experiments using real-world operational data collected from a photovoltaic system installed in Gökçeada, Turkey. The dataset comprises nearly one year of exported active energy data, accompanied by corresponding meteorological inputs and derived temporal features.

The forecasting task was performed at an hourly resolution with a 24-hour horizon. Each base predictor was trained independently as detailed in Section 4, and the ensemble was formed using constrained Ridge regression followed by a context-aware stacking layer. Finally, post-prediction filtering was applied to enforce physical plausibility. To avoid data leakage, all experiments follow a rolling-origin evaluation protocol: At each forecast origin $t$, models are trained on data up to $t$ and issue a 24-hour ahead forecast using only information available at $t$.

## 5.2. Evaluation Metrics

We use four complementary metrics to assess model performance. These indicators capture both numerical accuracy and temporal fidelity.

- **Root Mean Squared Error (RMSE)** penalizes large deviations between predicted and actual values. It is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{7}$$

- **Mean Absolute Error (MAE)** indicates the average magnitude of prediction errors, regardless of direction:

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

- **Mean Absolute Percentage Error (MAPE)** measures relative deviation, making it useful for evaluating proportional accuracy:

$$\mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{9}$$

- **Symmetric Mean Absolute Percentage Error (sMAPE)** provides a scale independent percentage error less sensitive to near-zero denominators:

$$\mathrm{sMAPE} = \frac{100}{N}\sum_{i=1}^{N}\frac{2\,|\hat{y}_i - y_i|}{|y_i| + |\hat{y}_i| + \varepsilon} \tag{10}$$

where $\varepsilon$ is a small constant to avoid division by zero.

- **Median Absolute Error (MedAE)** measures the typical absolute deviation and is robust to outliers:

$$\mathrm{MedAE} = \mathrm{median}\left(|\hat{y}_i - y_i|\right) \tag{11}$$

- **Pearson Correlation Coefficient** evaluates the temporal alignment and linear relationship between predicted and true values:

$$\rho = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{12}$$

These metrics were computed over the test period to capture both absolute accuracy and temporal alignment.

## 5.3. Quantitative Results

Table 1 summarizes the performance of the base models and the final ensemble. The physical model achieved high correlation ($\rho = 0.982$), but its RMSE and MAPE remained elevated due to sensitivity to unmeasured weather variability. The XGBoost regressor obtained moderate absolute accuracy but underperformed in temporal coherence, with a correlation of only 0.523. The Bi-LSTM achieved high correlation ($\rho = 0.962$) but suffered from relative errors during low irradiance periods leading to an inflated MAP.

The ensemble model outperformed all baselines, achieving a RMSE of **0.065**, MAE of **0.038**, and MAPE of **37.3%**, while maintaining a correlation of **0.979**. These results highlight the complementarity of the different predictors and the benefits of multi-level fusion.

TABLE 1. EVALUATION METRICS FOR EACH MODEL ON UEDAS-TR-B2 DATASET (DAYLIGHT HOURS).

| Model | RMSE | MAE | MAPE | sMAPE | MedAE | Corr. |
|---|---|---|---|---|---|---|
| Physical | 0.029 | 0.016 | 26.858% | 35.533% | 0.012 | 0.982 |
| XGBoost | 0.132 | 0.076 | 82.878% | 136.806% | 0.002 | 0.523 |
| Bi-LSTM | 0.308 | 0.176 | 482.0% | 104.7% | 0.124 | 0.962 |
| Ensemble | **0.065** | **0.038** | **37.342%** | **61.050%** | **0.004** | **0.979** |

## 5.4. Effect of Post-Prediction Filtering

To ensure robust deployment, the final prediction is passed through a lightweight filtering layer that:

- truncates model-specific outliers based on statistical bounds ($\mu + 3\sigma$),
- clips outputs exceeding 150% of the nominal capacity,
- zeroes predictions when solar elevation is negative.

This mechanism introduces negligible overhead while significantly increasing robustness. All predictions remained physically valid throughout the test period, whereas up to 5% of predictions from unfiltered models violated domain constraints.

### 5.5. Forecast Behavior Visualization

Figure 2 displays a forecast overlay over two representative days. The physical model provides a smooth curve, but tends to slightly overestimates midday production. XGBoost captures peaks levels but overshoots in some cases, while Bi-LSTM produces accurate ramp-up dynamics but remains less stable near peak hours. The ensemble balances these effects, producing a smooth and accurate trajectory aligned with the real data, across the full diurnal cycle.

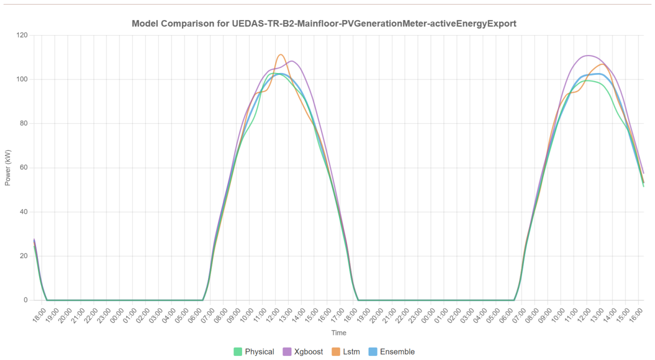


Figure 2. Forecast overlay comparison for UEDAS-TR-B2 over two days. The ensemble model (blue) corrects the over- and under-estimations of the base predictors.

### 5.6. Overall Performance Summary

Figure 3 illustrates the comparative performance across all metrics. The ensemble model achieves the best trade-off between accuracy, robustness, and physical plausibility, validating the proposed architecture for real-world PV forecasting applications.

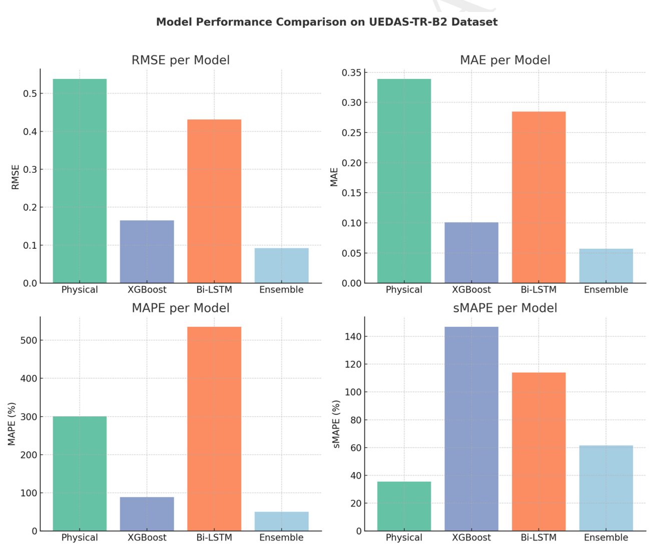


Figure 3. Comparison of RMSE, MAE, MAPE, and correlation across models.

## 6. Conclusion

In this work, we presented a robust photovoltaic forecasting system based on a multi-layer ensemble architecture combining a physical model, a Bi-LSTM network, and an XGBoost regressor. Each component was designed to capture distinct aspects of PV behavior: physical consistency, temporal dynamics, and short-term nonlinear interactions. The system leverages a constrained Ridge regression and a context-aware stacking layer to fuse the strengths of the base predictors. A post-prediction filtering module was added to enforce domain-specific constraints, ensuring physical plausibility and operational safety.

Extensive experiments on real operational data from the UEDAS-TR-B2 system demonstrated that the proposed ensemble model consistently outperforms individual predictors in both accuracy and robustness. Notably, it achieves substantial improvements in RMSE and MAE while maintaining high correlation with ground-truth observations. The filtering layer effectively mitigates the impact of prediction anomalies and guarantees compliance with system limits.

These results support the relevance of hybrid and ensemble strategies in solar forecasting, especially in the context of digital twins and grid integration where prediction reliability and interpretability are critical. Future work will explore online learning extensions, uncertainty quantification, and the integration of satellite-based forecasts for further generalization across sites and seasons.

## Acknowledgments


This work has been carried out within the framework of the European Union's Horizon Europe research and innovation programme under Grant Agreements No. 101096946 (FlexCHESS) and No. 101096836 (MASTERPIECE).


## References


[1] A. Mystakidis, P. Koukaras, N. Tsalikidis, D. Ioannidis, and C. Tjortjis, "Energy forecasting: A comprehensive review of techniques and technologies," *Energies*, vol. 17, no. 7, p. 1662, 2024, doi: 10.3390/en17071662.

[2] C. J. Ejiyi, D. Cai, D. Thomas, *et al.*, "Comprehensive review of artificial intelligence applications in renewable energy systems: current implementations and emerging trends," *J. Big Data*, vol. 12, p. 169, 2025, doi: 10.1186/s40537-025-01178-7.

[3] K. Barhmi, C. Heynen, S. Golroodbari, and W. van Sark, "A review of solar forecasting techniques and the role of artificial intelligence," *Solar*, vol. 4, no. 1, pp. 99–135, 2024, doi: 10.3390/solar4010005.

[4] W. Castillo-Rojas and J. Salinas, "Forecasting models applied in solar photovoltaic and wind energy: A systematic mapping study," *IEEE Access*, pp. 1–1, 2024.

[5] N. Rahimi, S. Park, W. Choi, B. Oh, S. Kim, Y.-h. Cho, S. Ahn, C. Chong, D. Kim, C. Jin, and D. Lee, "A comprehensive review on ensemble solar power forecasting algorithms," *J. Electr. Eng. Technol.*, published Jan. 12, 2023.

[6] W. F. Holmgren, C. W. Hansen, and M. A. Mikofski, "pvlib python: A python package for modeling solar energy systems," *J. Open Source Softw.*, vol. 3, no. 29, p. 884, 2018, doi: 10.21105/joss.00884.

[7] M. Gupta, A. Arya, U. Varshney, J. Mittal, and A. Tomar, "A review of PV power forecasting using machine learning techniques," *Prog. Eng. Sci.*, vol. 2, no. 1, p. 100058, 2025, doi: 10.1016/j.pes.2025.100058.

[8] S. Meisenbacher, B. Heidrich, T. Martin, R. Mikut, and V. Hagenmeyer, "AutoPV: Automated photovoltaic forecasts with limited information using an ensemble of pre-trained models," in *Proc. ACM Int. Conf.*, 2023, pp. 386–414, doi: 10.1145/3575813.3597348.

[9] L. Yang, Z. Huang, X. Mo, and T. Luo, "Enhanced GAIN-Based Missing Data Imputation for a Wind Energy Farm SCADA System," *Electronics*, vol. 14, no. 8, p. 1590, 2025, doi: 10.3390/electronics14081590.

[10] Y. Guo, Q. Han, T. Li, H. Fu, M. Liang, and S. Zhang, "Robust photovoltaic power forecasting model under complex meteorological conditions," *Mathematics*, vol. 13, p. 1783, 2025, doi: 10.3390/math13111783.

[11] W. Liu, C. Ren, and Y. Xu, "PV generation forecasting with missing data: A super-resolution perception approach," *IEEE Trans. Sustain. Energy*, 2020, doi: 10.1109/TSTE.2020.3029731.

[12] L. de Oliveira Santos, T. Alskaif, G. Barroso, and P. Carvalho, "Photovoltaic power estimation and forecast models integrating physics and machine learning: A review on hybrid techniques," *Solar Energy*, vol. 284, p. 113044, 2024, doi: 10.1016/j.solener.2024.113044.

[13] E. Roginek, J. Xu and D. F. Hsu, "InFusionLayer: A CFA-Based Ensemble Tool to Generate New Classifiers for Learning and Modeling," 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI), Herndon, VA, USA, 2024, pp. 756-763, doi: 10.1109/ICTAI62512.2024.00112.

[14] W. Khan, S. Walker, and W. Zeiler, "Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach," *Energy*, vol. 240, p. 122812, 2021, doi: 10.1016/j.energy.2021.122812.

[15] S. Shrestha, B. Buckley, A. S. Varde, and D. Cwynar, "Hybrid CNN-LSTM and domain modeling in climate-energy analysis for a smart environment," in *Proc. 2023 IEEE 35th Int. Conf. Tools Artif. Intell. (ICTAI)*, Atlanta, GA, USA, 2023, pp. 229–233, doi: 10.1109/ICTAI59109.2023.00041.

[16] X. Tang, H. Zhang, D. Li, D. Tang, C. Gong, and X. Yu, "A stacking ensemble model based on nonlinear feature selection for photovoltaic power prediction," in *Proc. 7th Asia Conf. Energy Electr. Eng. (ACEEE)*, Chengdu, China, 2024, pp. 345–349, doi: 10.1109/ACEEE62329.2024.10652171.

[17] H. Zhang and T. Zhu, "Stacking model for photovoltaic-power-generation prediction," *Sustainability*, vol. 14, no. 9, p. 5669, 2022, doi: 10.3390/su14095669.