Solar Energy Forecasting in Seattle Using Machine Learning Models

Tianshuo Wang

University of Washington, Seattle, United States

twang38@uw.edu

Abstract. Seattle faces challenging environmental issues, which encourage people to explore ways of improving the utility of renewable resources, such as Global Horizontal Irradiance (GHI). This research consists of methods of data cleaning, data visualization, and designs of Machine Learning models, like Linear Regression, Decision Tree, and Random Forest. Based on the results of R2 and RMSE values, the Random Forest has the best performance among other models, with an RMSE≈of 48.1 and an R^2≈of 0.82. The research is dedicated to providing a stable and reliable prediction of GHI Values. With GHI predictions, city planners and government are able to efficiently plan and manage solar resources, mitigate instability risks, and enhance building energy performance through improved control of lighting and shading systems. Moreover, the application of machine learning models is a good start in environmental sciences and gives a solid foundation in the application of data sciences in real life. Additionally, the findings offer a reference framework for integrating predictive solar energy data into urban planning and renewable energy policy-making. Future work could extend these models to incorporate real-time data for dynamic forecasting and decision support.

Keywords: Global Horizontal Irradiance (GHI); Machine Learning; Solar Energy Forecasting; Random Forest.

1. Introduction

With growing emphasis on environmental protection, increasing attention is being paid to solar energy and its applications in daily life. This research focuses on the prediction of the Global Horizontal Irradiation (GHI), defined as the total solar radiation received on a horizontal surface at ground level. The applications of GHI include estimating electricity generation from photovoltaic (PV) plants and improving solar forecasting to balance supply and demand. These functions support governments in managing and planning electricity distribution. The study aims to develop accurate GHI predictions using machine learning techniques. By integrating time series analysis with machine learning, it is possible to obtain daily interactive data and facilitate operational decisions based on predictive results. Seattle's climate, characterized by extensive cloud cover and rainfall, leads to high variability in GHI. The average annual solar radiation in Seattle is approximately 4.12 kWh/m²/day, and the average monthly GHI is about 3.46 kWh/m²/day (Solar Energy Local) [1]. However, Phoenix, AZ, averages over 6.5 kWh/m²/day. These values indicate both the importance and the difficulty of using solar energy in the Seattle area, emphasizing the need for reliable solar prediction. Seattle City Light, the city's public utility, sources over 88% of its electricity from renewable hydroelectric power. About 40–50% is generated internally, with the remainder supplied by the Bonneville Power Administration (BPA) and other renewables, including wind (5%), nuclear (4%), biogas (1%), and other sources (2%) [2]. The Clean Energy Institute (CEI) at the University of Washington, established in 2013, conducts research on solar batteries and grid systems to promote renewable energy adoption and environmental sustainability [2]. These factors highlight the importance of developing prediction algorithms to enhance local solar energy utilization. GHI serves as a fundamental indicator for solar energy potential. It informs site selection, system design, and the development of solar energy facilities. Accurate GHI prediction aids grid operators in balancing supply and demand, mitigating instability risks, and enhancing building energy performance through improved control of lighting and shading systems. With informed predictions, governments can establish efficient energy strategies and distribution plans that increase supply without raising carbon emissions [2].

2. Literature Review

Traditional statistical models, such as ARIMA, exhibit certain limitations when applied to GHI prediction [3]. As a relatively simple and interpretable model, ARIMA assumes that the time series is stationary and that variables are linearly dependent. GHI fluctuations are influenced by multiple factors. Notably, daytime and nighttime data display distinct characteristics: daytime GHI shows significant variation due to solar activity, while nighttime data remain relatively stable and stationary. Moreover, factors such as temperature and dew point are not necessarily linearly correlated. As a result, traditional statistical models are not well-suited for accurate GHI forecasting. Given these limitations, nonlinear models such as Random Forests and Decision Trees offer promising alternatives [2, 4]. These machine learning approaches can capture complex relationships without relying on linear assumptions, thereby improving prediction accuracy. Existing models, such as Support Vector Regression (SVR), are widely used in short-term GHI forecasting, while Random Forest (RF) performs effectively with multi-variable weather data and short-term predictions [2, 5]. Based on these considerations, this study will focus on Linear Regression (accounting for potential linear relationships), Decision Trees, and Random Forests (addressing nonlinear variable interactions) [2]. This research aims to provide policymakers and research institutions with valuable references for urban energy infrastructure planning, helping to avoid under- or over-investment in energy installations [3]. In addition, applied G prediction supports carbon reduction efforts, a topic of major global importance in recent years [1]. It further promotes interdisciplinary research integrating environmental science and artificial intelligence. AI-based models facilitate the processing of large-scale datasets, significantly reducing the time and effort required for data cleaning and multi-dimensional variable integration [2]. They also offer robust and accurate forecasting tools. Moreover, GHI prediction enables governments to formulate dynamic renewable energy policies and implement adaptive strategies in real time.

3. Dataset Description

3.1 Data Sources

Based on all the references, the data is sourced from the National Solar Radiation Database for Seattle 2023, which includes records from 1 January 2023 to 31 December 2023 and contains all features necessary for prediction [2]. This source is reliable and efficient, as it provides location-specific and timestamped data at regular intervals, making it suitable for time-series analysis.

3.2 Target Variable: Global Horizontal Irradiance (GHI)

The target variable is Global Horizontal Irradiance (GHI), which measures the total solar radiation—including both direct and diffuse sunlight—incident on a horizontal surface on the Earth.

3.3 Auxiliary Data and Splitting Strategies

3.3.1 Day/Night Dataset Splitting

The dataset is divided into daytime and nighttime subsets based on sunrise and sunset times in Seattle for 2023, obtained via a sunrise-sunset API.

3.3.2 Train and Test Data Split

The entire hourly dataset for 2023 is partitioned into 80% for training (approximately 7,000 hours) and 20% for testing (approximately 1,800 hours). This ensures sufficient data for model development while preserving an independent test set for performance evaluation.

3.4 Data Cleaning and Visualization

All missing and NA values were removed, and feature values were analyzed using consistent units. The dataset comprises 22 features in total: *Temperature, Alpha, AOD, Asymmetry, Clearsky DHI,*

Clearsky DNI, Clearsky GHI, Cloud Fill Flag, Cloud Type, Dew Point, DHI, DNI, Fill Flag, Ozone, Relative Humidity, Solar Zenith Angle, SSA*, Surface Albedo, Pressure, Precipitable Water, Wind Direction, Wind Speed.* Principal Component Analysis (PCA) was initially applied for dimensionality reduction. However, the result retained approximately 18 components, indicating limited linear correlation among variables and inefficacy of PCA for this dataset. This further underscores the limitations of traditional statistical models and supports a focused analysis on GHI. Given the significant difference between daytime and nighttime GHI patterns, the data were split accordingly to improve modeling. Day/night splitting was determined using Seattle's sunrise and sunset times for 2023. Visualization clearly shows that GHI fluctuates actively during the daytime and remains near zero at night. Figure 1 clearly shows that the GHI value is actively fluctuating during daytime, and almost at value 0 during nighttime [7].

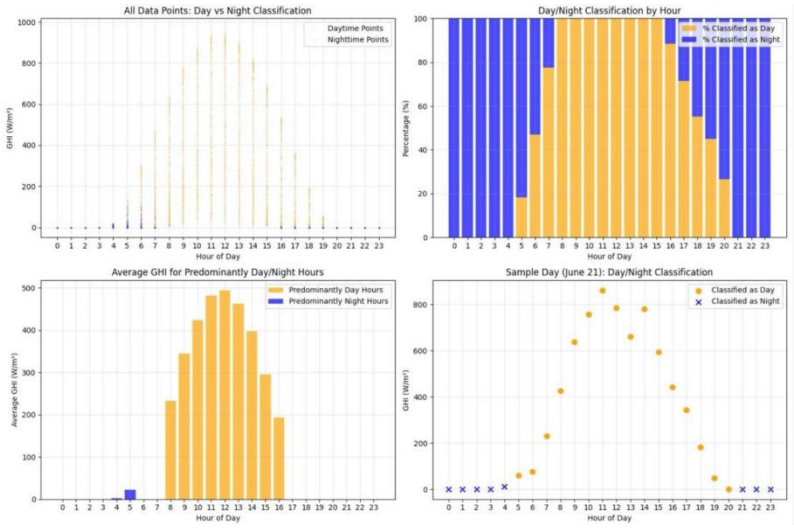

Figure 1. Day/Night classification

3.4.1 Short-term GHI Prediction Using Machine Learning (10 Days → 1 Day)

A sliding window method was employed, using the past 10 days (240 hours) of data to predict GHI values for the next 24 hours. This approach helps temporal pattern recognition in machine learning models.

3.4.2 Model Comparison and Performance Evaluation

The following information summarizes the performance of different models under day/night splitting and unified dataset conditions:

- Linear Regression performed best with day/night split models.
- Decision Trees also achieved the best performance under day/night models.
- Random Forest yielded the best results using the unified dataset (without day/night split).

Note: "Unified model" refers to the approach where the dataset is not split into daytime and nighttime subsets.

4. Methodology

4.1 Regression Models

Three regression approaches were applied to model GHI: Linear Regression, Decision Tree, and Random Forest.

4.1.1 Linear Regression

Linear Regression model serves as a baseline, providing interpretable results but limited capacity to capture the nonlinear variability of GHI.

- Ordinary least squares were used (no regularization).

- Ridge ($\alpha \in \{0.1, 1, 10\}$) and Lasso ($\alpha \in \{0.01, 0.1\}$) were also tried, but no improvement over OLS was observed.
- Thus, regularization was not useful in this case, and the ordinary least squares (OLS) method was applied.

4.1.2 Decision Tree

Decision Tree model captures nonlinear relationships in GHI data, offering flexibility, though it may suffer from overfitting on noisy patterns.

- Use DecisionTreeRegressor (max_depth=10, random_state=42) to model GHI as a function of the other weather features.
- After importing and instantiating the regressor with a fixed random_state for reproducibility, we call. fit () to train two separate trees on daytime and nighttime data.

4.1.3 Random Forest

Random Forest aggregates multiple Decision Trees, reducing overfitting and achieving higher accuracy in predicting complex GHI fluctuations.

- RandomForestRegressor (n_estimators=100, max_depth=10) on all-hours weather data → predicting continuous GHI (train + validation RMSE & R^2).
- on the same features → predicting the binary Is_Daylabel (train/validation accuracy).

4.2 Time-Series Forecasting via Sliding Window

To capture temporal dependencies in GHI, a sliding window approach was applied:

- A sliding window approach was applied using 240 hours of historical data to predict the next 24 hours of GHI values.
- The 2D input array (240 × number of features) was flattened into a 1D vector to be used as input for machine learning models.
- This method uses past data points as input to predict future values, capturing temporal patterns through sliding windows in machine learning models.

The sliding window method provides an effective way to transform sequential GHI data into a supervised learning format. By using 240 hours of past observations as input to predict the subsequent 24-hour output, the model captures short-term temporal dependencies in solar radiation. In this study, the approach is applied with Linear Regression, Random Forest, and Decision Tree models. Linear Regression provides a simple baseline but struggles with nonlinear dynamics, while Decision Trees and Random Forests capture more complex patterns. The data, provided in hourly resolution and preprocessed by the source database, was used without further standardization. Overall, the sliding window enables effective model training by converting time-series GHI data into structured input-output samples.

4.3 Evaluation Metrics

The performance of models was evaluated using Root Mean Square Error (RMSE) and coefficient of determination (R^2).

$$RMSE=\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\widehat{y_t}\right)^2} \quad (1)$$

$$R^2=1-\frac{\sum_{t=1}^{n}\left(y_t-\widehat{y_t}\right)^2}{\sum_{t=1}^{n}\left(y_t-\bar{y}\right)^2} \quad (2)$$

Y_t = actual observed value at time t
$\hat{y}_t$ = predicted value at time t
$\bar{y}$ = mean of the observed values
n = total number of observations.

5. Results

5.1 Hourly Forecasting Results (Day, Night, Unified Datasets)

The experimental results, summarized in **Table 1**, demonstrate clear differences in the performance of Linear Regression, Decision Tree, and Random Forests across day, night, and unified datasets.

Table 1. Performance of Linear Regression, Decision Tree, and Random Forest on Day, Night, and Unified Data

Model	Scenario	Train RMSE (W/m^2)	Test RMSE (W/m^2)	Train R^2	Test R^2
Linear Regression	Day	50.89	48.61	0.9651	0.9657
	Night	0.51	0.46	0.9962	0.9954
	Unified	38.20	37.52	0.9758	0.9760
Decision Tree	Day	41.79	45.18	0.9765	0.9704
	Night	0.70	0.85	0.9928	0.9846
	Unified	33.29	32.91	0.9816	0.9815
Random Forest	Day	3.51	8.88	0.9998	0.9989
	Night	0.38	0.53	0.9979	0.9939
	Unified	2.58	6.04	0.9999	0.9994

The experimental results demonstrate clear differences in the performance of Linear Regression, Decision Tree, and Random Forests under day, night, and unified datasets. Linear Regression serves as a simple baseline but shows limitations in capturing nonlinear patterns of GHI. Its daytime errors are high (Test RMSE = 48.61, R^2 = 0.9657), while nighttime performance improves significantly (Test RMSE = 0.46, R^2 = 0.9954) due to the stable irradiance conditions. The unified model reduces errors compared to the day case but remains less competitive than tree-based approaches. The Decision Tree model performs better at learning nonlinear relationships, with higher R^2 values across all scenarios. However, it shows signs of overfitting: training errors are very low, yet test errors remain relatively large (e.g., Day Test RMSE = 45.18). This indicates limited generalization despite a strong fit on training data. As shown in Table 1, Random Forest consistently outperforms the other models. By combining multiple Decision Trees, it achieves the lowest test errors (Day = 8.88, Night = 0.53, Unified = 6.04) and the highest R^2 values (>0.99), demonstrating both accuracy and robustness. Overall, Random Forest proves to be the most reliable model for GHI prediction, especially in handling the variability of daytime data while maintaining stable nighttime performance.

5.2 Comparative Model Analysis

Table 2. Comparative Model Analysis.

Model	RMSE ↓ (Better is Lower)	R^2 Score ↑ (Better is Higher)
Linear Regression	147.87	0.5913
Decision Tree	60.13	0.7147
Random Forest	48.10	0.8229
Summary	Random Forest outperformed all models, achieving the lowest RMSE and highest R^2, indicating strong accuracy and generalization. Decision Tree performs moderately well, significantly better than Linear Regression in both metrics, capturing nonlinear patterns. Linear Regression performs the worst, suggesting it fails to capture complex patterns in the data due to its linear nature.	

Table 2 shows the order of the values of our models, and we can notice that if we have the lower RMSE values, the results are better, and if we have the higher R^2 scores, the results are better. By comparison, we can figure out that our Random Forests work best.

5.3 Time-Series Forecasting Results (10-Day → 1-Day Prediction)

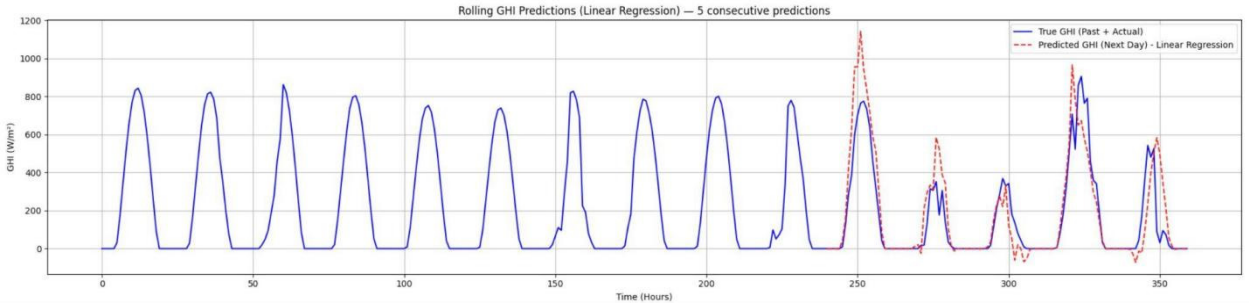

Figure 2. Linear Regression model with GHI forecasting.

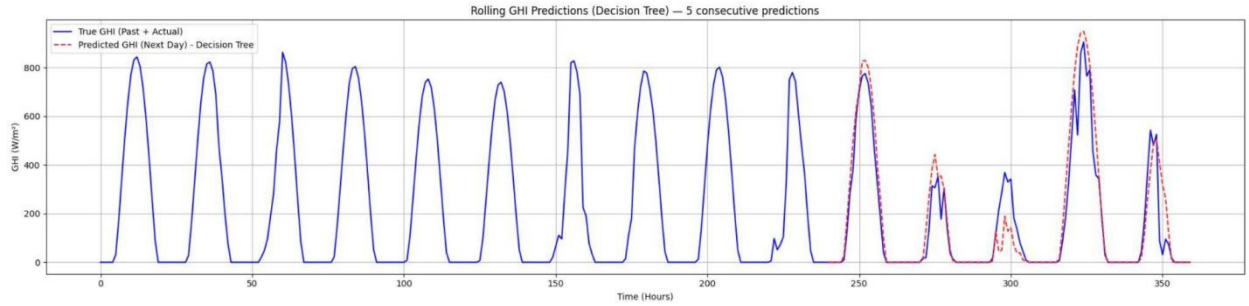

Figure 3. Decision Tree model with GHI forecasting.

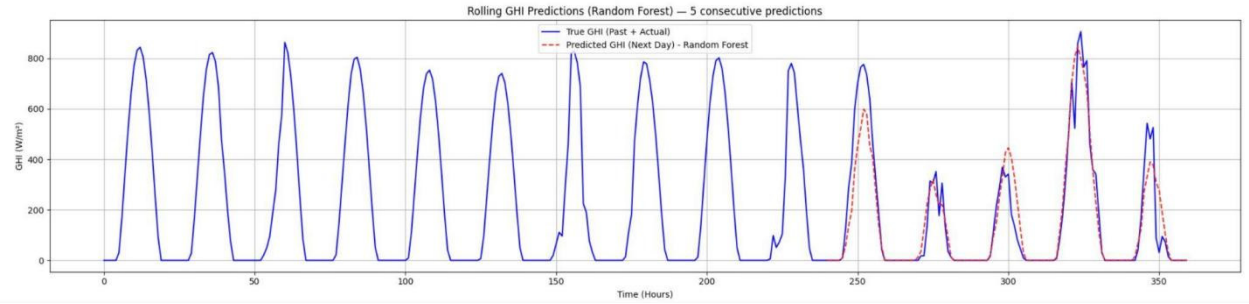

Figure 4. Random Forest with GHI forecasting.

Table 3. Short-Term GHI Forecasting Performance of Different Models.

Model	RMSE	R^2
Linear Regression	147.87	0.5913
Decision Tree	60.13	0.7147
Random Forest	48.10	0.8229

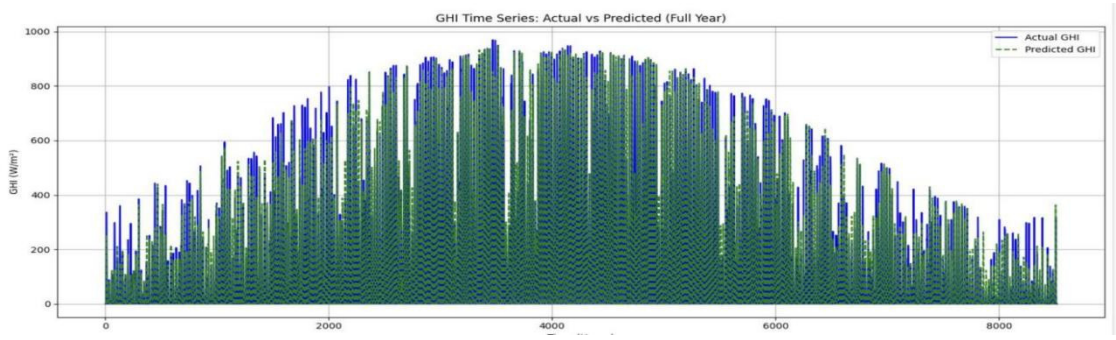

Figure 5. GHI Time Series: Actual vs Predicted (whole year).

As shown in Table 3, Linear Regression model for GHI forecasting yielded an RMSE of 147.87 and an R^2 of 0.5913. Decision Tree demonstrated improved performance metrics over the Linear Regression model. Random Forest achieved the highest predictive accuracy among all models tested. A full-year prediction was generated using the Random Forest. Using the Random Forest to get the whole year prediction:

6. Discussion

6.1 Main Findings

Based on all the discussions and research, we can figure out that our machine learning models are good for GHI prediction, and they have huge potential to improve the strategies. The performance of the Linear Regression model suggests that the variables may not have strong linear relationships with the target. The attempt to use PCA for dimensionality reduction was not successful, indicating that the variables are not necessarily interdependent and may be largely independent. While the Decision Tree model performed better than Linear Regression, which is suitable for capturing nonlinear relationships, its performance on unseen data suggests potential overfitting and associated uncertainties. The superior performance of the Random Forest is attributed to its inherent mechanism of averaging multiple Decision Trees, which mitigates overfitting and improves generalization. This model is recommended for applications where computational resources are sufficient. However, for very large datasets, the computational demands and longer processing times of the Random Forest may present practical limitations.

6.2 Comparison with State-of-the-Art Approaches (e.g., LSTM, SARIMA)

With the current discussions and research about the prediction of GHI values, we can try to figure out how people use deep learning and other models like LSTM. They combine them to analyze massive datasets and get more precise results. Our models are not really good enough. Our data is not large, which is only about one year of data, and is not sufficient for us to make a decision. And our models are sensitive to noise, so that our real-world data will affect the accuracy of results, since real-world data is random and not ideal as we expected. For the strong time sequence characteristics (such as self-correlation, periodicity) and spatial dependence (such as the influence of peripheral sites) in GHI prediction, the ability to directly capture standard Decision Trees and Random Forests is not as good as models specially designed for sequence modeling (such as LSTM, SARIMA) or models that consider spatial relationships [6]. We want to consider including more models and combining the advantages and avoiding the disadvantages with different models, and consider more situations together to get a more accurate result.

6.3 Research Limitations

No matter how we develop our models and the strategy we want to combine with, we usually face the challenging facts that operating models need huge computational resources and require highly calculating performance, which will cost a lot of time and effort. The strategy we try to use will still need to consider the abilities of calculations and the performance of the machines. During the selection process, choosing an accessible model is one factor to consider.

7. Future Work

Based on the current study, we plan to explore more advanced models, such as Deep learning, SVM, LSTM, etc. Deep learning models will represent the most advanced level of current GHI prediction, which can capture space-time features at the same time, greatly improving the accuracy of prediction, especially short-term and ultra-short-term forecasts. For SVM models, it efficiently deals with nonlinear problems: By selecting appropriate nuclear functions (such as RBF), SVM can very effectively capture the complex nonlinear relationship between GHI and meteorological factors

(such as cloud, temperature, humidity) without manual feature transformation. For the LSTM model, with the support of sufficient data, LSTM can usually achieve higher prediction accuracy than SVM and traditional machine learning models, especially when predicting and capturing complex dynamic changes in multiple steps. Finally, we intend to extend the prediction horizon beyond 24 hours, possibly forecasting up to 7 or 10 days. We hope to develop a lightweight web interface to visualize predictions and support real-time solar energy planning, since our goal is to help people and the government plan and develop strategies for energy resources. So, for people who have no idea about how the models perform and how we can build these models, but they can still experience and understand the GHI and how they can use the data from the GHI to plan their life, which is our goal to help others, even though they have no background in coding and machine learning.

8. Conclusion

This study evaluates the performance of various machine learning models for predicting global horizontal irradiance (GHI). Among Linear Regression, Decision Tree, and Random Forests, the Random Forest achieved the best performance, with the lowest RMSE (6.04 W/m²) and the highest R^2 score (0.9994) on the test set, demonstrating exceptionally high predictive accuracy and strong generalization capability. Although the Decision Tree model exhibited very high training accuracy, significant overfitting was observed, indicating limited generalization ability. The Linear Regression model performed the worst, confirming that the relationship between GHI and meteorological features is complex and nonlinear, and cannot be adequately captured by a simple linear approach. Furthermore, the study revealed that training a unified model on full-day data yields satisfactory predictions, making it unnecessary to separate daytime and nighttime data. These findings highlight the Random Forest as an efficient and reliable tool for GHI prediction. Its ensemble learning mechanism effectively captures complex nonlinear relationships, offering a practical solution for forecasting solar power generation.

References

[1] Singhal, R., Singhal, P., & Gupta, S. (2022). Solar-Cast: Solar power generation prediction from weather forecasts using machine learning. 2022 IEEE 10th Power India International Conference (PIICON), 1–6. https://doi.org/10.1109/PIICON56320.2022.10045237.

[2] Seattle City Light. (2025, March 21). Celebrating renewable energy. Powerlines. https:// powerlines. seattle. gov/2025/03/21/celebrating-renewable-energy/.

[3] Zeng, J. W., & Qiao, W. (2013). Short-term solar power prediction using a support vector machine. Renewable Energy, 52, 118–127. https://doi.org/10.1016/j.renene.2012.10.009.

[4] Hiremath, V., Naik, R., Naik, S., Shettar, S., & Chachadi, K. (2024). GHI prediction based on weather data using machine learning. 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), 1–6. https://doi.org/10. 1109/ ICITEICS 61368. 2024.10624921.

[5] Narváez, G., Giraldo, L. F., Bressan, M., & Pantoja, A. (2021). Machine learning for site-adaptation and solar radiation forecasting. Renewable Energy, 167, 333–342. https://doi.org/ 10.1016/ j.renene. 2020. 11. 089.

[6] Yu, Y., Cao, J., & Zhu, J. (2019). An LSTM short-term solar irradiance forecasting under complicated weather conditions. IEEE Access, 7, 145651–145666. https://doi.org/10.1109/ACCESS.2019.2946057

[7] Sunrise-Sunset API. Sunrise and sunset times for specific coordinates. https://api.sunrise-sunset.org/ json? lat={lat}&lng={lon}&date={date_str} &formatted=0.