Polo del Conocimiento

Pol. Con. (Edición núm. 105) Vol. 10, No 4
Abril 2025, pp. 2115-2127
ISSN: 2550 - 682X
DOI: https://doi.org/10.23857/pc.v10i4.9440

Pronóstico de energía solar a corto plazo usando modelos de aprendizaje automático: Un estudio comparativo

Short-term solar energy forecasting using machine learning models: A comparative study

Previsão de energia solar de curto prazo usando modelos de aprendizagem automática: um estudo comparativo

Alex Ricardo Guamán-Andrade [I]
alexr.guaman@espoch.edu.ec
https://orcid.org/0000-0001-8862-8350

Diego Alejandro García-Saraguro [II]
diego.garcia@espoch.edu.ec
https://orcid.org/0009-0008-7792-3969

Julio Francisco Guallo-Paca [III]
jguallo@espoch.edu.ec
https://orcid.org/0000-0002-8799-4735

Johanna Gabriela del Pozo-Naranjo [IV]
Johanna.delpozo@espoch.edu.ec
https://orcid.org/0009-0005-8680-8741

Correspondencia: alexr.guaman@espoch.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de Investigación

*** Recibido:** 18 de febrero de 2025 ***Aceptado:** 22 de marzo de 2025 *** Publicado:** 30 de abril de 2025

I. Escuela Superior Politécnica de Chimborazo, Ecuador.
II. Escuela Superior Politécnica de Chimborazo, Ecuador.
III. Escuela Superior Politécnica de Chimborazo, Ecuador.
IV. Escuela Superior Politécnica de Chimborazo, Ecuador.

Resumen

La creciente implementación de sistemas de energía renovable dentro de marcos energéticos descentralizados, incluyendo microrredes, instalaciones fuera de la red y entornos residenciales inteligentes, exige pronósticos precisos de energía solar a corto plazo para garantizar tanto la confiabilidad del sistema como la eficiencia operativa. Esta investigación presenta un análisis comparativo de cuatro algoritmos de aprendizaje automático para pronóstico: Bosque Aleatorio (Random Forest), Extreme Gradient Boosting, Regresión por Vectores de Soporte (Support Vector Regression) y Promedio Móvil Integrado Autorregresivo (ARIMA). Utilizando datos de generación solar recolectados en intervalos de 15 minutos durante un período de tres días, se evaluó la efectividad de los modelos en función del Error Cuadrático Medio (RMSE), el Coeficiente de Determinación (R^2) y la correspondencia visual con los patrones reales de generación. Los resultados de esta investigación revelan que las metodologías de aprendizaje automático basadas en ensamblado, específicamente RF y XGBoost, superan consistentemente tanto a los métodos estadísticos convencionales como ARIMA como a las técnicas basadas en kernel como SVR. Entre ellos, RF alcanzó el menor RMSE y el mayor R^2, lo que indica una precisión y capacidad de generalización excepcionales. Estos resultados resaltan la eficacia de los enfoques de ensamblado para capturar dinámicas no lineales y dependencias temporales inherentes a la generación de energía solar, lo que respalda su uso en aplicaciones de pronóstico en tiempo real relevantes para sistemas de energía renovable distribuidos.

Palabras clave: Pronóstico a corto plazo; Predicción de energía solar; Random Forest, XGBoost; Support Vector Regression; ARIMA.

Abstract

The increasing deployment of renewable energy systems within decentralized energy frameworks, including microgrids, off-grid installations, and smart residential environments, demands accurate short-term solar power forecasts to ensure both system reliability and operational efficiency. This research presents a comparative analysis of four machine learning algorithms for forecasting: Random Forest, Extreme Gradient Boosting, Support Vector Regression, and Autoregressive Integrated Moving Average (ARIMA). Using solar generation data collected at 15-minute intervals over a three-day period, the effectiveness of the models was evaluated based on the Root Mean

Square Error (RMSE), Coefficient of Determination (R^2), and visual correspondence with actual generation patterns. The results of this research reveal that ensemble-based machine learning methodologies, specifically RF and XGBoost, consistently outperform both conventional statistical methods such as ARIMA and kernel-based techniques such as SVR. Among them, RF achieved the lowest RMSE and the highest R^2, indicating exceptional accuracy and generalization capabilities. These results highlight the effectiveness of ensemble approaches in capturing nonlinear dynamics and temporal dependencies inherent in solar power generation, supporting their use in real-time forecasting applications relevant to distributed renewable energy systems.

Keywords: Short-range forecasting; Solar power prediction; Random Forest, XGBoost; Support Vector Regression; ARIMA.

Resumo

A crescente implantação de sistemas de energia renovável em estruturas de energia descentralizadas, incluindo micro-redes, instalações fora da rede e ambientes residenciais inteligentes, exige previsões precisas de energia solar a curto prazo para garantir a fiabilidade do sistema e a eficiência operacional. Esta investigação apresenta uma análise comparativa de quatro algoritmos de aprendizagem automática para previsão: Random Forest, Extreme Gradient Boosting, Support Vector Regression e Autoregressive Integrated Moving Average (ARIMA). Utilizando dados de geração solar recolhidos em intervalos de 15 minutos durante um período de três dias, a eficácia dos modelos foi avaliada com base na Raiz Quadrática Média do Erro (RMSE), no Coeficiente de Determinação (R^2) e na correspondência visual com os padrões de geração reais. Os resultados desta investigação revelam que as metodologias de aprendizagem automática baseadas em conjuntos, especificamente RF e XGBoost, superam consistentemente os métodos estatísticos convencionais, como o ARIMA, e as técnicas baseadas em kernel, como o SVR. Entre eles, o RF alcançou o RMSE mais baixo e o R^2 mais elevado, indicando uma precisão e capacidade de generalização excecionais. Estes resultados destacam a eficácia das abordagens de conjunto na captura de dinâmicas não lineares e dependências temporais inerentes à geração de energia solar, apoiando a sua utilização em aplicações de previsão em tempo real relevantes para sistemas de energia renovável distribuída.

Palavras-chave: Previsão de curto prazo; Previsão de energia solar; Floresta aleatória, XGBoost; Regressão de vetores de suporte; ARIMA.

Introduction

The escalating incorporation of renewable energy sources, particularly solar photovoltaic systems, has profoundly transformed contemporary power grids, notably within decentralized frameworks such as microgrids, off-grid configurations, and intelligent residential setups. In these scenarios, solar energy is pivotal in realizing energy independence and sustainability (Kocakusak et al., 2023). Nonetheless, the sporadic nature and fluctuations of solar irradiance present a significant obstacle to the dependability and efficacy of the power infrastructure. In this regard, precise forecasting of solar power generation is crucial for optimizing system performance and enhancing grid management, thereby facilitating the incorporation of renewable resources into the current energy frameworks. (Shi & Eftekharnejad, 2016).

Short-term forecasting—encompassing temporal spans ranging from several minutes to numerous hours—has ascended as a pivotal instrument for alleviating the operational uncertainties linked with solar energy generation. Accurate short-term forecasts enable optimal scheduling of energy storage systems, refined load management, and improved decision-making for energy trading and grid stability. (Lin et al., 2020) In the context of microgrid and off-grid applications, meticulous forecasting has the potential to diminish reliance on fossil fuel-dependent auxiliary generators, decrease fuel utilization, and enhance the resilience and sustainability of energy supply systems. (Guimarães et al., 2020).

Machine learning (ML) methodologies have attained significant recognition in the domain of solar forecasting owing to their capacity to elucidate intricate, nonlinear associations within meteorological and energy-related datasets. Among these methodologies, ensemble approaches including Random Forest (RF) and Extreme Gradient Boosting (XGBoost) have demonstrated significant predictive effectiveness. (Solano & Affonso, 2023) RF is esteemed for its robustness, simplicity of application, and immunity to overfitting, whereas XGBoost provides augmented precision through the mechanisms of gradient boosting, regularization, and meticulous oversight of learning processes. In addition to these ensemble algorithms, Support Vector Regression (SVR) offers a kernel-based framework adept at effectively modeling nonlinear trends within smaller datasets, although it frequently demonstrates diminished responsiveness to sudden variances. Conventional statistical methodologies, (Liu et al., 2017) including Autoregressive Integrated Moving Average (ARIMA), remain pivotal for establishing baseline comparisons, particularly within the framework of univariate time series forecasting. Collectively, these models embody a

continuum of forecasting strategies that differ in intricacy, adaptability, and computational efficiency, rendering them significant benchmarks for assessing short-term solar energy prediction methodologies within real-time operational frameworks. (Chodakowska et al., 2023)

In the current investigation, a comparative study of four predictive models RF, XGBoost, SVR and ARIMA for short-term solar power prediction in power energy systems are presented. For this purpose, solar generation data sampled at 15-minute intervals over a three-day period. Each model is evaluated based on predictive accuracy, computational robustness, and its ability to handle temporal features and nonlinear dynamics inherent in solar power output. Finally in the results, visual and quantitative metrics, as Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2) describes the optimal modeling strategy for implementation in real world applications, such as microgrids, nanogrids, and intelligent residential systems. The insights derived from this analysis aim to inform the selection of forecasting algorithms that enhance reliability, operational efficiency, and energy autonomy in renewable-powered distributed systems.

Methodology

This research utilized a dataset comprising short-term solar energy measurements captured at 15-minute intervals throughout a duration of three days. Each record contained a timestamp and the corresponding power output value. The high temporal resolution of these measurements is particularly suitable for modeling short-term solar generation dynamics in decentralized energy systems such as microgrids, off-grid installations, and smart homes (Ziyabari et al., 2020). The aim of this investigation was to assess and contrast the predictive efficacy of machine learning algorithms within authentic operational contexts, integrating temporal dependencies and environmental fluctuations.

Prior to model training, a sequence of data preprocessing and feature engineering steps was performed. Timestamps were converted into datetime objects to facilitate temporal decomposition. From these, time-based features are extracted, including the hour and minute of the day, and the day of the week, to account for diurnal patterns and weekday–weekend variations. Furthermore, lag features were constructed using the power output values recorded at 15, 30, and 45 minutes before each observation. These lagged variables were designed to capture short-term temporal dependencies and autocorrelation inherent in solar generation profiles, thereby enhancing model input richness.

Rows with missing values introduced during lag feature creation were removed to ensure data integrity. The cleaned dataset was then split into training and testing subsets using an 80:20 chronological split, without shuffling, to preserve the temporal order essential for time series forecasting. (Wu et al., 2015) This methodology ensures that model evaluation simulates real-world scenarios, where predictions are made using only past data. Such sequential splitting enhances the realism and reliability of the assessment by preventing data leakage from future observations. (Bourchtein & Bourchtein, 2008)

The primary analytical model employed in this research was the Random Forest (RF) Regressor, a sophisticated ensemble learning algorithm that operates under the principles of the bagging methodology. Random Forest methodology generates a comprehensive array of decision trees throughout the training process, with each tree constructed utilizing a random selection of both the training dataset and the feature set (Basavarajaiah & Narasimha Murthy, 2020) This approach reduces overfitting and improves generalization by minimizing variance without significantly increasing bias. Its ability to handle nonlinear relationships and interactively model complex feature dependencies make it particularly effective for regression problems involving environmental and time-based data such as solar power output. (Khan & Srivastava, 2020)

The subsequent model utilized was the Extreme Gradient Boosting Regressor (XGBoost), which represents a robust and scalable execution of the gradient boosting algorithm. Unlike bagging methods, boosting builds trees sequentially, where each tree attempts to correct the errors of its predecessor by minimizing a differentiable loss function. (Bentéjac et al., 2019) XGBoost enhances this procedure with techniques such as regularization (to avoid overfitting), shrinkage (learning rate control), and advanced tree pruning. Moreover, it supports parallelized computations and handles missing values intrinsically. These characteristics make XGBoost exceptionally efficient and accurate in predictive modeling, particularly in structured datasets where temporal and lagged relationships play a significant role. (Huang et al., 2019)

The third analytical framework examined was the Autoregressive Integrated Moving Average, a statistical approach for the forecast of time series data. ARIMA amalgamates autoregressive and moving average elements with differencing techniques to effectively encapsulate non-stationary patterns within temporal datasets. It is particularly suited for univariate forecasting problems where historical values of a series provide significant predictive power. (Tiao, 2001) The model parameters (p, d, q) were selected based on autocorrelation and partial autocorrelation function

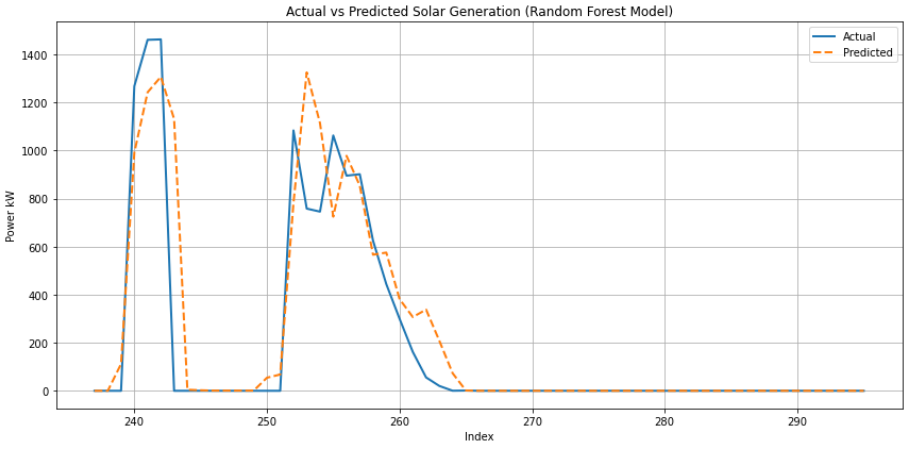
(ACF and PACF) plots. Although ARIMA lacks the flexibility of machine learning models in handling nonlinearities or exogenous inputs, it serves as a robust baseline for evaluating forecasting accuracy in stationary or near-stationary solar generation series. (Alsharif et al., 2019)

The fourth model employed was Support Vector Regression that is a kernel-based learning method derived from Support Vector Machines. SVR aims to find a function that approximates the output within a predefined margin of tolerance while minimizing model complexity. (WU, 2006) This model is particularly effective for small to medium sized datasets and can model nonlinear relationships using kernel functions such as radial basis function. SVR's capacity to generalize well from limited data makes it a suitable candidate for short-term solar forecasting when feature engineering is limited or when data availability is constrained. (Bisoi et al., 2022)

The efficiency of the model was evaluated through the application of two widely recognized regression metrics: Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). RMSE measures the average prediction error magnitude, while R^2 indicates the proportion of changing in the output variable that is captured by the model. (Willmott & Matsuura, 2005) In addition to numerical evaluation, forecasted and actual power output values were plotted over time to visually assess each model's tracking capability with respect to real generation trends (Narmadha et al., 2017). The experiment was also conducted to examine how variations in input parameters affected model accuracy and robustness. These comprehensive evaluations provided critical insights into each model's forecasting reliability, interpretability, and potential for deployment in solar energy applications. (Liebermann et al., 2021)

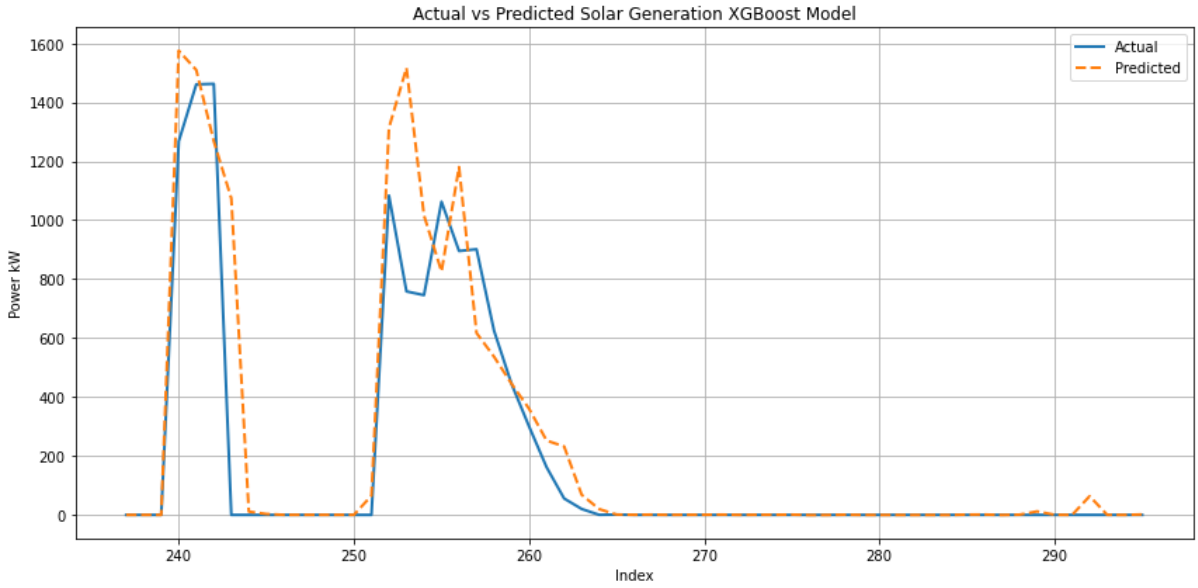
Analysis and results

Figure 1. *Solar Energy Forecasting using Random Fores Model*



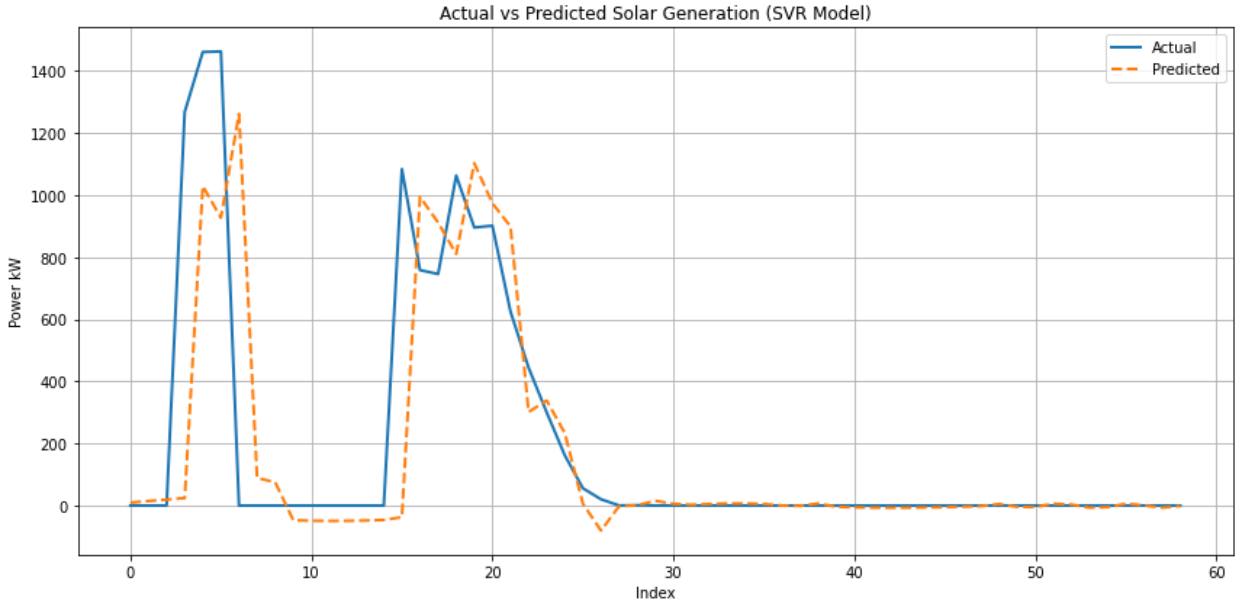
The performance of the Random Forest model in predicting short-term solar power generation is presented on figure 1. The predicted values align closely with the actual measurements, especially during peak generation hours. The model effectively tracks the rapid increase and decrease in solar output, with minor underestimations during maximum output periods. The smoothness and stability of the predictions suggest a strong ability of RF to handle the temporal dynamics and nonlinearity inherent in solar data, particularly in residential and microgrid contexts.

Figure 2. *Solar Energy Forecasting using XGBoost Model*



The results of the Extreme Gradient Boosting model are presented on figure 2. Like RF, XGBoost captures the general trend and fluctuations of the actual data. However, its predictions tend to slightly overestimate solar output during peak hours, leading to more pronounced deviations in some intervals. This overfitting behavior is a known characteristic of boosting models, especially when not finely tuned. Despite this, XGBoost maintains high reactivity and responsiveness, which makes it competitive for real-time forecasting in dynamic solar environments.

Figure 3. *Solar Energy Forecasting using SVR Model*



In figure 3, the forecast obtained from the Support Vector Regression model is presented. Compared to the ensemble methods, SVR exhibits greater deviation from the actual values, particularly at the beginning and end of the generation window. The model's predictions show delays in the onset and decline of solar generation, with visible under- and overestimations throughout the day. These limitations suggest that SVR struggles to adapt to complex nonlinearities and fast transitions typical of solar irradiance data sampled at high frequency.

Figure 4. *Solar Energy Forecasting using ARIMA*

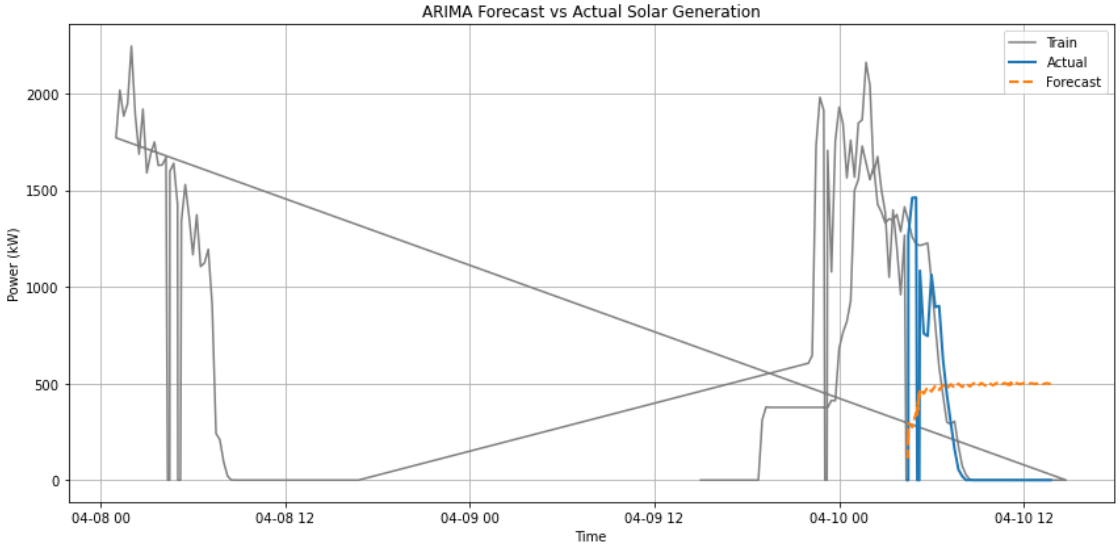


Figure 4 presents the forecast of the ARIMA model, which significantly diverges from the actual solar generation pattern. The model fails to replicate the cyclical and nonlinear nature of the time series, resulting in a nearly flat and overly smoothed forecast. The lack of responsiveness to sudden changes and the inability to follow peak dynamics demonstrate that ARIMA is not suitable for this type of application, especially when used without exogenous variables or advanced seasonal adjustments.

When comparing the visual outputs across all four models, clear distinctions emerge in terms of accuracy, responsiveness, and generalization. Both Random Forest and XGBoost closely follow the shape and peaks of the actual solar power output, with Random Forest offering slightly smoother transitions and better alignment during low- and no-generation periods. XGBoost, while equally reactive, tends to overestimate peak values, which may reflect a more aggressive fitting behavior. On the other hand, SVR exhibits noticeable misalignments, particularly in capturing the timing and magnitude of peaks, indicating its limited ability to handle sudden variations in solar irradiance. The ARIMA model demonstrates the weakest visual performance, as its forecast remains flat and disconnected from the actual generation profile, due to its linear nature and lack of adaptation to non-stationary trends.

Table 1. *Error comparison between machine learning models*

	Random Forest	**XGBoost**	**SVR**	**ARIMA**
RMSE	196.98	196.14	297.05	468.4
R^2	0.78	0.76	0.45	-0.57

The summary in Table 1 reinforces these qualitative observations with quantitative evidence. Random Forest emerges as the best-performing model, combining a low RMSE (196.98 kW) with the highest R^2 (0.78), thus ensuring both numerical accuracy and strong variance explanation. XGBoost exhibits nearly identical RMSE (196.14 kW) but slightly lower R^2 (0.76), implying that while its predictions are close in value to the actual ones, they deviate more in terms of variance structure. SVR's high RMSE (297.05 kW) and modest R^2 (0.45) suggest weaker predictive capability and limited utility for capturing the dynamics of short-term generation. ARIMA's RMSE of 468.4 kW and negative R^2 (-0.57) confirm its inadequacy, as it fails to outperform even a naive baseline. This comprehensive comparison highlights the superior adaptability of ensemble machine

Pol. Con. (Edición núm. 105) Vol. 10, No 4, Abril 2025, pp. 2115-2127, ISSN: 2550 - 682X

learning models—particularly RF and XGBoost—when dealing with high-resolution, nonlinear, and intermittent solar power data.

Conclusions

This research conducted an extensive assessment of four predictive models Random Forest, Extreme Gradient Boosting, Support Vector Regression, and ARIMA focused on short-term solar power forecasting employing high-resolution datasets derived from a decentralized energy framework. The findings unequivocally indicate the predominance of ensemble machine learning methodologies, particularly Random Forest and Extreme Gradient Boosting, in effectively capturing the nonlinear and dynamic characteristics inherent in solar generation data.

Among all models, the Random Forest algorithm exhibited the most optimal equilibrium between predictive precision and resilience, as indicated by its minimal Root Mean Square Error and maximal R^2 coefficient. XGBoost delivered comparable performance but exhibited slight overfitting during peak output periods. In contrast, SVR struggled to adapt to rapid fluctuations, resulting in delayed and less precise predictions. ARIMA, a traditional statistical model, proved unsuitable for this application, failing to capture both temporal variability and seasonal patterns.

The findings underscore the importance of incorporating temporal features, lag structures, and nonlinear modeling capabilities when designing forecasting systems for renewable energy applications. Ensemble learning methodologies, attributed to their inherent scalability and robustness against overfitting, present a pragmatic and efficacious approach for real-time solar energy prediction within microgrids, off-grid infrastructures, and intelligent residential systems. Future work may explore hybrid approaches that integrate weather forecasts or exogenous variables to further enhance prediction accuracy and operational reliability.

References

1. Kocakusak, D., Senick, J., & Andrews, C. J. (2023). Implementing the energy transition: lessons from New Jersey's residential solar industry.Climate Policy. https://doi.org/10.1080/14693062.2023.2202208
2. Shi, G., & Eftekharnejad, S. (2016, September 1). Impact of solar forecasting on power system planning.North American Power Symposium. https://doi.org/10.1109/NAPS.2016.7747909

3. Lin, Y., Duan, D., Hong, X., Cheng, X., Yang, L., & Cui, S. (2020, May 29).Very-Short-Term Solar Forecasting with Long Short-Term Memory (LSTM) Network. https://doi.org/10.1109/AEEES48850.2020.9121512
4. Guimarães, T., Costa, L. M., Leite, H., & Azevedo, L. F. (2020, September 1).A Hybrid Approach to Load Forecast at a Micro Grid level through Machine Learning algorithms. https://doi.org/10.1109/SEST48500.2020.9203308
5. Solano, E. S., & Affonso, C. M. (2023). Solar Irradiation Forecasting Using Ensemble Voting Based on Machine Learning Algorithms.Sustainability. https://doi.org/10.3390/su15107943
6. Liu, F., Qiao, R., Shen, C., & Luo, L. (2017).Designing ensemble learning algorithms using kernel methods. https://doi.org/10.1504/IJMISSP.2017.10009116
7. Chodakowska, E., Nazarko, J., Nazarko, Ł., Rabayah, H., Abendeh, R., & Alawneh, R. (2023). ARIMA Models in Solar Radiation Forecasting in Different Geographic Locations.Energies. https://doi.org/10.3390/en16135029
8. Ziyabari, S., Du, L., & Biswas, S. (2020, June 14). A Spatio-temporal Hybrid Deep Learning Architecture for Short-term Solar Irradiance Forecasting.Photovoltaic Specialists Conference. https://doi.org/10.1109/PVSC45281.2020.9300789
9. Wu, S.-F., Chang, C.-Y., & Lee, S.-J. (2015, March 2).Time series forecasting with missing values. https://doi.org/10.4108/ICST.INISCOM.2015.258269
10. Bourchtein, A., & Bourchtein, L. (2008).A Splitting Scheme for Large-Scale Atmosphere Dynamics Models. https://doi.org/10.1007/978-3-540-69848-7_51
11. Basavarajaiah, D. M., & Narasimha Murthy, B. (2020).Random Forest and Concept of Decision Tree Model. https://doi.org/10.1007/978-981-15-8210-3_3
12. Khan, M., & Srivastava, K. (2020, January 17). Regression Model for Better Generalization and Regression Analysis.International Conference on Machine Learning. https://doi.org/10.1145/3380688.3380691
13. Bentéjac, C., Csörgo, A., & Martínez-Muñoz, G. (2019). A Comparative Analysis of XGBoost.arXiv: Learning. https://doi.org/10.1007/S10462-020-09896-5
14. Huang, H., Zheng, Z., Xiao, J., Xu, G., & Wang, J. (2019).XGBoost disease probability predicting method, system and storage medium.

15. Tiao, G. C. (2001).Time Series: ARIMA Methods. https://doi.org/10.1016/B978-0-08-097086-8.42182-3
16. Alsharif, M. H., Younes, M. K., & Kim, J. (2019). Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea.Symmetry. https://doi.org/10.3390/SYM11020240
17. WU, T. (2006). SVM Based Algorithm for Regressive Modeling with Accurate Data Input-interval Number Output.Control and Decision.
18. Bisoi, R., Dash, D. R., Dash, P. K., & Tripathy, L. N. (2022). An efficient robust optimized functional link broad learning system for solar irradiance prediction.Applied Energy. https://doi.org/10.1016/j.apenergy.2022.119277
19. Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.Climate Research. https://doi.org/10.3354/CR030079
20. Narmadha, J., Kaavya, G. N., & Preethii, S. D. (2017, April 1). Analysis on electricity generation forecasting system.International Conference on Electric Information and Control Engineering. https://doi.org/10.1109/ICEICE.2017.8191901
21. Liebermann, S., Um, J.-S., Hwang, Y., & Schlüter, S. (2021). Performance Evaluation of Neural Network-Based Short-Term Solar Irradiation Forecasts.Energies. https://doi.org/10.3390/EN14113030