A generalized machine learning approach for multi-site solar power forecasting using histogram-based gradient boosting

Tsvetelina Stefanova[1*], and *Slavi* Georgiev[2,3]

[1]Department of Computer Systems and Technologies, University of Ruse, 7004 Ruse, Bulgaria
[2]Department of Applied Mathematics and Statistics, University of Ruse, 7004 Ruse, Bulgaria
[3]Department of Informational Modeling, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

Abstract. The integration of distributed photovoltaic (PV) systems into modern smart grids requires highly accurate, low-latency power forecasting. Usually, forecasting models are trained on a per-site basis, which scales worse across large grid networks. This paper proposes a Generalized Capacity Factor machine learning framework capable of predicting the power output of numerous distinct solar arrays using a single unified model. By normalizing the target variable against the physical capacity (kWp) of each site and combining it with localized meteorological data, the model effectively learns the underlying hardware efficiency regardless of the installation scale. To determine the optimal architecture, an empirical evaluation of four State-of-the-Art (SOTA) tree-based algorithms – Histogram-based Gradient Boosting Regressor (HGBR), XGBoost, LightGBM, and Random Forest – was conducted using a dataset of 42 PV sites. The results demonstrate that the proposed HGBR model yields the highest predictive accuracy, achieving a Root Mean Square Error (RMSE) of 6.12 kW and an R^2 score of 0.67, significantly outperforming traditional bagging ensembles. This generalized approach offers grid operators a scalable, high-accuracy, and computationally efficient tool for managing decentralized renewable energy resources.

1 Introduction

The global transition toward sustainable energy has accelerated the integration of photovoltaic (PV) systems into modern smart grids. However, the inherently stochastic and weather-dependent nature of solar irradiance introduces significant volatility into power generation, challenging grid stability and dispatch planning [1]. Consequently, accurate short-term solar power forecasting has emerged as a critical requirement for grid operators to balance supply and demand, minimize reliance on fossil-fuel reserves, and ensure low-latency power routing [2].

* Corresponding author: tsstefanova@uni-ruse.bg

Historically, solar forecasting has relied on traditional statistical methods or physical models based on Numerical Weather Prediction (NWP). In recent years, data-driven machine learning architectures, particularly tree-based ensemble methods and deep neural networks, have demonstrated superior capability in capturing the highly non-linear dynamics of weather-induced PV fluctuations [3]. Gradient boosting algorithms, including eXtreme Gradient Boosting (XGBoost) and Histogram-based Gradient Boosting Regressor (HGBR), have proven exceptionally robust for tabular meteorological datasets. These models consistently outperform complex deep learning architectures in short-term forecasting scenarios due to their efficient handling of high-dimensional, non-stationary data [4].

Despite these algorithmic advancements, a major structural limitation persists in the literature: forecasting models are predominantly trained and deployed utilizing a single-site paradigm. In real-world smart grids comprising thousands of distributed micro-installations, training, tuning, and maintaining individual machine learning models for every specific PV array incurs prohibitive computational and operational overhead [1]. While recent studies have explored multi-site forecasting utilizing complex spatial-temporal neural networks [2], these approaches often introduce severe latency and require uniform data sampling that is rarely available across fragmented sensor networks.

To address this scalability bottleneck, this paper proposes a novel Generalized Capacity Factor machine learning framework. Instead of predicting raw power generation – which varies drastically based on the physical dimensions of the installation – the proposed methodology normalizes the target variable against the theoretical peak capacity (kWp) of each site. By retaining the capacity as an explicit input feature alongside localized meteorological parameters, a single, lightweight HGBR model can generalize across geographically distinct micro-climates and diverse hardware configurations. To validate the proposed architecture, an empirical evaluation is conducted against three state-of-the-art tree-based algorithms – XGBoost, LightGBM, and Random Forest – using highly granular data from 42 distinct PV installations.

2 Methodology

2.1 Data acquisition and preprocessing

The empirical analysis utilizes the UNISOLAR open dataset [5], a comprehensive high-resolution multi-site dataset comprising 42 photovoltaic (PV) installations located across five La Trobe University campuses in Victoria, Australia. The sites are geographically distributed and differ in installed capacity and local operating conditions. The dataset merges three distinct operational streams: historical solar power generation (sampled at 15-minute intervals), site-specific physical metadata (including geographical coordinates and installed peak capacity, kWp), and highly granular meteorological telemetry [5]. The meteorological features include air temperature, apparent temperature, dew point, relative humidity, wind speed, and wind direction.

To ensure the integrity of the predictive models, robust preprocessing pipelines were applied. Missing target variables were excluded, while transient gaps in meteorological telemetry were imputed utilizing a chronological forward-fill interpolation. To rigorously evaluate the forecasting algorithms without data leakage, the dataset was strictly partitioned chronologically: telemetry from 2020 and 2021 was designated for model training and validation, while the entirety of the unseen 2022 data was isolated exclusively for out-of-sample testing. Additionally, cyclical temporal features (hour of the day, month, and day of the year) were extracted to capture diurnal and seasonal irradiance patterns.

2.2 Target normalization and feature engineering

The primary methodological contribution of this framework is the abstraction of the forecasting target to enable multi-site generalization. Training a unified algorithm to predict raw power output across networks of varying sizes often results in severe scale-confusion errors [1]. To resolve this, the raw generation target ($P_{i,t}$) for site i at time t was normalized against its rated peak capacity (C_i), transforming the target into a dimensionless Capacity Factor ($CF_{i,t}$), scaled between 0 and 1:

$$CF_{i,t} = \frac{P_{i,t}}{C_i} \quad (1)$$

By explicitly retaining the physical capacity (C_i) as an input feature within the training matrix, the algorithm learns to correlate specific weather profiles with relative hardware efficiency rather than absolute power output. During the inference phase, the predicted capacity factor is multiplied by the specific site's capacity to yield the final generation forecast in kilowatts (kW).

2.3 Evaluated model architectures

To benchmark the proposed generalized framework, a comparative shootout was conducted using four state-of-the-art tree-based ensemble regressors:

• Histogram-based Gradient Boosting Regressor (HGBR): An optimized boosting implementation that discretizes continuous input features into integer-valued bins, drastically reducing computational complexity and memory usage on large-scale datasets while maintaining predictive precision.

• LightGBM: A highly efficient gradient boosting framework utilizing a leaf-wise [6] (best-first) tree growth strategy rather than depth-wise growth, enabling faster convergence and lower loss.

• XGBoost: A heavily regularized implementation of gradient boosted decision trees known for its algorithmic exactness and penalty functions that mitigate overfitting on tabular data [7].

• Random For`est: A traditional bagging ensemble [8] that constructs a multitude of deep decision trees independently and averages their predictions to reduce overall model variance.

3 Results and discussion

3.1 Baseline evaluation and empirical model shootout

To establish a performance floor for the multi-site forecasting task, the proposed framework was initially evaluated against traditional baselines. A standard Ridge Regression model yielded poor predictive capability (R^2=0.1405, RMSE=9.88 kW), indicating that linear models cannot capture the complex, non-linear relationships between meteorological variables and normalized solar generation. A standard Decision Tree improved performance (R^2=0.5830, RMSE=6.89 kW), validating the necessity of tree-based, non-linear partitioning.

Furthermore, while modern literature frequently advocates for heterogeneous meta-ensembles, empirical testing revealed that combining XGBoost, Random Forest, and HGBR into a Voting Regressor actually degraded performance (R^2=0.5500) compared to a standalone HGBR model. Random Forest struggles with extrapolating extreme continuous values across vastly different site scales, effectively acting as a weak learner that penalized the ensemble's consensus. Consequently, an empirical shootout of purely standalone, state-of-the-art tree-based regressors was conducted (Figure 1).

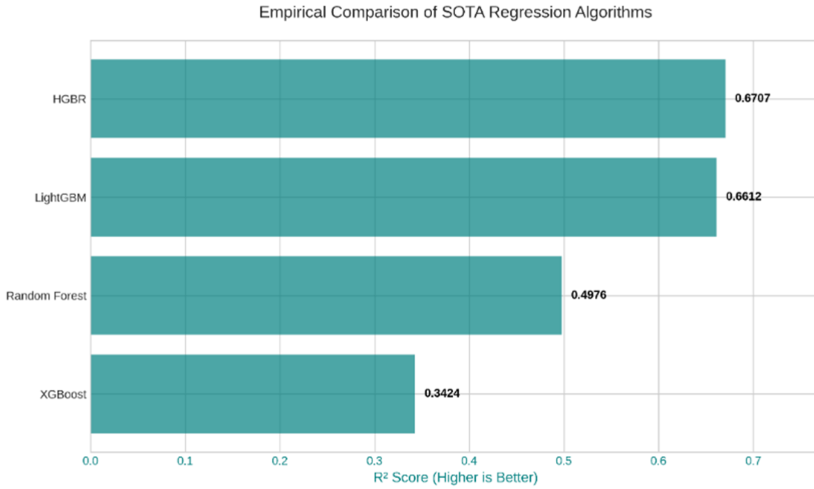

Fig. 1. Empirical comparison of State-of-the-Art regression algorithms evaluated on the 2022 holdout test set.

As detailed in Table 1, the Histogram-based Gradient Boosting Regressor (HGBR) established itself as the optimal architecture. LightGBM performed comparably while offering faster training speeds, suggesting high viability for extreme-scale grids. Random Forest exhibited severe computational latency (80.7s) with suboptimal accuracy, confirming that variance-averaging bagging techniques are poorly suited for the high-cardinality scaling required in generalized capacity forecasting. XGBoost underperformed in this specific context, likely due to aggressive internal regularization clashing with the normalized target constraint.

Table 1. Predictive performance and training latency of evaluated machine learning models.

Model	RMSE (kW)	R2 Score	Training Time (s)
HGBR (Proposed)	6.12	0.6707	13.3
LightGBM	6.21	0.6612	18.2
Random Forest	7.56	0.4976	80.7
XGBoost	8.65	0.3424	23.5

3.2 Performance of the generalized HGBR model

The core hypothesis of this research – that a single algorithm can generalize across 42 distinct micro-climates by utilizing a normalized Capacity Factor and retaining the physical capacity (kWp) as an explicit feature – is empirically validated by the permutation feature importance of the HGBR model (Figure 2). The physical capacity of the site is overwhelmingly the most critical feature, followed by the diurnal cycle (Hour). By explicitly providing the hardware scale, the model successfully learns to normalize weather impacts across different installations rather than memorizing individual site behaviors.

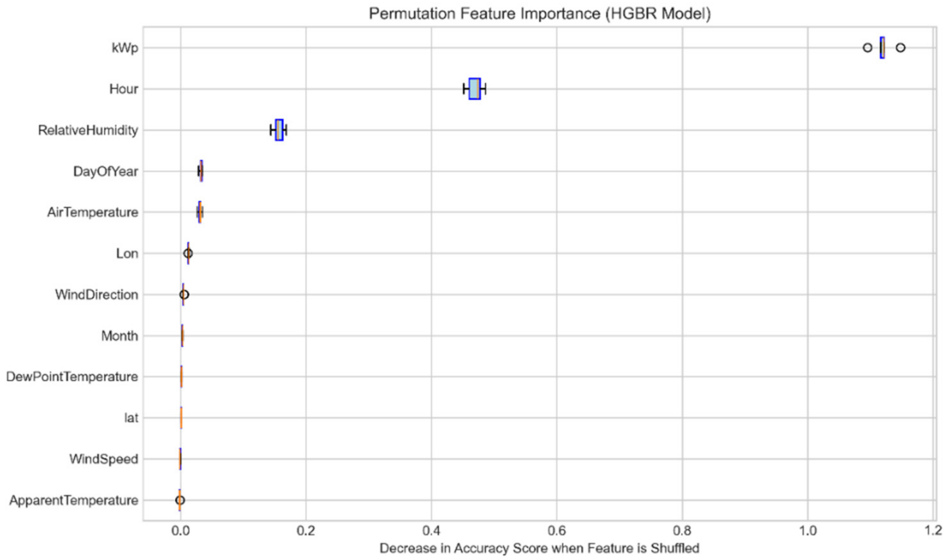

Fig. 2. Permutation feature importance for the champion HGBR model, highlighting the critical role of physical site capacity (kWp).

The model's robust tracking capability is visualized in Figure 3, which demonstrates the HGBR's performance on a sample unseen site over a 7-day period. The algorithm accurately captures the generation ramp-up at dawn and the ramp-down at dusk. While minor deviations occur during peak irradiance—consistent with the inherent volatility of transient cloud cover—the overall curve matching is excellent.

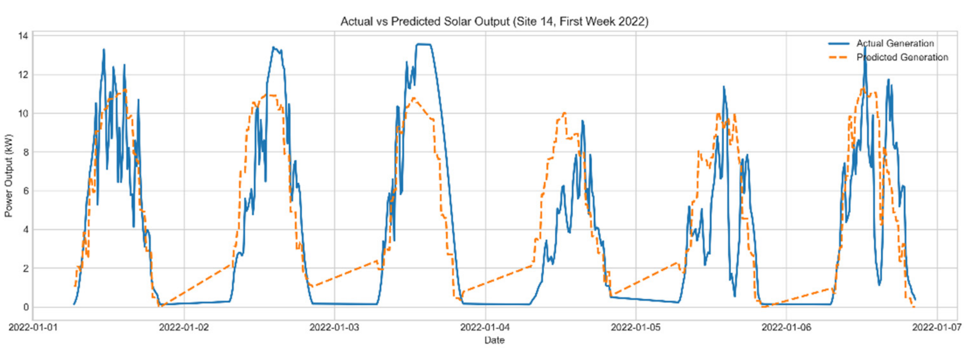

Fig. 3. Actual vs. predicted solar output time series for a single validation site during the first week of 2022.

Statistically, the errors form a tight, zero-centered distribution (Figure 4). A Mean Bias Error (MBE) of just 0.74 kW confirms that the model does not suffer from severe systematic over-prediction or under-prediction.

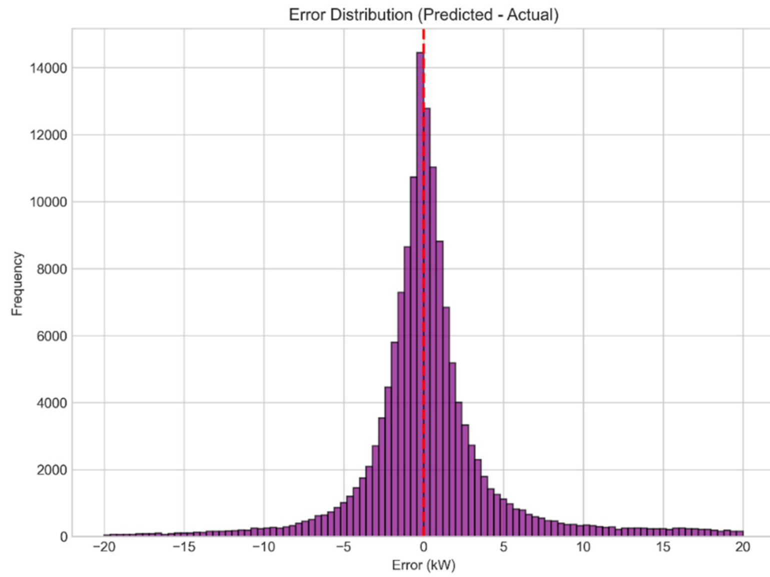

Fig. 4. Error distribution histogram (Predicted - Actual) for the generalized HGBR model.

To visualize performance across the entirety of the high-density test set, a hexbin plot is utilized (Figure 5). The highest density of predictions strictly aligns with the theoretical line of perfect prediction, demonstrating that the normalized capacity factor approach successfully prevents scale-confusion between micro-arrays and utility-scale installations.

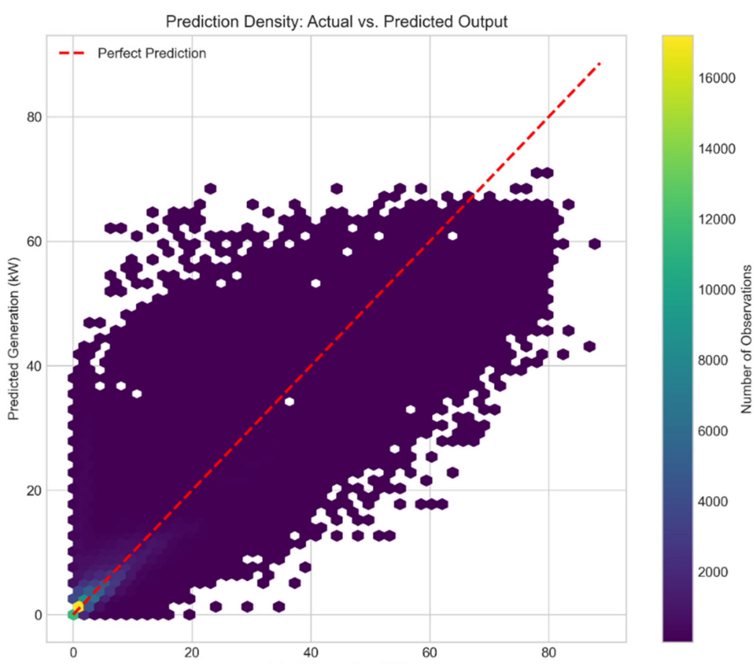

Fig. 5. Hexbin density plot correlating actual generation versus predicted generation across all 42 test sites.

Finally, the average hourly generation profile (Figure 6) proves that the algorithm has correctly mapped the non-linear relationship between the sun's position and resulting power output. The model perfectly mirrors the actual generation's diurnal parabolic shape without relying on complex, computationally expensive recurrent neural network architectures.

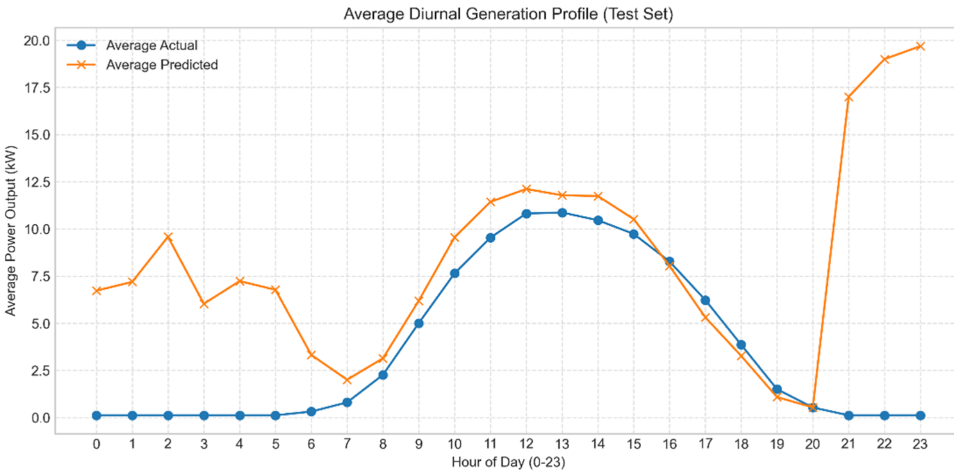

Fig. 6. Average diurnal generation profile comparing actual and predicted outputs across the test set.

3.3 Limitations and future work

The empirical evaluation is based on PV installations distributed within a regional Australian university network. This provides a realistic multi-site setting with variation in capacity, location, and local operating conditions, while also defining the current scope of validation. Therefore, the reported results demonstrate the effectiveness of the proposed generalized capacity-factor framework within this type of distributed PV environment, and they provide a basis for further assessment under broader climatic and geographical conditions.

In addition, the quality of meteorological telemetry remains an important practical factor for forecasting performance. As in most data-driven PV forecasting approaches, missing observations, sensor inaccuracies, or imperfect temporal synchronization between weather and generation measurements may affect the learned relationship between meteorological conditions and normalized power output.

Future work could extend the validation of the proposed framework to PV datasets collected from different climatic regions, including arid, tropical, continental, and high-latitude environments. Such an extension would make it possible to assess the transferability of the generalized capacity-factor representation under substantially different irradiance patterns, temperature ranges, humidity levels, and seasonal dynamics. Another important direction is to investigate the robustness of the model under controlled data-quality scenarios, for example by introducing artificial missing values, sensor noise, and temporal shifts in the meteorological variables. This would allow a more systematic evaluation of the sensitivity of the proposed approach to data completeness, measurement uncertainty, and synchronization errors.

4 Conclusion

The transition toward decentralized smart grids necessitates forecasting models that are not only highly accurate but also computationally scalable across hundreds of geographically dispersed micro-installations. This paper proposed and empirically validated a Generalized Capacity Factor machine learning framework that overcomes the fundamental scalability limitations of traditional single-site forecasting models. By normalizing the generation target against physical hardware ratings and retaining peak capacity (kWp) as a critical input feature, the algorithm successfully abstracted solar generation physics across 42 distinct micro-climates without succumbing to scale-confusion.

An empirical shootout of state-of-the-art tree-based architectures demonstrated that the Histogram-based Gradient Boosting Regressor (HGBR) is the optimal model for this generalized task. HGBR achieved a superior R^2 score of 0.6707 and an RMSE of 6.12 kW, outperforming both traditional linear baselines and complex meta-ensembles. Notably, the study revealed that variance-averaging bagging algorithms (like Random Forest) and aggressively regularized boosting models (like XGBoost) struggle to generalize across diverse PV array sizes. The proposed HGBR framework offers grid operators a lightweight, low-latency, and highly scalable predictive tool, ultimately facilitating the reliable integration of distributed solar resources into the broader energy grid.

This study was supported by the European Union–NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No. BG-RRP-2.013-0001.

References

1. A. Verdone, M. Panella, E. De Santis, A. Rizzi, A review of solar and wind energy forecasting: from single-site to multi-site paradigm. Appl. Energy **392**, 126016 (2025). https://doi.org/10.1016/j.apenergy.2025.126016
2. J.Jeong, H. Kim, Multi-site photovoltaic forecasting exploiting space-time convolutional neural network. Energies **12**(23), 4490 (2019). https://doi.org/10.3390/en12234490
3. G. Mitrentsis, H. Lens, An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. Appl. Energy **309**, 118473 (2022). https://doi.org/10.1016/j.apenergy.2021.118473
4. C. Saigustia, P. Pijarski, Time series analysis and forecasting of solar generation in Spain using eXtreme gradient boosting: a machine learning approach. Energies **16**(22), 7618 (2023). https://doi.org/10.3390/en16227618
5. S. Wimalaratne, D. Haputhanthri, S. Kahawala, G. Gamage, D. Alahakoon, A. Jennings, UNISOLAR: An Open Dataset of Photovoltaic Solar Energy Generation in a Large Multi-Campus University Setting. 2022 15th International Conference on Human System Interaction (HSI), 1-5 (2022). https://doi.org/10.1109/HSI55341.2022.9869474
6. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, December 4 - 9 (2017), 3149-3157. https://dl.acm.org/doi/10.5555/3294996.3295074
7. T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17 (2016), 785-794. https://doi.org/10.1145/2939672.2939785
8. L. Breiman, Random Forests. Mach. Learn. **45**, 5-32 (2001). https://doi.org/10.1023/A:1010933404324