Journal of
Sensor and Actuator Networks

MDPI

Article

Generalizable Solar Irradiance Prediction for Battery Operation Optimization in IoT-Based Microgrid Environments

Ray Colucci and Imad Mahgoub *

Department of Electrical Engineering and Computer Science, Florida Atlantic University, 777 Glades Rd, Boca Raton, FL 33431, USA; rcolucci@fau.edu
* Correspondence: mahgoubi@fau.edu

Abstract: The reliance on fossil fuels as a primary global energy source has significantly impacted the environment, contributing to pollution and climate change. A shift towards renewable energy sources, particularly solar power, is underway, though these sources face challenges due to their inherent intermittency. Battery energy storage systems (BESS) play a crucial role in mitigating this intermittency, ensuring a reliable power supply when solar generation is insufficient. The objective of this paper is to accurately predict the solar irradiance for battery operation optimization in microgrids. Using satellite data from weather sensors, we trained machine learning models to enhance solar irradiance predictions. We evaluated five popular machine learning algorithms and applied ensemble methods, achieving a substantial improvement in predictive accuracy. Our model outperforms previous works using the same dataset and has been validated to generalize across diverse geographical locations in Florida. This work demonstrates the potential of AI-assisted data-driven approaches to support sustainable energy management in solar-powered IoT-based microgrids.

Keywords: machine learning; solar energy; ensemble learning; cross-validation; solar irradiance prediction; battery energy storage system; renewable energy

Academic Editor: Stefan Fischer

Received: 5 November 2024
Revised: 10 December 2024
Accepted: 21 December 2024
Published: 27 December 2024

Citation: Colucci, R.; Mahgoub, I. Generalizable Solar Irradiance Prediction for Battery Operation Optimization in IoT-Based Microgrid Environments. *J. Sens. Actuator Netw.* **2025**, *14*, 3. https://doi.org/10.3390/jsan14010003

1. Introduction

In our previous work, we proposed to predict solar irradiance accurately to provide input to battery operation optimization [1]. In this paper, we use cross-validation to further enhance solar irradiation prediction results and generalize the model to predict solar irradiance in different geographical locations in Florida.

The global shift towards renewable energy sources, including solar and wind, is redefining the landscape of electrical power generation. Currently, the world relies heavily on fossil fuels—natural gas, coal, and oil—which are finite resources with unsustainable consumption levels. At the current rate of use, carbon emissions are projected to reach between 8 and 12 gigatons by 2030, representing a 62% increase from 2002 levels [2]. This looming increase underscores the urgent need for research to incorporate technologies that mitigate environmental impacts and harness renewable energy.

One essential area of renewable energy research involves the prediction of solar irradiance, which measures power density over a given area. Accurate solar energy forecasting is critical for optimizing control mechanisms in energy systems that depend on inherently intermittent power sources [3]. For instance, predicted solar irradiance values can serve as key input constraints for optimizing battery energy storage systems (BESS), where maintaining an adequate state of charge (SOC) is crucial to meet demand without risking depletion [4].

Potential applications of this research span residential and commercial buildings, which can utilize microgrids—self-contained power grids that can optionally integrate with the main utility grid. These microgrids, comprising renewable energy sources and BESS, enable efficient, cost-effective energy supply solutions [5]. Additionally, determining the optimal sizing of photovoltaic (PV) solar panels and BESS remains a critical research focus to ensure sufficient energy while minimizing overall system costs [6]. This focus is driven by a growing demand for precise modeling and forecasting of weather parameters to support the broader deployment of renewable energy technologies [3].

The broader environmental impact of using fossil fuels as an energy source is significant and multifaceted, encompassing various ecological, climatic, and health-related challenges such as greenhouse gas emissions, climate change, air pollution, health hazards, environmental degradation, non-renewability, resource depletion, energy inefficiency and economic costs. However, due to the intermittency of renewable energy sources, they are a challenge to deploy in an effective manner. A viable alternative to solve this issue is to utilize a BESS so that when the energy input is not available, an optimized BESS operation strategy will provide a reliable way to provide energy on demand, reduce costs, and lower carbon emissions.

For BESS, further objectives include extending battery lifespan and reducing maintenance costs. Effective optimization algorithms can maximize battery life and economic return, which are vital for the sustainability of renewable energy systems [4]. A range of approaches has been explored, including multi-objective methods, iterative techniques, graphical analyses, and artificial intelligence-based strategies [2].

Recent studies show that historical solar irradiance data are often used as input for optimization, with machine learning algorithms incorporated to account for the inherent variability in solar radiation [2]. Machine learning models have been developed for solar irradiance prediction. These models leverage historical data and advanced algorithms to forecast solar energy potential, aiding in the efficient operation of renewable energy systems. By incorporating features such as temperature, humidity, wind speed, and atmospheric clarity, these models can effectively account for the complex and nonlinear relationships influencing solar energy generation. Advanced techniques like ensemble learning and feature selection further enhance prediction accuracy, making these models robust tools for renewable energy management. Accurate solar irradiance forecasts not only optimize energy generation but also support strategic battery energy storage system (BESS) operations, enabling better alignment of energy supply with demand. This predictive capability plays a critical role in mitigating the challenges posed by the intermittency of solar energy [1].

In [7], a Bayesian optimization-based regression tree algorithm was employed to predict global horizontal solar irradiance, enabling the calculation of an optimal PV system size for residential applications. Similarly, ref. [8] introduced a microgrid solar energy prediction model using weather input data and a Robust Optimized Functional Link Broad Learning System (FLBLS) enhanced by exponential trigonometric functions. However, this approach relied on only one year of historical data. In [9], a comparison between ANN and multiple linear regression (MLR) forecast models was conducted using PV panel and weather data, with micro inverter technology improving power output forecasts for battery control, although the training data spanned just six months. Additionally, ref. [10] utilized deep neural networks and extreme gradient boosting forest algorithms with inputs such as relative humidity, clear sky index, and temperature for solar irradiance prediction. A bidirectional long short-term memory network combined with wavelet transform was further proposed in [11] for enhanced GHI prediction.

The integration of solar energy into power grids depends on accurate solar irradiance forecasting, and a study [12] conducted in Douala, Cameroon, introduces an innovative approach combining Bayesian Optimized Attention-Dilated Long Short-Term Memory (LSTM) and Savitzky-Golay filtering. The proposed methodology enhances data quality through pre-processing and optimizes deep learning models, achieving exceptional performance with a Symmetric Mean Absolute Percentage Error of 0.6564 and a Normalized Root Mean Square Error of 0.2250. This research not only surpasses prior studies but also offers significant contributions through novel hybrid deep-learning architectures and benchmark datasets, benefiting both researchers and solar energy managers.

The integration of photovoltaic (PV) systems into the energy sector has accelerated due to environmental concerns and advancements in renewable energy technologies. Accurate prediction of temperature and solar irradiance is critical for optimizing PV system performance and grid integration, with machine learning (ML) emerging as a powerful tool for enhancing forecast accuracy. A comprehensive review [13] compares various ML algorithms, including decision trees, random forests, XGBoost (version 2.1.3), and support vector machines, highlighting their advantages over traditional meteorological models in improving PV power generation forecasts.

Photovoltaic panels offer a sustainable solution for generating electricity by converting solar radiation into power, reducing dependence on fossil fuels. The efficiency of this process depends on factors such as panel quality and accurate forecasting of region-specific solar irradiance, which is essential for designing and managing effective solar power systems. A study [14] utilizing three years of data from Izmir, Turkey, compares various machine learning and deep learning algorithms for solar irradiance prediction, with the multilayer perceptron emerging as the most effective model.

Managing energy produced by microgrids requires intelligent, efficient strategies that maximize profitability while meeting consumer demand. Ideally, such strategies should reduce reliance on the public utility grid, especially during peak demand periods, by enabling the microgrid to sell excess energy back to the grid, thereby lowering electricity prices. Embracing renewable energy not only fosters environmental preservation but also provides an inexhaustible energy supply. This transition motivates the adoption of green energy buildings, which is achievable through home energy management systems and smart microgrid integration, which optimize resource use and operational efficiency.

2. Materials and Methods

To develop solar irradiance prediction models, we utilized time series data sourced primarily from the NASA POWER database, which collects satellite-derived measurements. This choice was made due to the limited accessibility of ground-based stations, and the satellite data provided a comprehensive view by capturing varied atmospheric conditions.

While the dataset's comprehensiveness and global coverage are significant strengths, we acknowledge certain limitations inherent to satellite-derived data. Weather satellites rely primarily on remote sensing technology and spatial imagery to gather weather data, enhanced by localized ground measurements for calibration and improved accuracy. Satellite data, including that from the NASA POWER database, often have lower spatial and temporal resolution compared to ground-based measurements. This may lead to discrepancies in regions with highly localized weather phenomena or rapidly changing atmospheric conditions.

The main output parameter was the total solar irradiance, which includes both direct and diffuse irradiance on a horizontal or tilted plane at the Earth's surface under all sky conditions, measured in kW-h/m^2/day. We accessed six years of data, spanning 2015 to 2020, using the POWER Data Access Viewer for the SSE-Renewable Energy Community.

The dataset we used is from Bhopal, India and can be accessed through the following URL https://power.larc.nasa.gov/data-access-viewer/ (accessed on 9 June 2024).

The dataset includes key atmospheric and environmental input variables found in Table 1:

Table 1. Input parameters used for machine learning algorithms.

Input Parameter	Description
Temperature at 2 m	average air temperature at 2 m above ground
Relative Humidity at 2 m	ratio of actual partial water vapor pressure to saturation pressure, in percent
Precipitation Corrected	bias-corrected average total surface precipitation
Wind Speed at 10 m	average wind speed at 10 m above ground
All Sky Insolation Clearness Index	fraction representing atmospheric clearness; the ratio of all-sky irradiance reaching the surface to the top-of-atmosphere solar irradiance
Earth Skin Temperature	average ground surface temperature
Surface Pressure	average surface atmospheric pressure

2.1. Machine Learning Algorithms

Five supervised machine learning regression algorithms were chosen for this study to map these seven predictor variables to output solar irradiance values. The model was trained on data from 2015 to 2019, with 2020 reserved for validation. Therefore, 83% of the dataset was used for training and the remaining 17% for testing. One record was created for each day of the year. The algorithms are as follows:

1. Random Forest (RF): This ensemble learning method constructs multiple decision trees during training and averages their outputs to make predictions, reducing overfitting and increasing accuracy. It performed exceptionally well on our dataset, generally achieving the highest R^2 value and the lowest error metrics due to its robustness in handling nonlinear patterns and capturing feature interactions.
2. Extreme Gradient Boosting (XGBR): This method builds sequential models that correct errors from previous iterations, making it efficient for handling complex datasets. XGBR demonstrated strong performance, with metrics closely aligned with those of RF, particularly in capturing subtle variations in solar irradiance data.
3. Kernel Ridge Regression (KR): KR combines ridge regression with kernel functions, enabling it to model nonlinear relationships effectively. It performed moderately well, reflecting its ability to generalize but with higher sensitivity to parameter tuning compared to ensemble methods.
4. Support Vector Regression (SVR): SVR aims to find a hyperplane that fits the data within a margin of tolerance. In this study, SVR was implemented with hyperparameter tuning using GridSearchCV, optimizing parameters such as C, gamma, epsilon, kernel, and degree to enhance its performance. After tuning, SVR achieved a strong R-squared value of 0.927. This result demonstrates that, with appropriate parameter optimization, SVR can effectively model the data and provide reliable predictions. Given this performance, SVR was included in the ensemble testing phase to further leverage its predictive capabilities within the broader framework of machine learning models.
5. Linear Regression (LR): This simplest regression technique assumes a linear relationship between inputs and outputs. While its performance was relatively modest, LR served as a baseline to compare the more complex models.

We analyzed each algorithm using key metrics, including R^2, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Random Forest generally achieved a superior R^2 value, outperforming the other algorithms by an average of 10%. The results validate RF's suitability for this application, given the dataset's complexity and the inherent variability of solar irradiance.

In our study, we utilized GridSearchCV to systematically explore the optimal combination of hyperparameters for each of the 5 models. We performed parameter tuning on the SVR model. The parameter grid for the SVR model included:

- C: Regularization parameter with values [0.1, 1, 10, 100]
- Gamma: Kernel coefficient with values [1, 0.1, 0.01, 0.001]
- Epsilon: Tube within which no penalty is associated with the training loss function with values [0.01, 0.1, 1, 10]
- Kernel: Types of kernel functions including ['linear', 'poly', 'rbf']
- Degree: Degree of the polynomial kernel function for "poly" kernels with values [3, 6]

These ranges were selected based on common practices in SVR applications and literature to ensure thorough coverage of the parameter space. The optimization process for hyperparameter selection was carried out using GridSearchCV with a 5-fold cross-validation scheme. This method ensures that the best-performing combination of hyperparameters is chosen based on the cross-validation results, balancing bias and variance.

After the GridSearchCV process, the best hyperparameters identified for our dataset are:

- C: 10
- Gamma: 1
- Epsilon: 0.01
- Kernel: 'poly'
- Degree: 3

These parameters were applied to the SVR model to generate predictions on the test set. The performance of the tuned SVR model was evaluated using standard metrics:

- Mean Squared Error (MSE): 0.333
- Mean Absolute Error (MAE): 0.441
- R-squared Score (R^2): 0.927

Figure 1 shows the scatterplot of the actual vs. predicted values after hyperparameter tuning.

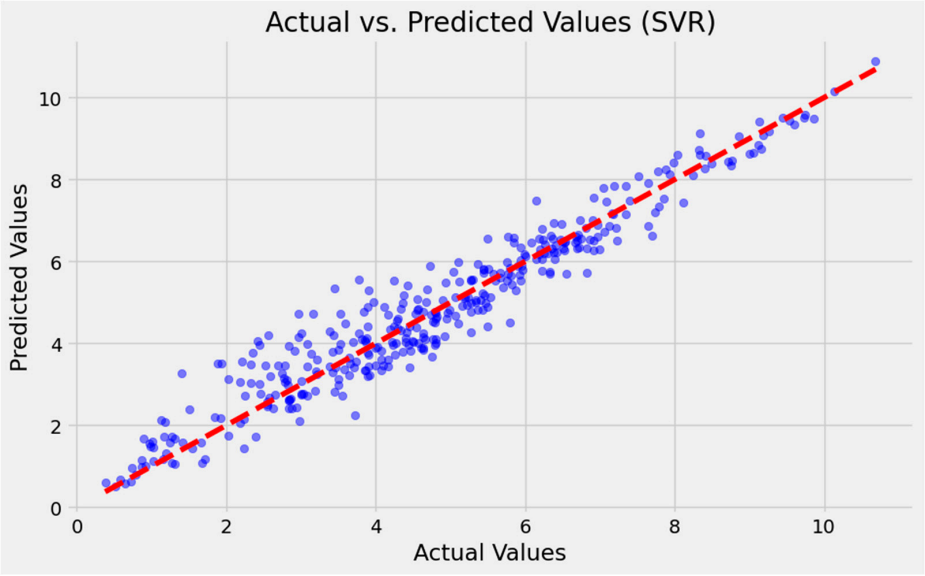

Figure 1. Scatterplot of True vs. Predicted Values for SVR.

Tables 2–4 provide the value and range of each hyperparameter for each tuning of the following 3 models. An optimizer was not explicitly used for Random Forest since the model leverages decision tree-splitting heuristics.

Table 2. Random Forest.

Hyperparameters	Range of Values	Value Used
n estimators	[10, 100, 200]	100
max depth	[3, 5, 10, none]	none
max features	[1, 3, 5, 7]	auto
min samples split	[1, 2, 3]	2
min samples leaf	[1, 2, 3]	1

Table 3. Kernel Ridge.

Hyperparameters	Range of Values	Value Used
alpha	[0.1, 1.0, 10]	1.0
kernel	[linear, poly, rbf]	Linear
gamma	[0.1, 1.0, 10]	none
degree	[3, 6]	3

Table 4. XGBoost.

Hyperparameters	Range of Values	Value Used
learning rate	[0.01, 0.1, 0.3]	0.3
n estimators	[100, 200, 500]	100
max depth	[3, 6, 10]	6
subsample	[0.6, 0.8, 1.0]	1.0
colsample bytree	[0.6, 0.8, 1.0]	1.0
reg alpha	[0, 0.1, 1.0]	0
Reg lambda	[0.1, 1.0, 10]	1.0

An optimizer was not explicitly used for Kernel Ridge since it relies on solving linear systems using the kernel trick. No hyperparameters were used for Linear Regression as it fits a linear model without regularization. The optimizer used was Ordinary Least Squares (OLS).

The optimizer used for XGBoost was the Gradient Boosting Algorithm.

2.2. Machine Learning Methods

To further enhance accuracy, we used ensemble learning techniques that combine multiple models, improving robustness and reducing potential biases:

- Voting: combines predictions from several models to produce a final prediction.
- Stacking: learns an optimal combination of predictions from at least two models using a meta-learner.
- Bagging: uses data resampling with replacement, which reduces model variance and enhances stability.

Datasets were validated across different locations. First, a model was trained on a dataset from India [15] so that we could compare results with existing work using the same

dataset. Once we trained the model, we could check its accuracy and train the model on a dataset with solar irradiance output from tilted solar panels in India.

Using the output parameter for tilted solar panels, All Sky Surface Shortwave Downward Direct Normal Irradiance (DNI) provides more precise and relevant information for solar energy applications, enabling better performance prediction, optimization, and operation of solar energy systems. The output parameter for horizontal solar panels is named All Sky Surface Shortwave Downward Irradiance (SWD).

DNI, as compared to SWD, is relevant to solar energy systems, has a higher concentration of solar energy, has better performance, has optimized tracking systems, and utilizes standardized measurements [16].

The formula to convert solar irradiance captured by solar panels to the solar energy produced for electricity is based on the efficiency and area of the solar panels. The relationship can be expressed as:

$$E = I \cdot A \cdot \eta \quad (1)$$

where:

E: Solar energy produced (in kilowatt-hours or watt-hours)
I: Solar irradiance (in kilowatts per square meter or watts per square meter)
A: Area of the solar panel (in square meters)
η: Efficiency of the solar panel (as a decimal)

We then utilized an implementation of five-fold cross-validation. In our experiment, each fold represents one year of data. Cross-validation is a valuable technique in machine learning for improving its generalization ability. The motivation to use cross-validation is based on its inherent advantages [17–21]. It produces more reliable performance estimates, provides better utilization of data, reduces bias, improves model selection and hyperparameter tuning, overfitting detection, is robust to data imbalance, and allows for easier comparison of models.

Overall, cross-validation is a valuable tool for model evaluation, selection, and improvement in machine learning, contributing to more reliable and robust predictive models. Our goal is to improve the prediction results by utilizing cross-validation.

Since we're interested in predicting solar irradiance in our geographical area and improving performance, we built a model on the Boca Raton, Florida, dataset for tilted panels using five-fold cross-validation. The advantage of using DNI (tilted solar panels) over SWD (horizontal panels) lies in its ability to provide more accurate and precise information about the solar radiation that directly impacts solar energy systems, particularly concentrating solar power (CSP) and photovoltaic (PV) systems [22–24].

India and Florida share several climatic similarities that are relevant for developing machine learning models for solar irradiance prediction. These similarities make them valuable case studies for validating predictive models across diverse but comparable conditions such as abundant solar radiation, tropical and subtropical climates, seasonal variations, high humidity levels, frequent cloud cover and atmospheric changes, and high solar energy potential. By incorporating these shared climatic features into machine learning models, researchers can design predictive frameworks that are transferable between regions, enhancing their generalizability and robustness for solar irradiance forecasting across different geographies.

During the data pre-processing stage, we identified and removed solar irradiance values less than -1 as outliers. Moreover, the machine learning models employed, such as Random Forest and XGBoost, inherently exhibit robustness to outliers due to their reliance on tree-based structures. These models divide the data into smaller subsets during decision tree construction, minimizing the impact of individual extreme values. For the Support

Vector Regression (SVR) model, we performed hyperparameter tuning (e.g., selecting an appropriate kernel and regularization parameters) to mitigate the influence of outliers.

Ensuring the reproducibility of our research is a priority, as it allows for transparency and facilitates further advancements in the field. We applied standard machine learning models, included descriptions of the features we used, and the datasets were available to the public. Therefore, our results can be reproduced using these NASA datasets. A detailed description of the data pre-processing steps, model configurations, and hyperparameters is included in the manuscript to further support reproducibility efforts.

There are uncertainties associated with solar irradiance predictions. Our future work involves developing an algorithm that utilizes solar irradiance predictions as input to optimize BESS operation, which will mitigate these uncertainties. The algorithm is based on fuzzy logic, which we selected because it deals with this uncertainty. This feature is critical for decision-making in energy systems.

After training the model on tilted solar panel datasets, we tested the model to show generalizability by testing the model on datasets belonging to different geographical locations.

We started in a nearby location, Orlando, Florida. The same trained model was tested again using tilted panel output in the Orlando dataset. To generalize the model, we tested datasets belonging to different locations in Florida, namely Orlando, Miami, Tampa, Jacksonville, and Tallahassee.

3. Results

3.1. Machine Learning Algorithms Trained on 2015–2019 Datasets from India Using Horizontal Solar Panel Output

We originally started our experiment with solar irradiance prediction from a horizontal solar panel in a location in India to compare our results against existing work that used horizontal panels and the same dataset [15]. We used the 2015–2019 datasets for training and the 2020 datasets for testing. The prediction results for our machine learning models are presented in Figures 2–6.

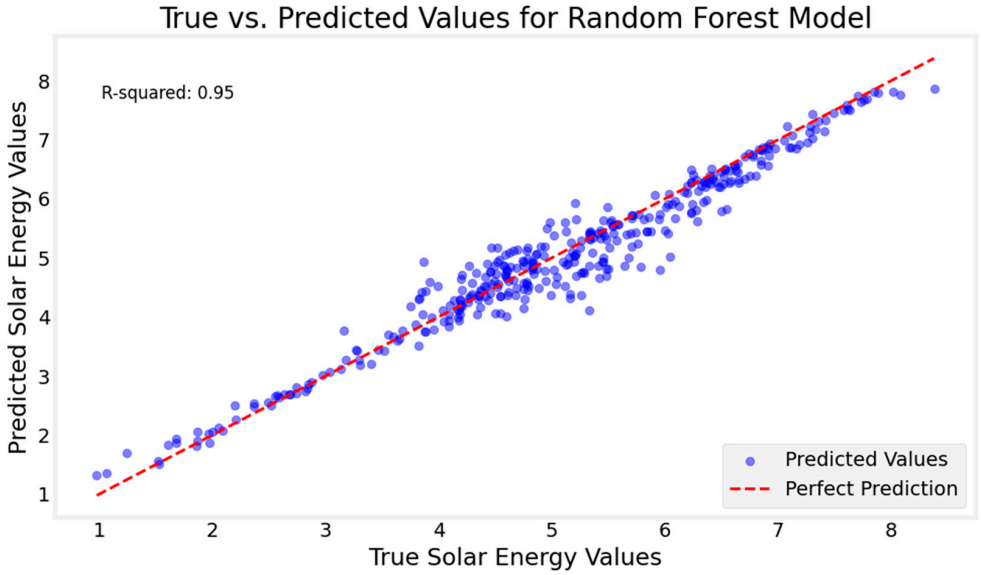

Figure 2. Scatterplot of True vs. Predicted Values for Random Forest Model.

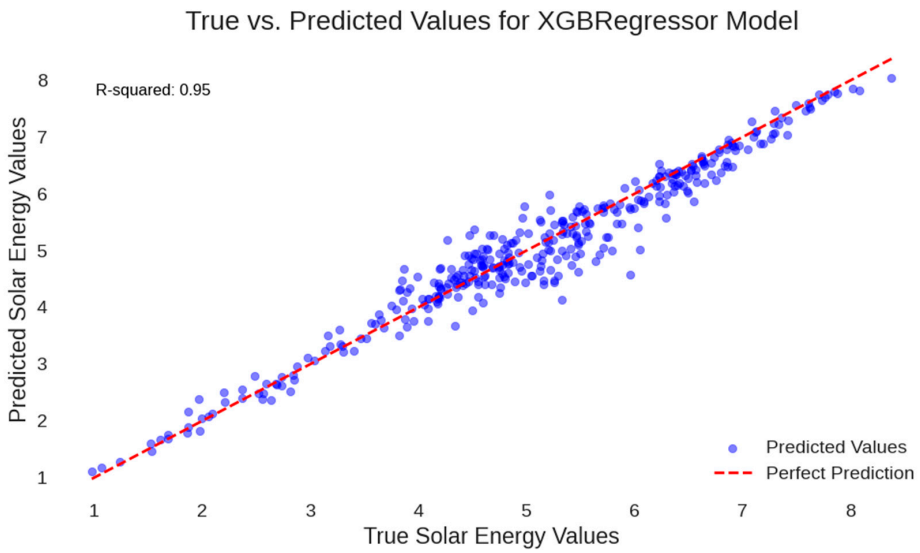

Figure 3. Scatterplot of True vs. Predicted Values for XGBRegressor Model.

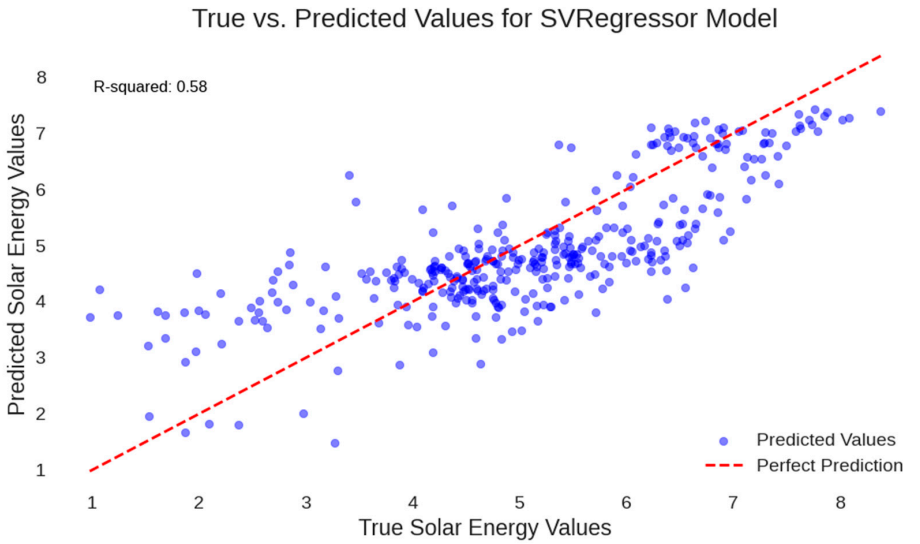

Figure 4. Scatterplot of True vs. Predicted Values for SVRegressor Model.

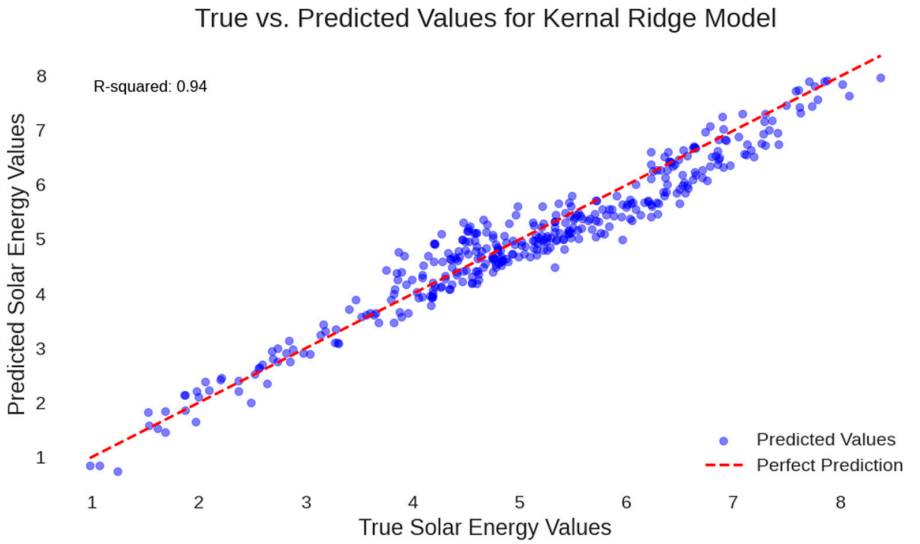

Figure 5. Scatterplot of True vs. Predicted Values for Kernal Ridge Model.

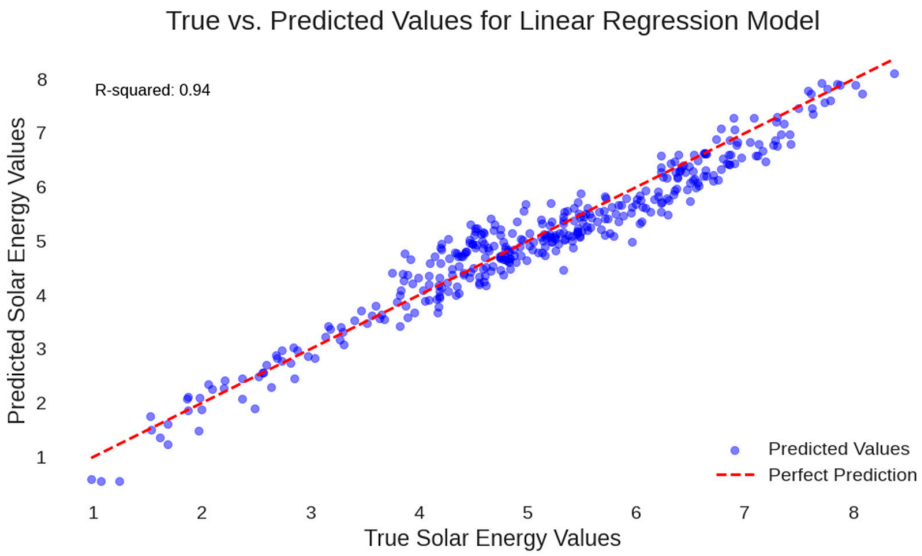

Figure 6. Scatterplot of True vs. Predicted Values for Linear Regression Model.

Table 5 provides four metrics used to compare the five algorithms.

Table 5. Metrics results for algorithms.

ML Algorithm	MSE	RMSE	MAE	R-Squared
Random Forest	0.100	0.317	0.228	0.952
XGBRegressor	0.101	0.318	0.231	0.951
SVRegressor	0.867	0.931	0.735	0.582
Kernal Ridge	0.126	0.355	0.289	0.939
Linear Regression	0.118	0.343	0.278	0.943

The following metrics were used to assess and compare the predictive accuracy of the five algorithms:

- MSE (Mean Squared Error): MSE quantifies the average squared difference between observed and predicted values, with lower values indicating fewer errors. An ideal MSE is zero, increasing as the model's error grows.
- RMSE (Root Mean Squared Error): RMSE is the square root of MSE, offering an interpretable metric in the same units as the data, with lower values showing better model performance.
- MAE (Mean Absolute Error): MAE measures the average magnitude of prediction errors, disregarding direction, calculated as the mean absolute difference between predicted and actual values. Lower MAE values reflect higher prediction accuracy.
- R^2 (Coefficient of Determination): R^2 indicates the proportion of variance in the target variable explained by the model. An R^2 close to 1 suggests that the model effectively explains variability in the data.

To further understand the contribution of input features to the model's predictions, we conducted an analysis of feature importance using the Random Forest algorithm, which inherently provides a measure of feature significance based on the decrease in impurity (Gini importance). The results revealed that temperature at 2 m and the all-sky insolation clearness index were the most significant predictors, as they strongly influence solar irradiance variability. Features such as wind speed at 10 m and surface pressure showed moderate importance, while relative humidity and precipitation had a lesser but still notable impact. The Random Forest algorithm evaluates the significance of each input feature by measuring its contribution to reducing impurity in the decision trees. Figure 7

provides a summary of the metrics and their importance scores for the seven input features used in the model:

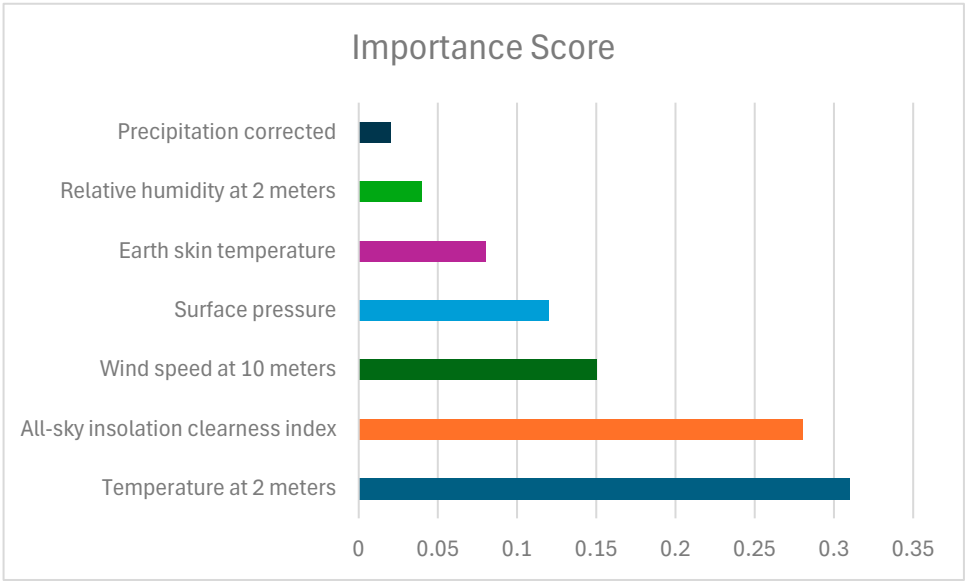

Figure 7. Summary of the importance scores for the input features used in the model.

3.2. Ensemble Models Trained on 2015–2019 Datasets from India Using Horizontal Solar Pane Output

To enhance predictive performance beyond that of individual machine learning algorithms, we employed ensemble methods. Rather than relying on a single model, ensemble approaches combine multiple models to boost the accuracy of key metrics.

The stacking regressor achieved the highest performance among the three ensemble learning methods. Stacking, or stacked generalization, integrates multiple regression models by training a meta-regressor on their combined predictions. This approach excels due to several factors: it leverages model diversity, effectively integrates the meta-learner, improves overall performance, handles nonlinear and complex patterns, and enhances robustness. Success in stacking is typically influenced by the diversity and quality of the base models and the fine-tuning of the meta-learner, which blends the base models' predictions effectively. The efficacy of a stacked model can vary based on the selection of base models, data characteristics, and integration within the stacking framework. Table 6 summarizes the comparative results of our individual algorithms, ensemble methods, and the work in [15] using the same dataset as our study sourced from NASA's POWER database. To the best of our knowledge, the only additional research utilizing this dataset is the work in [15], which serves as a direct comparison to our methodology.

Table 6. Comparison Of Our Work and Other Work That Used the Same Dataset [15].

Dataset Used	Method	R-Squared Results
Our work using NASA POWER Dataset [1]	Stacking (Ensemble)	0.9546
	Voting (Ensemble)	0.9536
	Bagging (Ensemble)	0.9511
	Random Forest	0.9513
	XGBR	0.9510
	LR	0.9430
	KR	0.9392
	SVR	0.5817

Table 6. *Cont.*

Dataset Used	Method	R-Squared Results
Other work that used the NASA POWER Dataset [15]	Bidirectional LSTM	0.7064
	GRU	0.7032
	CNN	0.7025
	Attention LSTM	0.6965
	LSTM	0.6906

3.3. Five-Fold Cross-Validation Using 2015–2019 Datasets from India Using Horizontal Solar Panel Output

Five-fold cross-validation is performed in 2015–2019 datasets, with each year serving as one-fold. Table 7 illustrates the mean R-squared values of the five folds for the four machine learning algorithms that produced the best prediction results with five-fold cross-validation. Random Forest gave the most accurate results of all machine learning algorithms.

Table 7. R-Squared Values for Machine Learning Algorithms With Five-Fold Cross-Validation.

Method	Mean R-Squared
Kernal Ridge	0.9531
Linear Regression	0.9564
XGBoost	0.9677
Random Forest	0.9700

Table 8 illustrates the mean R-squared values of the Five folds for the three ensemble methods with five-fold cross-validation. The stacking regressor provided the best prediction results of the ensemble learning methods.

Table 8. R-squared values for Ensemble Methods With Five-Fold Cross-Validation.

Ensemble Method	Mean R-Squared
Stacking	0.9730
Bagging	0.9698
Voting	0.9703

Figure 8 presents a graphical representation of the comparison of the mean R-squared values of the top four machine learning algorithms and three ensemble methods using five-fold cross-validation using the dataset from India.

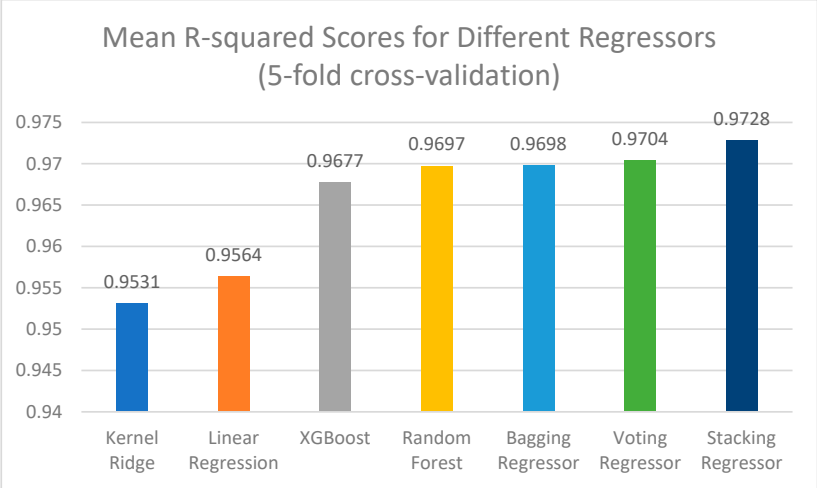

Figure 8. Comparison of R-squared scores of Machine Learning algorithms and Ensemble methods using five-fold cross-validation.

The figure shows that the stacking regressor ensemble method produced the best results of the seven methods with a mean R-squared value of 0.9728.

3.4. Five-Fold Cross-Validation Trained on 2015–2019 Datasets from Boca Raton Using Tilted Solar Panel Output

Since the models were validated with a particular dataset, we applied five-fold cross-validation algorithms for the same four machine learning algorithms and three ensemble methods on a different dataset from a local city, Boca Raton, FL, USA. To reflect a real-world application, the solar irradiance output variable for this dataset was changed from measurements from a horizontal solar panel to the Earth's surface to a tilted angle. The name of this output parameter is All Sky Surface Shortwave Downward Direct Normal Irradiance.

In practice, tilted solar panels are deployed to face the sun and increase solar irradiance. This is demonstrated by Figure 9, which compares the actual solar irradiance captured by both tilted solar panels and horizontal panels in New York City.

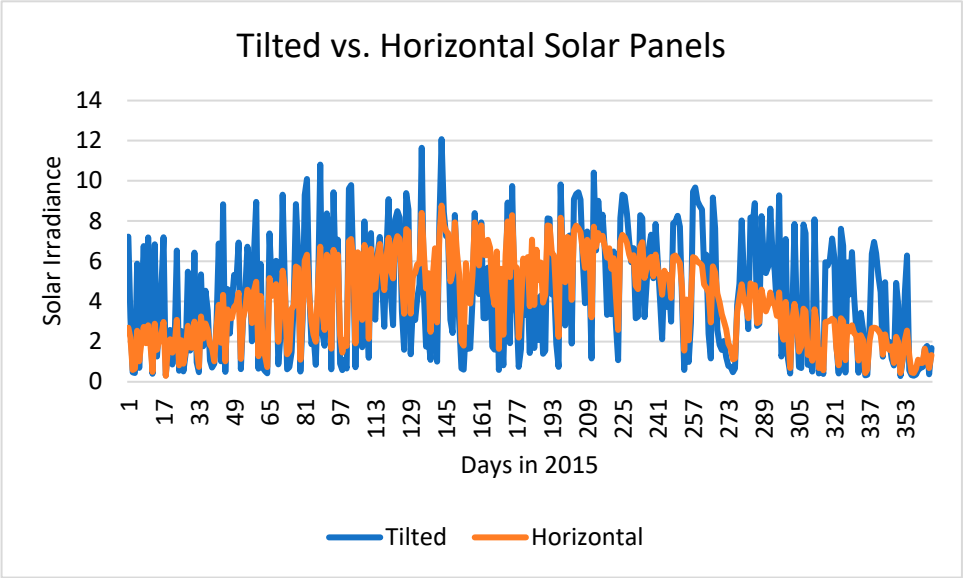

Figure 9. Comparison of solar irradiance received by tilted and horizontally aligned solar panels in New York City, USA, in the year 2015.

During the year 2015, New York City received 7% more solar irradiance from tilted panels than from a horizontal installation.

Table 9 presents the comparison of the mean R-squared values of the prediction results for each algorithm using tilted solar panels for Boca Raton, FL.

Table 9. R-Squared Values for Methods With Five-Fold Cross-Validation on The Boca Raton, Fl Dataset.

Machine Learning Method	Mean R-Squared
Kernal Ridge	0.8600
Linear Regression	0.8639
XGBoost	0.9165
Random Forest	0.9165
Ensemble Method	**Mean R-squared**
Bagging Regressor	0.9206
Voting Regressor	0.9065
Stacking Regressor	0.9190

The table shows that the bagging regressor ensemble method produced the best results of the seven methods with a mean R-squared value of 0.9206. Our next experiment was to generalize the model built from the Boca Raton dataset. We tested the model using the Orlando dataset with tilted solar panels for the year 2020. To generalize the model, we tested it on different locations in Florida: Orlando, Miami, Tampa, Jacksonville, and Tallahassee.

Figure 10 presents the comparison of the mean R-squared values of the prediction results for each algorithm for the Orlando dataset.

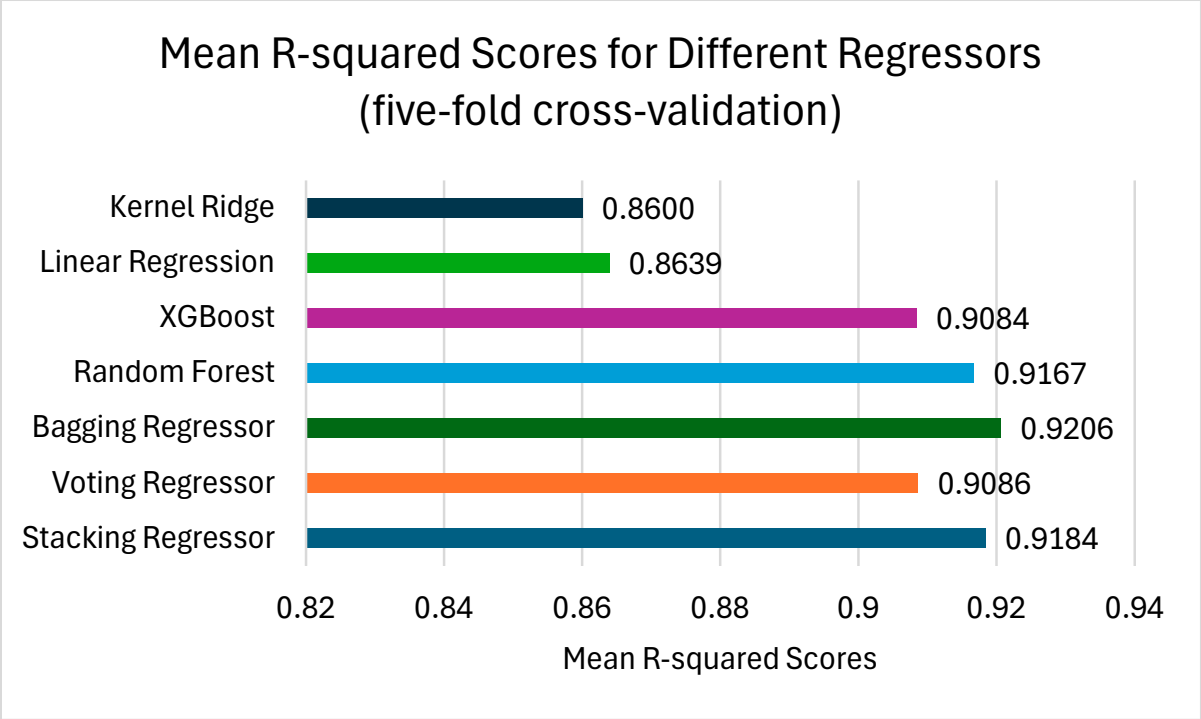

Figure 10. Comparison of R-squared scores of Machine Learning algorithms and Ensemble methods using five-fold cross-validation tested on Orlando, FL, dataset from 2020.

The figure shows that the bagging regressor ensemble method produced the best results of the seven methods with a mean R-squared value of 0.9206. Then, we expanded our experiment to other cities in the state of Florida using solar irradiance data from tilted solar panels in 2019. Our goal is to increase the generalizability of the model. Ridge regression was incorporated into all machine learning methods using the datasets from Boca Raton, Florida, as the training data for our model to address the critical issue of overfitting. In addition, five-fold cross-validation was performed on the dataset for each city. By penalizing large coefficients, ridge regression introduces a regularization term to the loss function, which helps to balance model complexity and generalizability. The use of ridge regression aligns with the primary goal of this research: to develop predictive models that generalize well to unseen data. By integrating ridge regression into all methods, this study ensures that the models are not only accurate but also robust and interpretable. This choice demonstrates a commitment to producing models that can withstand the complexities of real-world data while avoiding the pitfalls of overfitting. The prediction results for the city of Orlando using the Random Forest machine learning model are presented in Figure 11. Random Forest produced the best results.

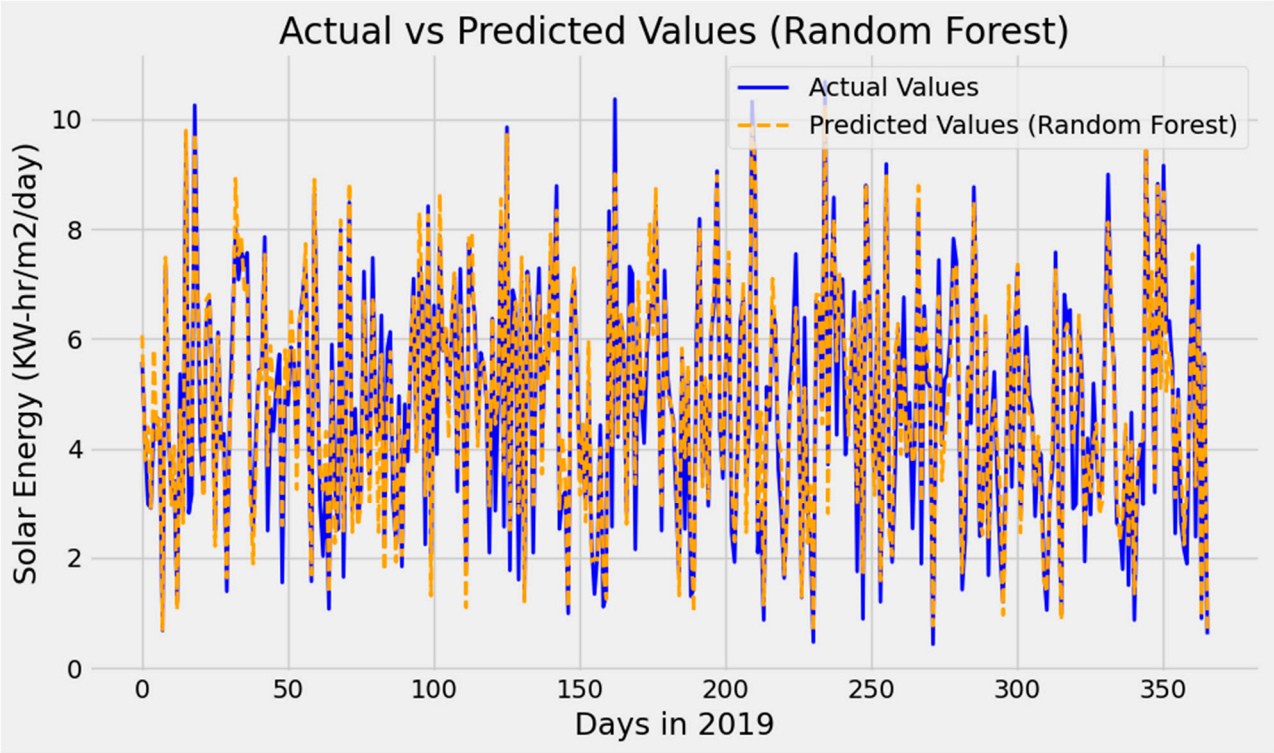

Figure 11. Prediction Results for Random Forest Model (Orlando).

The Random Forest algorithm produced an R-squared value of 0.9245. The prediction results for the city of Miami using the XGBoost machine learning model are presented in Figure 12.

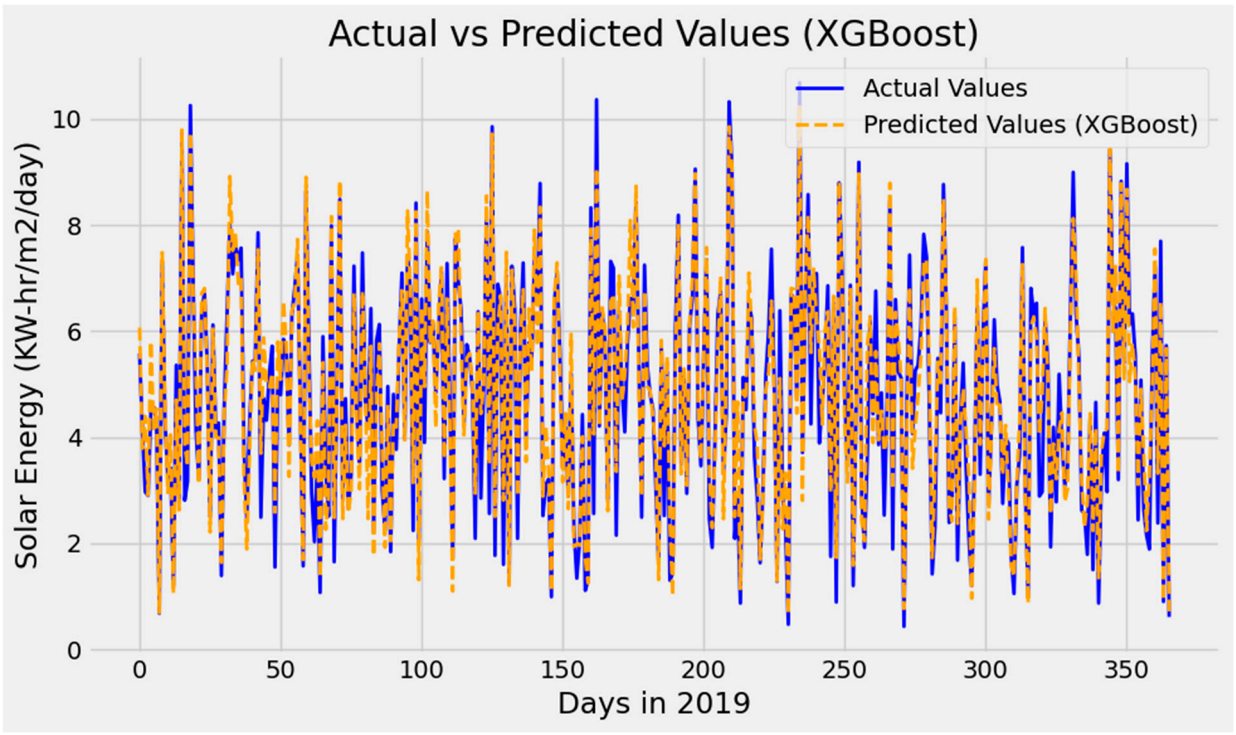

Figure 12. Prediction Results for XGBoost Model (Miami).

The XGBoost algorithm produced an R-squared value of 0.9242. The prediction results for the city of Tampa using the Random Forest machine learning model are presented in Figure 13.

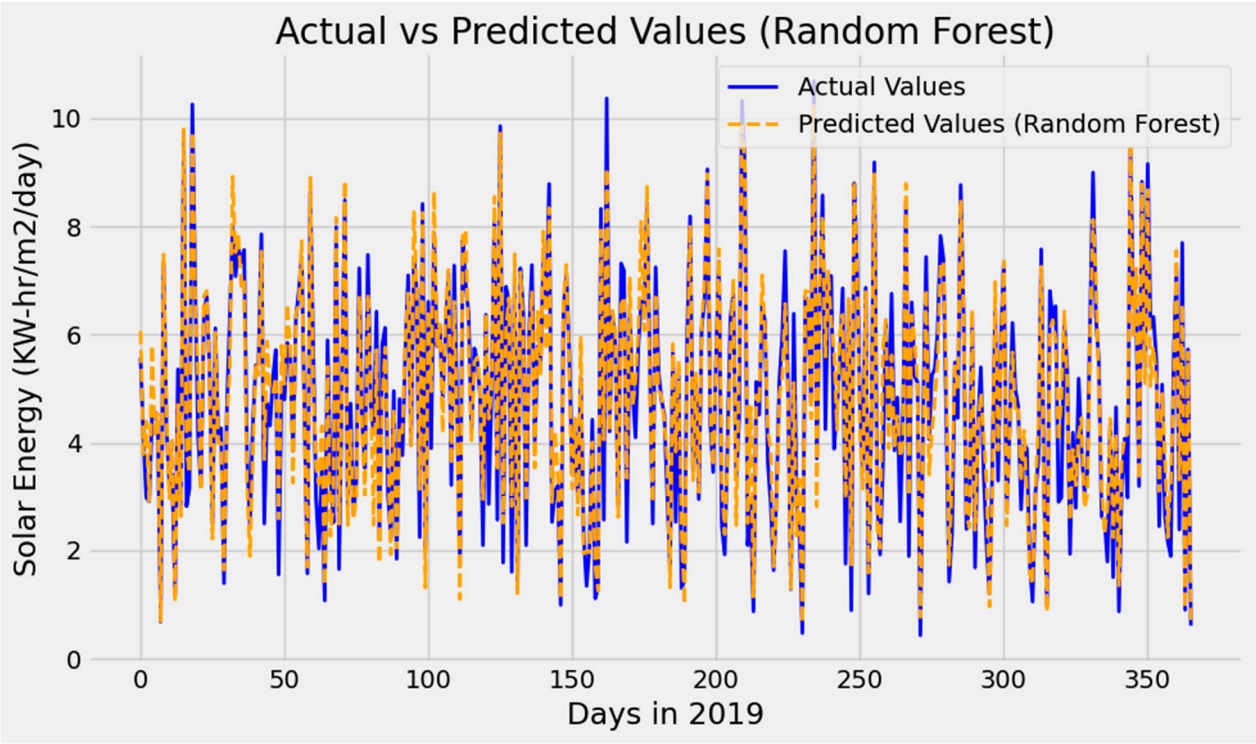

Figure 13. Prediction Results for Random Forest Model (Tampa).

The Random Forest algorithm produced an R-squared value of 0.9246. Next, we selected two major cities in Northern Florida to assess the prediction results of a region far from the other cities tested. The prediction results for the city of Jacksonville using the Random Forest machine learning model are presented in Figure 14.

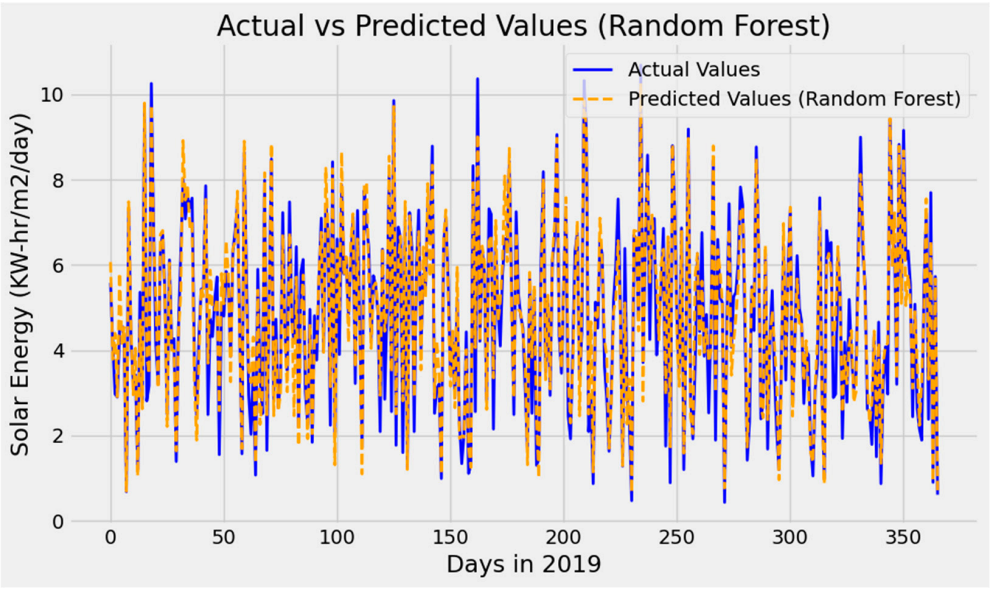

Figure 14. Prediction Results for Random Forest Model (Jacksonville).

The Random Forest algorithm produced an R-squared value of 0.9243. The prediction results for the city of Tallahassee using the Random Forest machine learning model are presented in Figure 15.

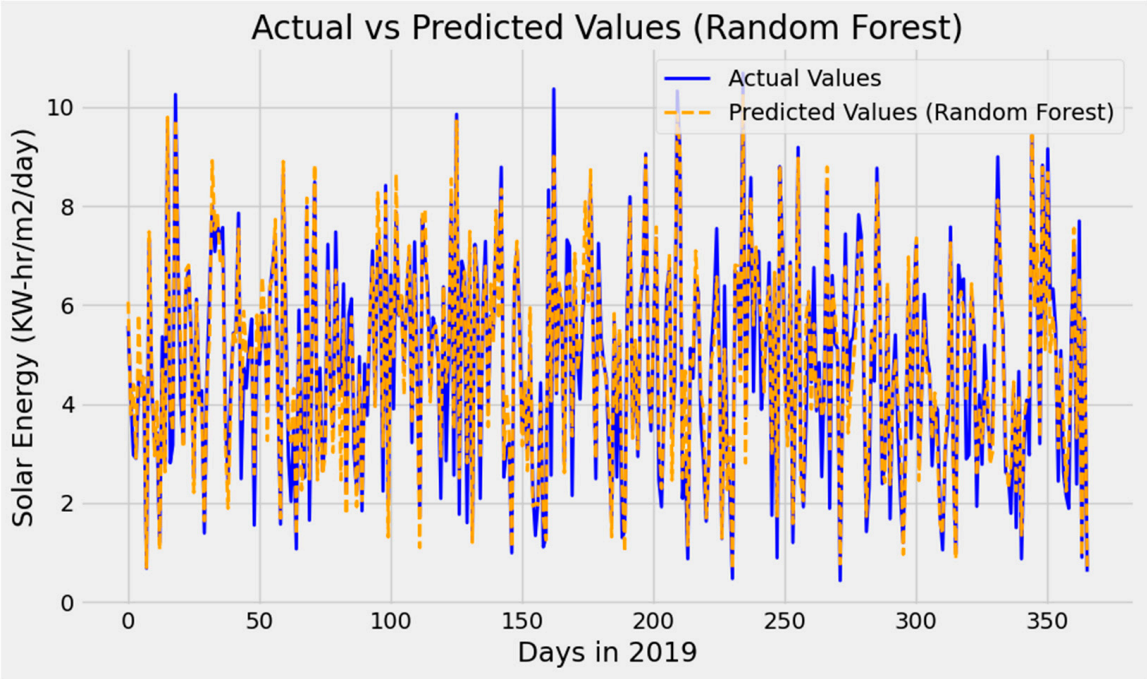

Figure 15. Prediction Results for Random Forest Model (Tallahassee).

The Random Forest algorithm produced an R-squared value of 0.9249.

Figure 16 presents the bagging ensemble regressor model using a scatter plot with a regression line that represents a match between the forecasted and actual values for Orlando. The stacking regressor performed the best out of the three ensemble learning methods with an R-squared value of 0.9339.

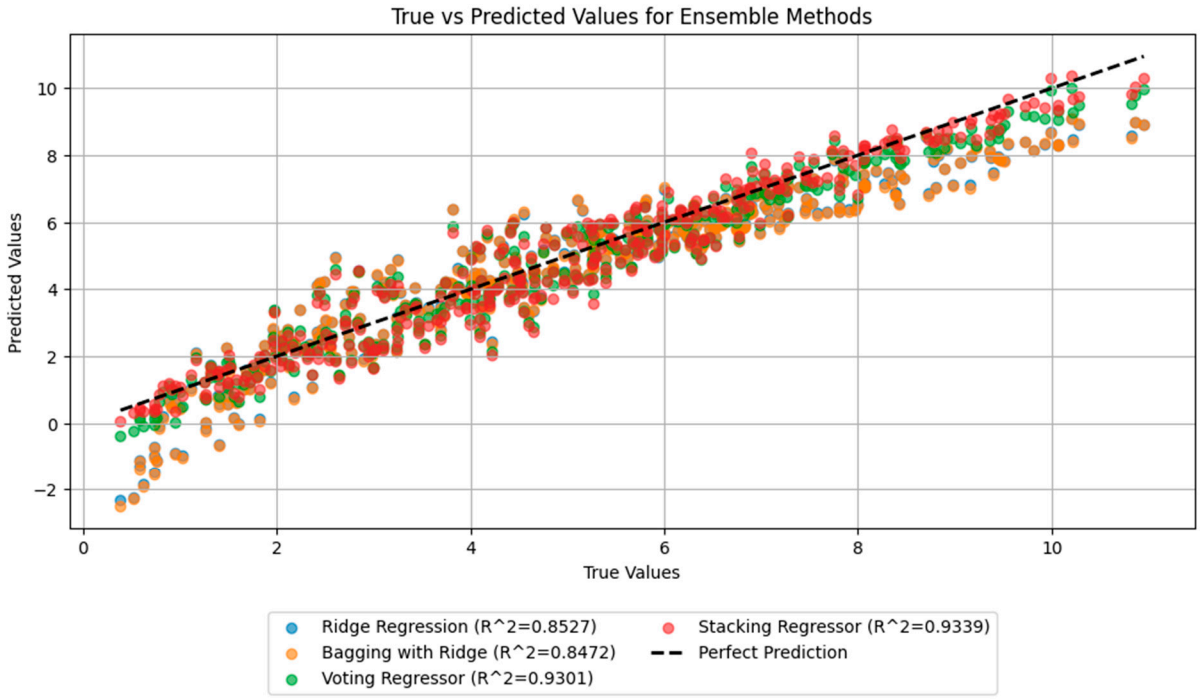

Figure 16. Scatterplot of True vs. Predicted Values for Ensemble Methods (Orlando).

Figure 17 presents the bagging ensemble regressor model using a scatter plot for Miami. The stacking regressor performed the best out of the three ensemble learning methods with an R-squared value of 0.9263.

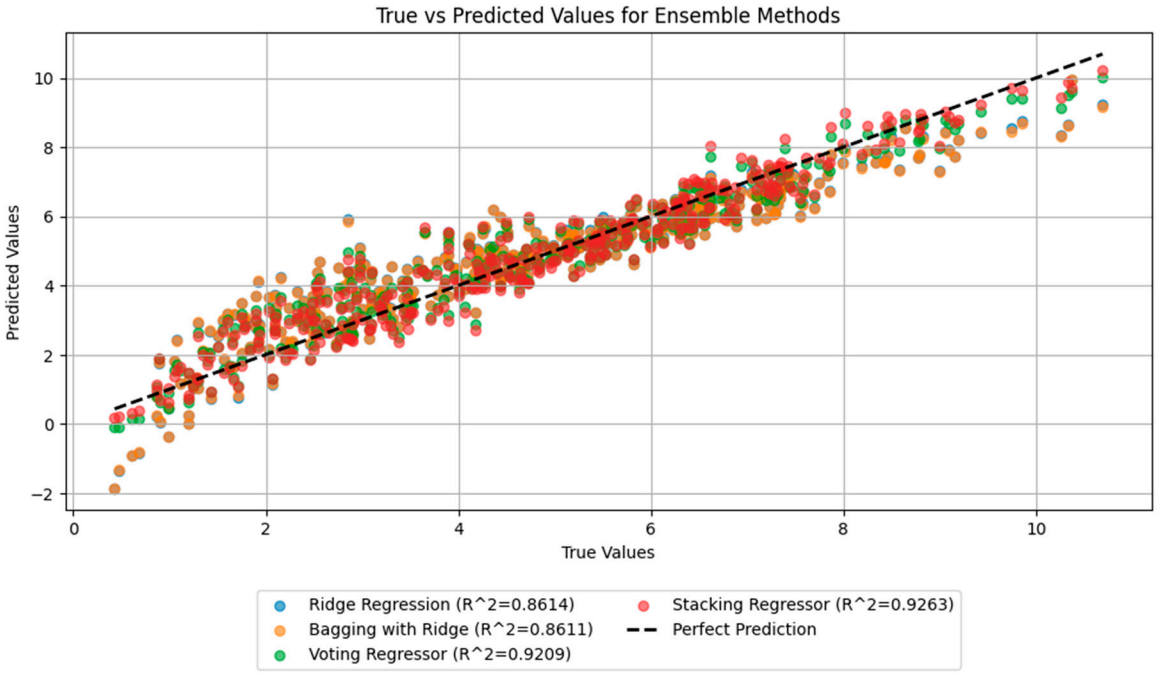

Figure 17. Scatterplot of True vs. Predicted Values for Ensemble Methods (Miami).

Figure 18 presents the bagging ensemble regressor model using a scatter plot for Tampa. The stacking regressor performed the best out of the three ensemble learning methods with an R-squared value of 0.9571.

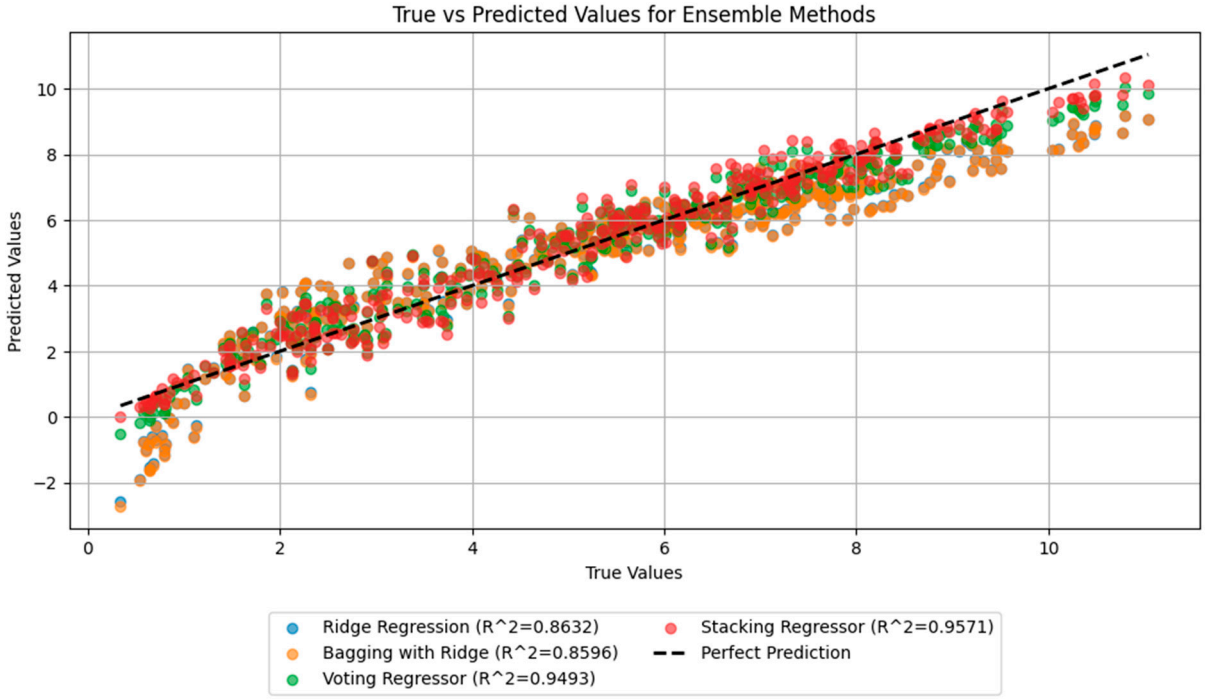

Figure 18. Scatterplot of True vs. Predicted Values for Ensemble Methods (Tampa).

Figure 19 presents the stacking ensemble regressor model using a scatter plot for Jacksonville. The stacking regressor performed the best out of the three ensemble learning methods with an R-squared value of 0.9403.

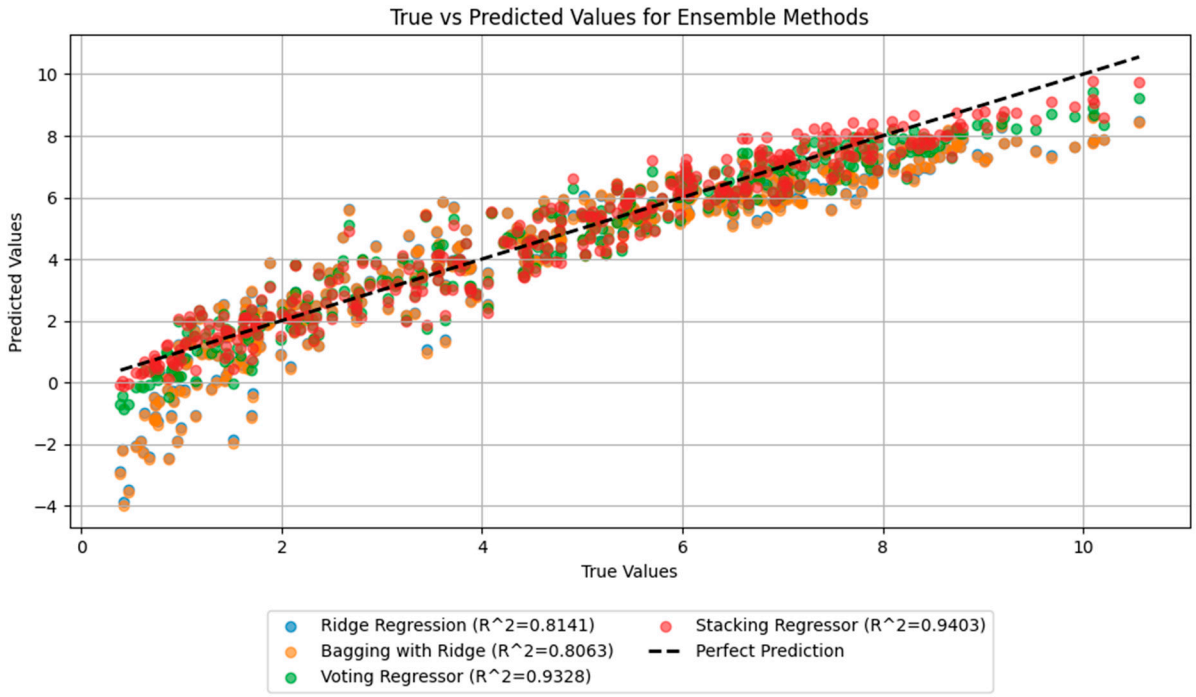

Figure 19. Scatterplot of True vs. Predicted Values for Ensemble Methods (Jacksonville).

Figure 20 presents the stacking ensemble regressor model using a scatter plot for Tallahassee. The stacking regressor performed the best out of the three ensemble learning methods with an R-squared value of 0.9425.

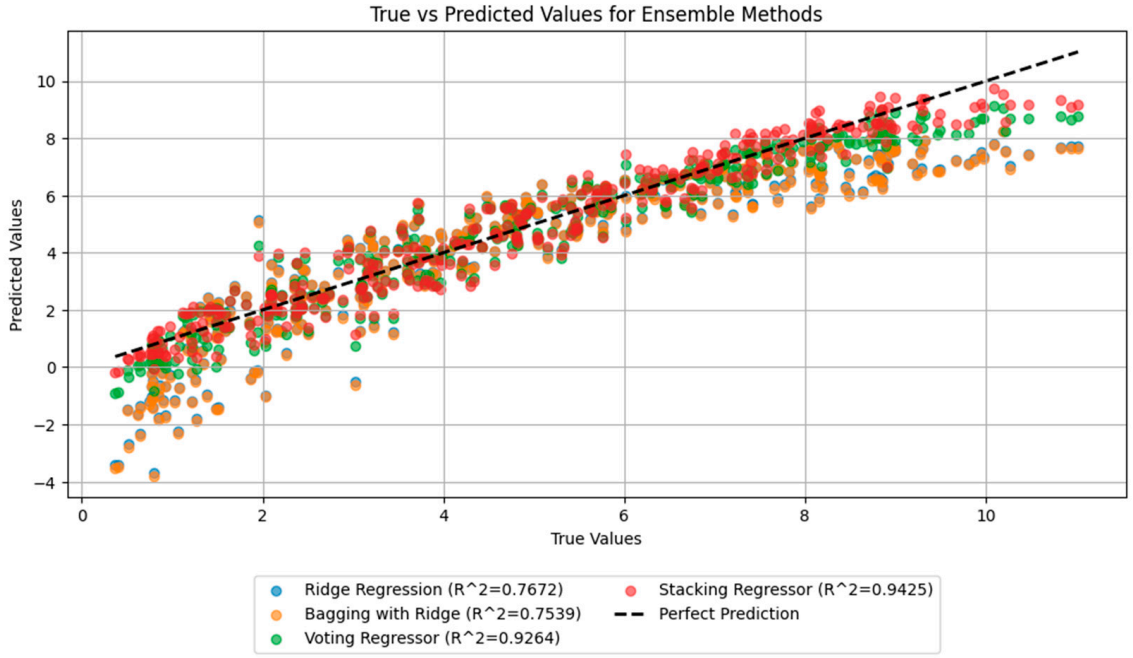

Figure 20. Scatterplot of True vs. Predicted Values for Ensemble Methods (Tallahassee).

Table 10 provides a concise comparison of key performance metrics for the machine learning models and ensemble methods across the tested cities: Orlando, Miami, Tampa,

Jacksonville, and Tallahassee. This table highlights the performance of Random Forest and Stacking Regressor as the best-performing methods in most cities, with a key performance metric (R-squared) value ranging from 0.92 to 0.96.

Table 10. R-squared values for the tested cities.

City	Best Machine Learning Model	R-Squared (ML)	Best Ensemble Method	R-Squared (Ensemble)
Orlando	Random Forest	0.9245	Stacking Regressor	0.9339
Miami	XGBoost	0.9242	Stacking Regressor	0.9263
Tampa	Random Forest	0.9246	Stacking Regressor	0.9571
Jacksonville	Random Forest	0.9243	Stacking Regressor	0.9403
Tallahassee	Random Forest	0.9249	Stacking Regressor	0.9425

Overestimation of solar irradiance could lead to an overly aggressive reliance on solar power, potentially depleting the battery prematurely, while underestimation could result in missed opportunities to optimize battery usage during peak solar periods.

In our work, the predicted solar irradiance serves as a critical input to our proposed fuzzy logic algorithm specifically designed for battery operation optimization. The fuzzy logic algorithm processes the solar irradiance predictions along with other key variables, such as the state of charge (SOC) and electricity price, to make real-time decisions on charging, discharging, and energy sourcing.

4. Discussion

The global reliance on fossil fuels as a primary energy source has led to substantial environmental concerns, including pollution and climate change. A shift toward renewable energy sources, particularly solar and wind, has gained momentum; however, these alternatives present challenges due to their inherent intermittency and installation costs. In this study, five machine learning algorithms—Random Forest, XGBR, SVR, KR, and LR—were trained and evaluated using data from the NASA POWER satellite database. By integrating machine learning with satellite-derived weather data, our approach provides a novel and efficient method for optimizing energy management within microgrids, facilitating smarter and more responsive operations in renewable energy systems. This methodology underscores the potential of remote satellite sensing to advance sustainability in energy generation.

The novelty of this study lies in our use of a combination of machine learning and ensemble models to enhance prediction accuracy through cross-validation and testing on different datasets. The model still performed well in five different cities in Florida using solar irradiance data from tilted solar panels. Additionally, we successfully applied machine learning models to generalize solar irradiance predictions across diverse geographical locations in Florida, achieving high accuracy with R-squared values between 0.90 and 0.9249. Importantly, our ensemble methods, including voting, stacking, and bagging, produced good results, with R-squared values exceeding 0.9262, outperforming previous studies on the same dataset. This demonstrates the robustness and efficiency of our approach.

For the machine learning models, Random Forest emerged as the best-performing individual algorithm in four out of the five cities, with improvements averaging 10% for R-squared, 21% for MSE, 18% for RMSE, and 16% for MAE compared to other algorithms. This success can be attributed to its ensemble structure, which effectively reduces variance and prevents overfitting. XGBoost, another ensemble method, showed competitive performance, though its computational complexity was higher compared to Random Forest.

Scalability and computational efficiency are critical factors when deploying machine learning models in BESSs. In our study, we utilized a range of machine learning and ensemble methods to balance predictive accuracy and computational efficiency. Methods like Random Forest and XGBRegressor, while computationally intensive during training, operate efficiently during inference, which aligns with the needs of IoT-based microgrids.

The ensemble methods we used further enhanced performance, with the stacking regressor achieving the highest accuracy in all scenarios. These approaches integrate multiple models to improve accuracy, and their scalability can be tailored by adjusting the number and complexity of base learners.

Research in [25] underscores solar irradiance as the primary factor in determining PV panel output. We plan to develop a battery optimization algorithm [26] that integrates solar irradiance predictions to optimize battery operations effectively. Future research can explore hybrid approaches that combine machine learning and deep learning models, which may uncover additional opportunities to enhance prediction accuracy while maintaining computational efficiency.

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

Funding: This research received no external funding.

Data Availability Statement: The dataset we used is from Bhopal, India and can be accessed through the following https://power.larc.nasa.gov/data-access-viewer (accessed on 9 June 2024).

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

References

1. Colucci, R.; Mahgoub, I. Solar Irradiance Prediction with Ensemble Learning Method as Input for Battery Operation Optimization. In Proceedings of the IEEE Texas Power and Energy Conference, College Station, TX, USA, 12–13 February 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
2. Gupta, R.A.; Kumar, R.; Bansal, A.K. BBO-based small autonomous hybrid power system optimization incorporating wind speed and solar radiation forecasting. *Renew. Sustain. Energy Rev.* **2015**, *41*, 1366–1375. [CrossRef]
3. Maleki, A.; Khajeh, M.G.; Rosen, M.A. Weather forecasting for optimization of a hybrid solar-wind–powered reverse osmosis water desalination system using a novel optimizer approach. *Energy* **2016**, *114*, 1120–1134. [CrossRef]
4. Mirzapour, F.; Lakzaei, M.; Varamini, G.; Teimourian, M.; Ghadimi, N. A new prediction model of battery and wind-solar output in hybrid power system. *J. Ambient. Intell. Humaniz. Comput.* **2019**, *10*, 77–87. [CrossRef]
5. Gheouany, S.; Ouadi, H.; Giri, F.; El Bakali, S. Experimental validation of multi-stage optimal energy management for a smart microgrid system under forecasting uncertainties. *Energy Convers. Manag.* **2023**, *291*, 117309. [CrossRef]
6. Al-Ja'Afreh, M.A.A.; Amjad, B.; Rowe, K.; Mokryani, G.; Marquez, J.L.A. Optimal planning and forecasting of active distribution networks using a multi-stage deep learning based technique. *Energy Rep.* **2023**, *10*, 686–705. [CrossRef]
7. Vijay, M.; Saravanan, M. Solar Irradiance Forecasting Using Bayesian Optimization Based Machine Learning Algorithm to Determine the Optimal Size of a Residential PV System. In Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 7–9 April 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 744–749.
8. Bisoi, R.; Dash, D.R.; Dash, P.K.; Tripathy, L. An efficient robust optimized functional link broad learning system for solar irradiance prediction. *Appl. Energy* **2022**, *319*, 119277. [CrossRef]
9. Kallio, S.; Siroux, M. Photovoltaic power prediction for solar micro-grid optimal control. *Energy Rep.* **2023**, *9*, 594–601. [CrossRef]
10. Kumari, P.; Toshniwal, D. Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. *J. Clean. Prod.* **2021**, *279*, 123285. [CrossRef]
11. Singla, P.; Duhan, M.; Saroha, S. An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network. *Earth Sci. Inform.* **2022**, *15*, 291–306. [CrossRef]

12. Molu, R.J.J.; Tripathi, B.; Mbasso, W.F.; Naoussi, S.R.D.; Bajaj, M.; Wira, P.; Blazek, V.; Prokop, L.; Misak, S. Advancing short-term solar irradiance forecasting accuracy through a hybrid deep learning approach with Bayesian optimization. *Results Eng.* **2024**, *23*, 102461. [CrossRef]
13. Tercha, W.; Tadjer, S.A.; Chekired, F.; Canale, L. Machine Learning-Based Forecasting of Temperature and Solar Irradiance for Photovoltaic Systems. *Energies* **2024**, *17*, 1124. [CrossRef]
14. Allal, Z.; Noura, H.N.; Chahine, K. Machine Learning Algorithms for Solar Irradiance Prediction: A Recent Comparative Study. *e-Prime-Adv. Electr. Eng. Electron. Energy* **2024**, *7*, 100453. [CrossRef]
15. Brahma, B.; Wadhvani, R. Solar irradiance forecasting based on deep learning methodologies and multi-site data. *Symmetry* **2020**, *12*, 1830. [CrossRef]
16. Srivastava, R.; Tiwari, A.N.; Giri, V.K. Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. *Heliyon* **2019**, *5*, e02692. [CrossRef]
17. Cros, S.; Badosa, J.; Szantaï, A.; Haeffelin, M. Reliability predictors for solar irradiance satellite-based forecast. *Energies* **2020**, *13*, 5566. [CrossRef]
18. Javed, A.; Kasi, B.K.; Khan, F.A. Predicting solar irradiance using machine learning techniques. In Proceedings of the 15th international wireless communications & mobile computing conference (IWCMC), Tangier, Morocco, 24–28 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1458–1462.
19. Chiteka, K.; Enweremadu, C.C. Prediction of global horizontal solar irradiance in Zimbabwe using artificial neural networks. *J. Clean. Prod.* **2016**, *135*, 701–711. [CrossRef]
20. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582. [CrossRef]
21. Wu, Y.K.; Phan, Q.T.; Zhong, Y.J. Overview of Day-ahead Solar Power Forecasts Based on Weather Classifications and a Case Study in Taiwan. *IEEE Trans. Ind. Appl.* **2024**, *60*, 1409–1423. [CrossRef]
22. Boutahir, M.K.; Farhaoui, Y.; Azrour, M.; Zeroual, I.; El Allaoui, A. Effect of feature selection on the prediction of direct normal irradiance. *Big Data Min. Anal.* **2022**, *5*, 309–317. [CrossRef]
23. Blanc, P.; Espinar, B.; Geuder, N.; Gueymard, C.; Meyer, R.; Pitz-Paal, R.; Reinhardt, B.; Renné, D.; Sengupta, M.; Wald, L.; et al. Direct normal irradiance related definitions and applications: The circumsolar issue. *Sol. Energy* **2014**, *110*, 561–577. [CrossRef]
24. Law, E.W.; Prasad, A.A.; Kay, M.; Taylor, R.A. Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting—A review. *Sol. Energy* **2014**, *108*, 287–307. [CrossRef]
25. Al-Bashir, A.; Al-Dweri, M.; Al-Ghandoor, A.; Hammad, B.; Al-Kouz, W. Analysis of effects of solar irradiance, cell temperature and wind speed on photovoltaic systems performance. *Int. J. Energy Econ. Policy* **2019**, *10*, 353–359. [CrossRef]
26. Colucci, R.; Mahgoub, I.; Yousefizadeh, H.; Al-Najada, H. Survey of strategies to optimize battery operation to minimize the electricity cost in a microgrid with renewable energy sources and electric vehicles. *IEEE Access* **2024**, *12*, 8246–8261. [CrossRef]