



Article

Forecasting Photovoltaic Performance: A Comparative Assessment of Machine Learning Methods

Ayman Mdallal

Department of Civil Engineering, McMaster University, 1280 Main St. W., Hamilton, ON L8S 4L8, Canada; mdallala@mcmaster.ca

How To Cite: Mdallal, A. Forecasting Photovoltaic Performance: A Comparative Assessment of Machine Learning Methods. *Renewable and Sustainable Energy Technology* 2026, 2(1), 4. <https://doi.org/10.53941/rset.2026.100001>

Received: 7 December 2025

Revised: 4 February 2026

Accepted: 12 February 2026

Published: 2 March 2026

Abstract: The increasing global need for renewable energy sources has identified photovoltaic systems as essential to clean energy transition initiatives. This study examines the predictive reliability of machine learning models in assessing and forecasting the output and thermal performance of a medium-scale photovoltaic facility, employing both monofacial and bifacial modules in Canada. A simulation model was created utilizing the System Advisor Model (SAM) with five years of weather data to produce hourly outputs, including power and photovoltaic cell temperature. The datasets were analyzed utilizing multiple regression-based machine learning algorithms, such as Linear Regression, Polynomial Regression, Decision Tree, Random Forest, XGBoost, and regularization methods. Key performance indicators, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2), were assessed to compare model accuracy. Results demonstrate substantial enhancements in prediction accuracy through the utilization of ensemble models such as Random Forest and XGBoost, attaining R^2 values over 97% for both power output and cell temperature forecasts. Bifacial systems exhibited superior energy generation efficiency and tolerance to thermal fluctuations relative to monofacial systems, owing to their dual-sided light capture capability. Analysis of feature importance indicated that Global Horizontal Irradiance (GHI), temperature, and wind speed were the primary determinants affecting performance.

Keywords: solar energy; photovoltaics; machine learning; performance prediction

1. Introduction

Solar energy is a key component in combating climate change by reducing our reliance on fossil fuels, thereby decreasing carbon dioxide and other detrimental emissions, and mitigating global warming and its associated impacts. Moreover, solar energy technologies, such as photovoltaic (PV) systems and concentrated solar power (CSP), are continuously advancing, improving efficiency and cost-effectiveness, thereby increasing accessibility and affordability for various applications, ranging from residential to industrial uses [1]. The decreasing costs of solar technology made it a feasible alternative to traditional energy sources, hence accelerating global adoption. Besides its environmental benefits, solar energy significantly contributes to economic development by creating diverse employment opportunities in manufacturing, installation, maintenance, and research and development, particularly in regions with high unemployment rates, thus fostering economic growth and stability [2]. The decentralization of energy production through solar installations enhances energy security and reduces dependence on imported fuels, especially beneficial for countries with limited natural resources [1]. Solar energy fosters social development by improving electricity access in rural and remote regions, as off-grid solar solutions provide reliable and economical energy to communities far away from primary power grids, thereby improving living standards



Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

and enabling access to modern amenities and services. This is particularly vital in developing countries, where energy poverty remains an important barrier to social and economic advancement [2,3].

Monofacial and bifacial modules are the two primary types of PV technology, each exhibiting different features and applications that facilitate the advancement of solar energy. Figure 1 illustrates the performance difference [4]. Monofacial PV modules are the standard type of solar panels, designed to capture sunlight from one side and often oriented towards the sun to maximize direct solar irradiation. These panels are widely employed due to their developed technology, ease of installation, and lower initial costs, making them effective in multiple applications, from residential rooftops to large solar farms. However, their efficiency is constrained by the amount of sunlight that reaches their surface, which could be a disadvantage in areas limited by high reflectivity or diffuse light. Bifacial module design employs innovative materials and configurations, including transparent backsheets or dual glass designs that provide light penetration from the rear, hence enhancing energy capture and reinforcing structural integrity and durability [5,6]. Bifacial modules can be integrated with tracking sensors that adjust the panel's orientation to align with the sun, thereby enhancing energy production [6]. Bifacial modules, despite their initial higher cost and slightly more complicated installation, offer long-term benefits through increased energy production and potentially lower levelized cost of electricity (LCOE), which makes them useful in large-scale solar projects where maximizing energy yield per unit area is significant.

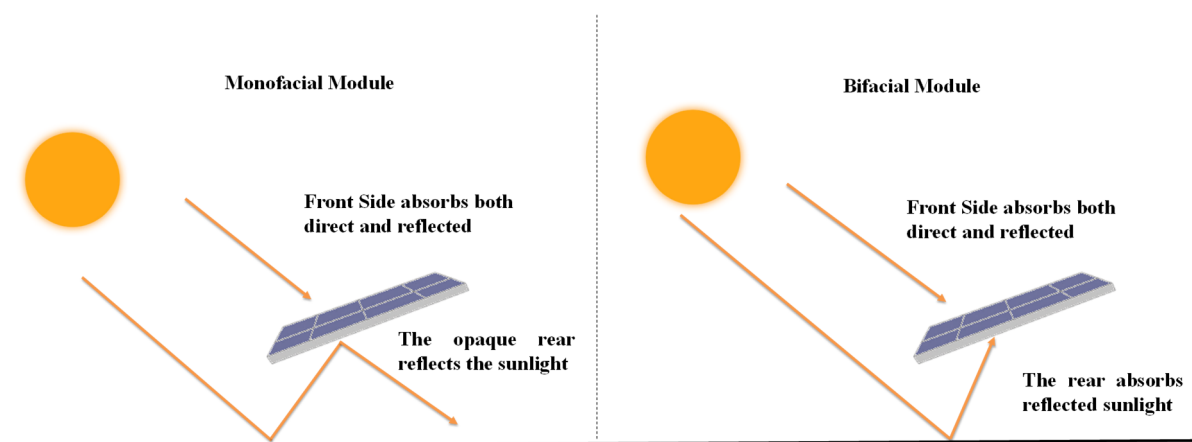


Figure 1. The variation in functional concepts between monofacial and bifacial modules.

Solar panels function most well when the sun passes directly on them. It cannot operate at all at night or when it's cloudy or rainy. Given that solar PV systems have to store energy to keep the power supply stable, batteries are needed. However, current battery technologies are expensive and have limited storage capacities, which raises the overall costs of solar PV systems [7,8]. Also, the costs of keeping the system running and fixing it if it fails could make the cost greater [9]. This means that it requires a lot of space to produce a lot of power, which can be a problem in cities or other crowded places. Solar power generation can be unstable and fluctuating, which can make it hard to connect to the electrical grid. This is especially true in areas where there is extremely high or low solar power. To fix these problems, it is essential to have a grid infrastructure that is more expensive and complicated [7,8]. The temperature has a big impact on how well PV cells work. For instance, silicon-based cells have a standard temperature coefficient of -0.4% to -0.5% per degree Celsius. This means that higher temperatures make them less effective because they cause more electron-hole recombination. Also, high temperatures lower the open-circuit voltage (V_{oc}) while the short-circuit current (I_{sc}) goes up a little bit. This means that less power is produced in total [10,11].

Machine Learning (ML) may bring a revolution within the energy sector as it enhances the performance of the energy sector in regard to the aspects of efficiency, reliability, and the ability to withstand changes. In energy management, machine learning algorithms process the enormous information available from the smart meter and sensor networks to predict the supply and demand for energy effectively. In this regard, machine learning technologies result in the effective distribution of energy in the sector, which leads to the reduction of losses and the attainment of a balance within the energy grid [12]. Furthermore, machine learning enhances the prediction of solar irradiance and wind speed; therefore, it enhances the integration of solar and wind energy within the energy sector [13], thus the aspect of grid integration, reliability, and the mitigation of greenhouse gases within the sector [14]. Predictive maintenance marks an essential aspect where machine learning algorithms process the information available from the sensor networks attached to the machines to recognize the patterns associated with the devices. In this regard, the aspect of machine learning within the sector ensures the early maintenance of the devices, thus

the reduction of the cost associated with the process while ensuring the devices are functioning at their optimal levels within the energy sector. Additionally, machine learning algorithms predict the failure of wind turbines based on the information from the temperature and vibration measurements within the sector, thus enabling the sector to undertake the process of repairing the turbines within the necessary time for the reduction of losses within the sector. A smart grid uses information within the sector to handle operations related to the balancing of loads, the diagnosis of the problems within the sector, and the response approach within the energy sector for effective functioning within the sector. In this regard, machine learning technologies form the crucial aspect of smart grids. Additionally, machine learning aspects within the sector enhance the flexibility and robustness of the energy sector via the control and adaptation of changes within the sector associated with changes in the power demand within the sector and the inherent distributed resources within the sector such as rooftop photovoltaic power and electric vehicles [12,13].

Machine learning empowered the photovoltaic power plant by increasing their predictability and optimizing the efficiency of the design and functioning process. The prediction of solar irradiance and power production needs to be very precise to manage the grid stability and the effective management of energy resources effectively. Machine learning algorithms such as the neural network and support machine rely on past meteorological data, images from satellites, and sensor inputs for the generation of very precise predictions for solar irradiance and power production levels [15]. Such uniqueness proves to be very effective for the grid management team to manage the supply-demand ratio effectively without the need for the backup power sources for power production purposes [16]. In addition to such aspects, machine learning proves to be very critical for the enhancement of both the design and functioning of the photovoltaic power plant. Such an aspect occurs through the analysis of massive data from the photovoltaic power plant installations to generate information based on the pattern analysis of the data for the generation of better designs and strategies for power plant installations [17].

Classical approaches, such as single-diode models and empirical temperature-power correlations often fail to capture the nonlinear interactions among irradiance, temperature, wind speed, and module operating conditions, particularly under rapidly fluctuating weather. In contrast, recent comparative reviews show that machine learning (ML) models consistently outperform statistical baselines across multiple forecasting horizons. For example, Gaboitaolelwe et al. (2023) [18] reports that ensemble methods and deep learning architectures achieve substantially lower forecasting errors than linear and autoregressive models, especially under high-variability irradiance conditions. Similarly, Radzi et al. (2023) [17] demonstrate that deep learning models (e.g., LSTM, CNN-LSTM hybrids) outperform shallow ML models when large datasets are available, while tree-based ensembles remain competitive in data-limited scenarios. These findings are reinforced by Benitez and Singh (2025) [19] showing that ML models provide superior generalization across climates compared to physics-only models, particularly for short-term and day-ahead PV forecasting.

A second major development in the literature is the emergence of system-specific ML modeling, especially for bifacial PV systems, where rear-side irradiance, albedo, and ground-reflected components introduce nonlinearities that traditional models cannot represent. Recent studies on bifacial systems show that ML-based models significantly improve prediction accuracy by learning rear-side gain behavior directly from operational data. For instance, recent work on real-time bifacial PV power prediction demonstrates that ML models outperform conventional irradiance-based analytical models under complex shading and albedo variability [20]. Likewise, ML-based thermal models for monofacial and bifacial modules show that bifacial systems require distinct feature sets, including rear-side irradiance proxies and albedo inputs and that models trained on monofacial data do not generalize well to bifacial configurations [21].

In a study [15] the performance of five machine learning algorithms; linear Regression, Polynomial Regression, Random Forest Algorithm, Neural Network, and the combination of both Random Forest (RF) Algorithm and Neural Network (NN), both without and with Hyperparameter Optimization, was evaluated. The RMSE of 9.21×10^{-7} was achieved by the RF algorithm, indicating very high precision, whereas the Hyperparameters improved the effectiveness of the NN algorithm substantially.

In another study, the attention was on the short-term interval forecasting of PV power, where a hybrid approach was used, combining the fuzzy information granulation technique with the CNN-BiGRU framework [22]. The CNN component of the hybrid approach was used for local information retrieval, while the bidirectional gated recurrent unit component (BiGRU component), on the other hand, was used for the analysis of the pattern features in the data, which aided in the exact prediction of the short-term intervals of the PV power.

A comparative study [23] compared the performance of three machine learning models, namely Linear Regression (LR), RF, and Gradient Boosting (GB), for short-term forecasting of PV energy production. From the results, it is clear the GB method performs better than LR and RF in terms of RMSE (1380.13), R-squared (0.8),

and MAPE (4.3%). Therefore, the GB technique is the most effective machine learning technique for short-term forecasting of PV energy production at the Melaka solar power plant.

A major contribution analyzed the deep learning models for the effective handling of time series data in the forecast of solar PV power generation, summarizing existing models and establishing new ones such as SVM, GRU, FFNN, and LSTM. Regions are evaluated based on the variability of the data, reflecting the effect of weather on power generation. The contribution proposed a hybrid approach of SVM and GRU in parallel, optimized by the ACO algorithm to enhance the forecasting results. The hybrid approach proposed the benefits of both SVM and GRU by leveraging their results together, and the ACO algorithm optimized the combination parameters for better results. The proposed approach achieved a correlation coefficient of 0.9986, reflecting the high forecast precision achieved [24].

Another research [25] used transfer learning and a deep neural network for the prediction of the power output of the PV plant one day in advance. Transfer learning technique improved the power prediction for newly installed photovoltaic power stations by leveraging the knowledge obtained from the previously installed photovoltaic power stations. The adjusted long short-term memory approach greatly improved the power prediction results, and the mean square error and weighted mean absolute percentage error values decreased from 0.55 to 0.168 and from 47.07% to 32.04%, respectively.

Recent advances in PV temporal modelling emphasize two complementary directions; generative, sequence-level simulation for long-horizon planning—exemplified by the Weather Diffusion Transformer (Weather-DiT), which couples a diffusion-Transformer architecture with trend/seasonality decomposition and a frequency probability joint loss to synthesize realistic, diverse 8760-h weather and PV scenarios and outperform GAN/probabilistic baselines and high-fidelity, multi-site ultra-short-term forecasting, typified by the PV-AFGNN framework that jointly applies bias-corrected NWP (NWP-CorrNet) and an Adaptive Fourier Graph Neural Network to capture cross-site spatiotemporal dependencies efficiently in the frequency domain [26,27].

In this research, a comparative analysis among various machine learning algorithms such as Linear Regression, Polynomial Regression, Support Vector Machines, Decision Tree, Random Forest, XGBoost, Lasso Regression, Elastic Net, and Ridge Regression are carried out to predict the power output and temperature of a medium-scale solar plant in Canada for both monofacial and bifacial solar technologies in order to enhance the forecasting processes for the management of the power production, thus minimizing the effects associated with the intermittency and unpredictability of the power production process. Additionally, the research will enhance the proactive approach during heatwave conditions to avoid the possible damage associated with the solar cells.

2. Methods

Several software tools exist that help estimate the performance of a PV system, some of which include PVWatts, PVsyst, System Advisor Model (SAM), HelioScope, PV*SOL, among others. This work will use System Advisor Model (SAM) to validate the extent of correspondence between the software estimate values of measured performance data.

After this validation process, a simulation model of a medium-scale PV plant having a power output of 1 MW will be developed in the software SAM. For this simulation model, the weather data will be obtained from the National Solar Radiation Database (NSRDB), which is developed by the National Renewable Energy Laboratory (NREL). This data will cover a time period of five years (2018–2022), with the location set to McMaster University in Canada.

After running the simulation in SAM, hourly data output, including power production and the temperature of the PV cells, will be exported into a CSV file. This dataset will then be processed using Jupyter Notebook to apply machine learning techniques for predictive modeling. The goal is to forecast PV power output (y_1) and PV cell temperatures (y_2) based on weather-related parameters. The selected independent variables for the predictive models include:

- Ambient Temperature (x_1)
- Global Horizontal Irradiance (GHI) (x_2)
- Relative Humidity (x_3)
- Atmospheric Pressure (x_4)
- Surface Albedo (x_5)
- Wind Speed (x_6)

Machine learning algorithms such as Linear Regression, Polynomial Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, Lasso Regression, Elastic Net, and Ridge Regression will be employed to construct predictive models. The study will analyze the performance of two types of PV modules (monofacial

and bifacial) to determine their efficiency and reliability under varying conditions. Figure 2 shows the steps followed in this work.

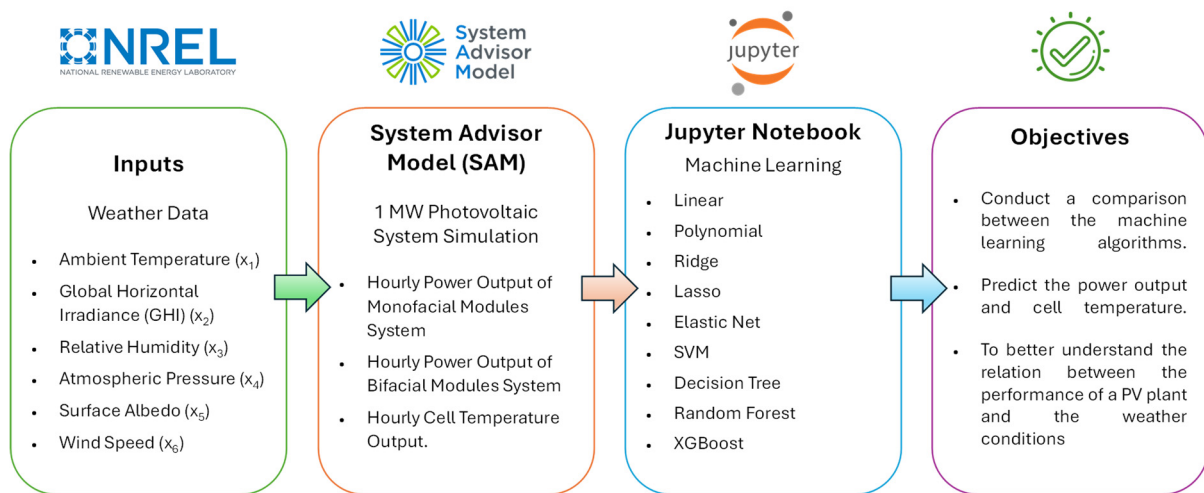


Figure 2. Method steps for this work.

Four key performance metrics will be used to evaluate the accuracy and efficacy of the predictive models:

1. Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

2. Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

3. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

4. Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

where n is the number of data points, y_i is the actual value, \hat{y}_i is the predicted value and \bar{y} is the mean of actual values.

2.1. Validation

To assess the accuracy of the software's estimates in simulating PV system performance, real measured power output data from a 9.6 kWp PV system installed in Egaleo, Greece will be utilized. This system is a roof-mounted configuration employing SunPower E19 320 EU PV modules paired with a Sunny Boy 3300-11 inverter with an efficiency rating of 94.50%. The system has a 30° tilt angle (slope) and a 0° azimuth angle (true south).

Weather data required for the validation will be obtained from the National Solar Radiation Database (NSRDB) for the specified location. The purpose of this analysis is to ensure that the power produced is compared to the predictions that this software makes in terms of performance.

A summary of the results of validation is presented in Table 1, which shows a comprehensive comparison between the system performance results obtained in simulation and in practice.

Table 1. The difference between real measured data and SAM results.

Location	Egaleo	Simulation	Relative Error (%)
	Measurements (Real Data) [28]	SAM Results	$\left(\frac{\text{Simulated} - \text{Measured}}{\text{Measured}}\right) \times 100\%$
System Size: 9.6 kWp	kWh/kWp	kWh/kWp	
January	81	79.06	-2.40
February	84.2	83.63	-0.67
March	136.7	138.08	1.01
April	152.5	145.80	-4.39
May	163.2	158.80	-2.69
June	173	171.84	-0.67
July	174.7	175.94	0.71
August	172.8	172.86	0.04
September	151.3	151.93	0.42
October	120.2	120.39	0.16
November	82.9	83.62	0.86
December	70	69.84	-0.23
Annual	1562.4	1549	-0.86

The data in Table 1 shows that the SAM is effective at predicting how a PV system performs. The validation shows that SAM's predictions are very close to the real measured data, with a maximum monthly relative error of 4.4% and an annual relative error of 0.86%. These results show that SAM is an appropriate tool for modeling PV systems in the real world.

2.2. PV System Specifications

Weather data from 2018 to 2022 will be used for the area around McMaster University in Canada (43.2614° N, 79.9197° W) to simulate a 1 MWp PV system. The simulation will use SunPower SPR-P19-400 (Mono-c-Si) modules, which each have a power output of 400 Wp, for both monofacial and bifacial systems. The system will have a tilt angle of 43.25° and an azimuth angle oriented towards the true south.

For the bifacial simulations, a bifaciality factor of 0.85 will be applied, with the modules installed at a ground clearance height of 1.5 m. The inverter used for the system will be a Canadian Solar Inc. CSI-50KTL-GS, with an efficiency of 98.5%.

Table 2 summarizes the average weather data values for the five-year period used in the simulation. Figure 3 illustrates a sample of the simulated power output and PV cell temperature results over a one-year period for the system.

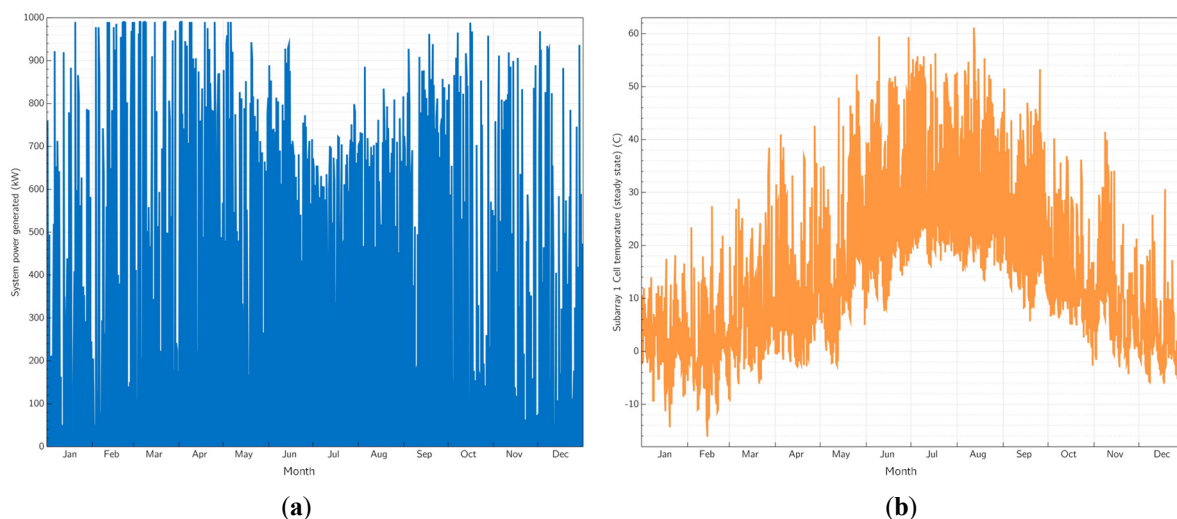
**Figure 3.** The hourly results of the bifacial system for year 2022; (a) power output, (b) cell's temperature.

Table 2. Average weather data for 5 years (2018–2022).

Parameter	Average Value
Temperature (°C)	9.53
GHI (W/m ²), excluding nighttime	312.4
Relative Humidity (%)	75.4
Pressure (mbar)	996.3
Surface Albedo	0.257
Wind Speed (m/s)	3.48

2.3. Regression Models

In this study, several machine learning regression techniques predict the performance of the PV systems, focusing on bifacial and monofacial. To provide a robust and interpretable comparison, it was chosen to include a variety of models that, taken collectively, cover the range of: (i) simple parametric baselines, (ii) regularized linear techniques for handling multicollinearity and solving the problem of implicit feature selection, (iii) non-linear regression using a kernel, and finally (iv) tree-based ensemble learning techniques, which have been observed to efficiently capture even the most complex non-linear patterns and relationships in the data. The selected models to be included in the comparison include: Ordinary Least Squares linear and polynomial regression, Ridge regression, Lasso regression, and Elastic Net Regularization, which reduce overfitting and multicollinearity, Support Vector Regression, a non-linear regression using a kernel that is robust to noise and outliers, Decision Trees, a non-parametric regression technique for modeling threshold, and finally, the Random Forest and XGBoost implementation of the tree-based ensemble learning techniques, which have been known to efficiently capture the complex patterns and relationships in the data. These techniques have been selected because they have been observed to be efficiently utilized for modeling the complexities of the PV market in all relevant studies.

These models will be implemented in Jupyter Notebook, and their performance will be evaluated by comparing their prediction accuracy and error metrics, such as MSE, MAE, RMSE and R².

2.3.1. Linear Regression

Linear regression is a supervised learning algorithm used to predict the value of a continuous dependent variable (Y) based on one or more independent variables (X) [29]. In its simplest form, for a single independent variable (simple linear regression), the relationship can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad (5)$$

For multiple variables (multiple linear regression), the equation generalizes to:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon \quad (6)$$

Y is the dependent variable (output or target), and X is the independent variable (input or feature). The line's intercept is β_0 , and the slope, β_1 , shows how significantly Y changes when X changes by one unit. Also, ϵ is the error term, which takes into account noise or changes that the model doesn't explain [29]. The objective function is:

$$\text{Minimize: } \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \quad (7)$$

2.3.2. Polynomial Regression

Polynomial regression is an extension of linear regression that models the relationship between the independent variable (X) and the dependent variable (Y) as a polynomial. While linear regression assumes a straight-line relationship, polynomial regression captures non-linear relationships, making it suitable for datasets where trends follow a curve [30].

Polynomial regression represents the relationship between the dependent and independent variables as an n -degree polynomial. The general equation for polynomial regression is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_n X^n + \epsilon \quad (8)$$

Polynomial regression minimizes the sum of squared residuals, similar to linear regression:

$$\text{Minimize: } \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_n X_i^n))^2 \quad (9)$$

For multiple independent variables (X_1, X_2, \dots, X_p), the polynomial regression equation of degree d can be expressed as:

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i + \sum_{i=1}^p \sum_{j=1}^p \beta_{ij} X_i X_j + \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p \beta_{ijk} X_i X_j X_k + \dots + \epsilon \quad (10)$$

2.3.3. Ridge Regression

Ridge regression is a regularized version of linear regression that helps prevent overfitting by adding a penalty term to the cost function. The penalty discourages the model from fitting excessively large coefficients, which can lead to overfitting, especially when the model has many features or when the features are highly correlated [31].

Ridge regression modifies the standard linear regression by adding an L_2 regularization term to the cost function. The general equation for ridge regression is:

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i \quad (11)$$

However, the key difference is in the optimization of the coefficients. In ridge regression, the cost function is:

$$\text{Minimize: } \sum_{i=1}^p (Y_i - \hat{Y}_i)^2 + \alpha \sum_{i=1}^p (\beta_i)^2 \quad (12)$$

Y_i represents the observed value of the dependent variable for the i -th sample, while \hat{Y}_i is the predicted value for the i -th sample. The coefficients of the model are denoted by $\beta_0, \beta_1, \dots, \beta_p$, and α is the regularization parameter, or penalty term. X_i refers to the input feature for the i -th sample, and the sum of squared coefficients, $\sum_{i=1}^p (\beta_i)^2$, serves as the penalty term that regularizes the model.

The term $\alpha \sum_{i=1}^p (\beta_i)^2$ penalizes the model for large coefficients, encouraging the model to shrink the coefficients toward zero but not exactly to zero (unlike Lasso regression, which uses L_1 regularization).

When $\alpha = 0$, ridge regression reduces to regular linear regression without any penalty term. As α becomes large, the penalty term dominates, causing the coefficients to shrink toward zero, which reduces the impact of each feature and may lead to underfitting if α is too large. On the other hand, when α is small, the regularization effect weakens, potentially causing the model to overfit by learning large coefficients for features, especially in cases of multicollinearity. Thus, α plays a crucial role in balancing the bias-variance tradeoff: a large α results in high bias and low variance (underfitting), while a small α leads to low bias and high variance (overfitting).

By shrinking coefficients, ridge regression reduces model complexity and prevents overfitting, especially when the dataset has many features or noise.

2.3.4. Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso), is a regularized form of linear regression that applies L_1 regularization to the cost function. Unlike ridge regression (which uses L_2 regularization), lasso regression encourages sparsity in the model by shrinking some coefficients exactly to zero, which effectively performs feature selection [32].

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i \quad (13)$$

However, the cost function in lasso regression becomes:

$$\text{Minimize: } \sum_{i=1}^p (Y_i - \hat{Y}_i)^2 + \alpha \sum_{i=1}^p |\beta_i| \quad (14)$$

where $\sum_{i=1}^p |\beta_i|$ is the sum of the absolute values of the coefficients (penalty term). The penalty term encourages the coefficients β_i to shrink toward zero. The key feature of lasso is that some coefficients can become exactly zero, resulting in a simpler, more interpretable model.

Lasso encourages sparsity, meaning that it tends to shrink some coefficients exactly to zero. This effectively selects a subset of the most important features and discards the irrelevant ones, making it useful for high-dimensional data.

2.3.5. Elastic Net

Elastic Net regression is a type of regularized linear regression that combines the penalties of both ridge regression (L_2) and lasso regression (L_1). It also addresses some of the issues in the LASSO, especially in those instances where the variables involved are correlated. Elastic Net is particularly advantageous in instances where there are numerous features that are inter-correlated since this approach can effectively combine the benefits by both L_1 and L_2 regularization [33].

The cost function in Elastic Net regression is changed to include both lasso and ridge penalties:

$$\text{Minimize: } \sum_{i=1}^p (Y_i - \hat{Y}_i)^2 + \alpha_1 \sum_{i=1}^p |\beta_i| + \alpha_2 \sum_{i=1}^p (\beta_i)^2 \quad (15)$$

where α_1 is the weight of the L_1 penalty (lasso) and α_2 is the weight of the L_2 penalty (ridge).

Elastic Net is effective at addressing multicollinearity since it uses the ridge penalty to spread shrinkage across predictors that are highly correlated to each other. It also uses the lasso penalty, which allows it to choose features by making some coefficients exactly zero. Elastic Net makes the model less complex and helps keep it from overfitting by regularizing the coefficients. It also combines the optimal aspects of lasso and ridge, balancing the sparsity of lasso with the stability of ridge. This solves problems that each method has when used independently [33].

2.3.6. Support Vector Machine (Regression)

Support Vector Machine Regression (SVR) is a variant of the Support Vector Machine (SVM) algorithm that is made for use in regression problems. SVR is different from other regression methods because it tries to predict a continuous target variable while keeping the error within a certain range, called the epsilon ϵ -tube [34]. It is a reliable algorithm that works well when the data has complicated relationships or isn't linear.

The goal of SVR is to find a function $f(X)$ that gets as close as possible to the target values Y while keeping a margin of error (ϵ) around the predictions. The general equation of a linear regression function in SVR is:

$$f(X) = \omega^T X + b \quad (16)$$

X represents the input features, ω is the weight vector that defines the orientation of the regression hyperplane, and b is the bias term that adjusts the offset of the hyperplane.

The optimization in SVR seeks to minimize the complexity of the model (measured by the magnitude of ω) while ensuring the predictions fall within a tolerance level ϵ from the true values. This is expressed as:

$$\text{Minimize: } \frac{1}{2} \|\omega\|^2 \quad (17)$$

Subject to:

$$|Y_i - f(X_i)| \leq \epsilon \quad \forall i \quad (18)$$

However, when some predictions fall outside the ϵ -tube, slack variables ξ_i^+ and ξ_i^- are introduced to penalize these deviations [34]. The optimization problem becomes:

$$\text{Minimize: } \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i^+ + \xi_i^-) \quad (19)$$

Subject to:

$$\begin{aligned} |Y_i - f(X_i)| &\leq \epsilon + \xi_i^+ \\ |Y_i - f(X_i)| &\leq \epsilon + \xi_i^- \\ \xi_i^+, \xi_i^- &\geq 0 \end{aligned} \quad (20)$$

C is the regularization parameter that controls the trade-off between model complexity (smoothness) and tolerance for deviations, while ϵ specifies the width of the margin of tolerance around the predicted values.

2.3.7. Decision Tree

Decision Tree Regression is a non-parametric supervised learning algorithm used for predicting continuous target values. Unlike linear models that assume a specific relationship between features and target variables, decision trees divide the feature space into distinct regions based on decision rules. This allows them to model complex, non-linear relationships [35].

A decision tree is a hierarchical model that splits the data into subsets based on feature thresholds. Each split corresponds to a condition (e.g., $X_i > t$), and the process continues recursively, forming a tree structure. The leaf nodes of the tree represent the predicted value, which is typically the mean of the target values in that region.

The regression tree predicts a target value Y using a piecewise constant approximation. The model is expressed as:

$$\hat{Y}_i = \sum_{j=1}^M c_j I(X \in R_j) \quad (21)$$

where M represents the number of terminal (leaf) nodes, and R_j denotes the region corresponding to the j -th leaf node. The constant c_j is the predicted value for observations in R_j . The indicator function $I(X \in R_j)$ equals 1 if X belongs to region R_j and 0 otherwise.

The purpose of a regression tree is to divide the feature space into regions R_1, R_2, \dots, R_M so that the sum of squared errors (SSE) in each region is as low as possible. The cost function for a given split is:

$$\text{Minimize: } \sum_{j=1}^M \sum_{i \in R_j} (Y_i - \hat{Y}_j)^2 \quad (22)$$

Here, Y_i is the observed target value for the i -th sample, and \hat{Y}_j is the predicted value, calculated as the mean of Y_i for region R_j . The algorithm picks split that lower the SSE as much as possible. This is called variance reduction.

Initialization is the first step in the process. In this step, the whole dataset is treated as the root node, and the best split is found by using a feature and threshold that lowers the SSE. During splitting, the dataset is recursively divided into two child nodes at each step, based on the chosen feature and threshold, until a stopping criterion such as maximum depth, minimum samples per leaf, or minimum SSE improvement is reached. For prediction, an input is passed through the tree from the root to a leaf node following the decision rules, and the predicted value is the mean of the target values within the corresponding leaf node.

2.3.8. Random Forest

Random Forest Regression builds multiple independent decision trees during training and aggregates their predictions to make the final output. The predicted value is typically the average of the predictions from all individual trees [36].

The regression function can be expressed as:

$$\hat{Y}_t = \frac{1}{T} \sum_{t=1}^T f_t(X) \quad (23)$$

Here, T represents the number of trees in the forest, and $f_t(X)$ denotes the prediction from the t -th tree for the input X . The final predicted value, \hat{Y}_t , is the mean of the predictions from all the trees in the forest.

The optimization goal for each decision tree in the random forest is to minimize the SSE within its regions, just like a single decision tree. However, the final output aggregates predictions across multiple trees, reducing the variance of the model. The aggregated cost function is not directly minimized but emerges as a result of individual tree optimizations.

2.3.9. Extreme Gradient Boosting (XGBoost)

XGBoost regression builds an ensemble of weak learners (decision trees) to predict continuous target variables, optimizing predictions iteratively by minimizing errors. XGBoost builds an additive model where predictions are made by combining outputs of multiple decision trees [37]. Each tree is trained to correct errors from the previous trees. The prediction for an input X is given by:

$$\hat{Y} = \sum_{t=1}^T f_t(X) \quad (24)$$

In the context of an ensemble model, T represents the number of trees in the ensemble, while $f_t(X)$ denotes the prediction made by the t -th tree for a given input X . The last predicted value, Y , is the total of all the trees' predictions in the ensemble. The training process is meant to find the best parameters for each $f_t(X)$ tree so that it can make accurate predictions. The total objective function is:

$$\text{Objective} = \sum_{i=1}^n L(Y_i, \hat{Y}_i) + \sum_{t=1}^T \Omega(f_t) \quad (25)$$

where L is the loss function (MSE):

$$L(Y_i, \hat{Y}_i) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (26)$$

Regularization Term (Ω) Penalizes the complexity of the model to prevent overfitting. Regularization is applied to each tree (f_t):

$$(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (27)$$

In this model, γ is a parameter controlling the number of leaf nodes in the underlying model and λ is a parameter that controls the size of the weights (w_j) allocated to the leaf nodes.

During training, XGBoost employs a gradient boosting framework in which first- and second-order derivatives (gradients and Hessians) of the loss function with respect to the predictions are computed at each iteration. These quantities guide tree construction by determining how predictions should be adjusted to minimize the objective. Candidate splits are evaluated using a gradient-based gain metric, which measures the reduction in the regularized loss achieved by splitting a node into left (L) and right (R) nodes:

$$\text{Gain} = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma \quad (28)$$

where G and H are the sums of gradients and Hessians of the loss function for the samples in each node. This formulation favors splits that yield the greatest reduction in loss while accounting for model regularization, thereby improving generalization when handling noisy meteorological inputs.

Predictions are updated at each boosting iteration according to:

$$\hat{Y}^{(t)} = \hat{Y}^{(t-1)} + \eta f_t(X) \quad (29)$$

where η is the learning rate that scales the impact of each tree.

XGBoost is well suited for photovoltaic power and cell-temperature prediction because PV system behavior exhibits regime-dependent and nonlinear characteristics that cannot be adequately represented by global linear models. While PV output is approximately linear with irradiance under moderate operating conditions, efficiency losses at high cell temperatures, interaction effects between irradiance and temperature, and additional dependencies on wind speed and albedo introduce nonlinear responses. Tree-based boosting models capture these localized nonlinearities by partitioning the feature space into regions corresponding to distinct operating regimes, enabling the model to represent efficient roll-off at high temperatures and complex interactions without explicit feature engineering.

3. Results and Discussion

The number of data points per hour used in the analysis is 43,800, calculated based on 5 years \times 8760 per year. Out of the data points, 70% was used in training the models, while 30% was set for testing. This training and testing split was consistently applied across all algorithms, with a random state value of 65 to ensure reproducibility. For the cell temperature models, all data points corresponding to nighttime or zero GHI during daytime were excluded to enhance the accuracy and reliability of the predictions. Since the PV cell temperature will be the same as the ambient temperature when power generation is zero.

An iterative grid-search method was used on the training dataset only to avoid data leakage to find the best hyperparameters for the ensemble models was. The model was trained on the training subset and tested using cross-validated error metrics for each possible set of hyperparameters. The main measure of performance was RMSE. The grid search methodically examined predetermined parameter ranges, preserving the configuration that reduced validation error while ensuring consistent generalization performance. This iterative evaluation process made sure that the model's complexity was adjusted progressively based on actual performance instead of fixed assumptions made before the search. The final hyperparameter sets listed above are the best-performing configurations found through this systematic search.

For Ridge Regression, the optimal α values were determined to be:

- 100 for the power output model
- 0.01 for the cell temperature model

In Lasso Regression, the optimal α values were:

- 21.87 for the monofacial power output model
- 23.44 for the bifacial power output model
- 1.29 for the cell temperature model

For Decision Tree Regression, a Grid Search was performed to identify the optimal hyperparameters, including maximum depth, minimum samples per leaf, and minimum samples per split. The results were:

- For power output models:
 - Maximum depth: 10
 - Minimum samples per leaf: 4
 - Minimum samples per split: 12
- For cell temperature models:
 - Maximum depth: 10
 - Minimum samples per leaf: 4
 - Minimum samples per split: 10

In Random Forest model, the number of estimators was set to be 100 for all cases. For XGBoost, all models used a Grid Search was performed to identify the optimal hyperparameters. It is found to be; Learning rate: 0.1, max depth: 7 and number of estimators: 200, while the sub sample varies.

These parameter values were selected to achieve the best predictive performance for both power output and cell temperature across all algorithms used.

3.1. Monofacial Models

Analysis of the performance of different regression models in accurately forecasting the output of monofacial PV systems exhibits certain patterns with regard to accuracy and error measurements. Linear, Ridge, and Elastic Net regression models demonstrate balanced performance with MAE ranging between 58.7–58.9, overall MSE ranging between 7796.1–7800.7, and RMSE of 88.3. These models result in an R^2 of 86.7%, defining their moderate performance and failure to deal with non-linearities in the data. Lasso regression models result in higher MAE and MSE with an R^2 of 83.3%, defining lower accuracy in dealing with the data.

Polynomial regression, especially in the second and third degrees, presents improvements in terms of accuracy. The 2nd degree polynomial regression improves the MAE to 34.1, RMSE to 59.2, with an R^2 of 94.0%, indicating improvements in accuracy due to fitting the data points to non-linear patterns. The 3rd degree polynomial regression increases accuracy even further with an MAE of 30.54, RMSE of 51.5, and R^2 of 95.5%, establishing the accuracy in identifying complex patterns in the data. There is, however, an increasing possibility of overfitting in polynomial models due to the complexity.

The models with the strongest performance in the current study include Decision Tree Regression, Random Forest, and XGBoost, which give very accurate results when their hyperparameters are properly optimized. The optimized Decision Tree model, with depth 10 and certain parameters in node split, results in an MAE of 20.9, RMSE of 46.6, and R^2 of 96.3%, which outperforms other regression models. The Random Forest model further enhances accuracy, with MAE of 17.7, RMSE of 39.5, and R^2 of 97.3%, showing the efficiency of the model in combining multiple decision trees. Table 3 shows results of all regression models for monofacial PV.

Table 3. Key performance indicators result for monofacial power output prediction models.

Regression Type	MAE	MSE	RMSE	R^2 (%)
Linear	58.9	7796.1	88.3	86.7
Polynomial 2nd Degree	34.1	3507.5	59.2	94.0
Polynomial 3rd Degree	30.54	2651.3	51.5	95.5
Ridge	58.7	7800.7	88.3	86.7
Lasso	56.5	9787.2	98.9	83.3
Elastic Net	58.8	7796.8	88.3	86.7
Support Vector	38.9	8192.9	90.5	86
Decision Tree	23.2	3019.1	54.9	94.8
Decision Tree (max depth: 10, min samples leaf: 4, min samples split: 12)	20.9	2175.4	46.6	96.3
Random Forest	17.7	1564.3	39.5	97.3
XGBoost	18.85	1565.98	39.57	97.33
XGBoost				
Learning rate: 0.1				
max depth: 7	17.585	1427.97	37.79	97.56
number of estimators: 200				
sub sample: 0.8				

Figure 4 representing monofacial power output, shows that RMSE is initially very high at low polynomial degrees, indicating that a simple linear model cannot capture the complexity of the data. As the polynomial complexity increases, RMSE decreases significantly for both training and test datasets, reaching its lowest value between degrees 4 and 6. This suggests that the model can effectively capture the underlying patterns of monofacial power output within this range. However, as the polynomial degree exceeds 6, the test RMSE begins to increase slightly while the training RMSE continues to decrease, indicating the onset of overfitting. At these higher degrees, the model starts to fit the noise in the training data rather than generalizing well to unseen data. Therefore, for monofacial power output prediction, a polynomial degree of around 5 is optimal, balancing accuracy and generalizability.

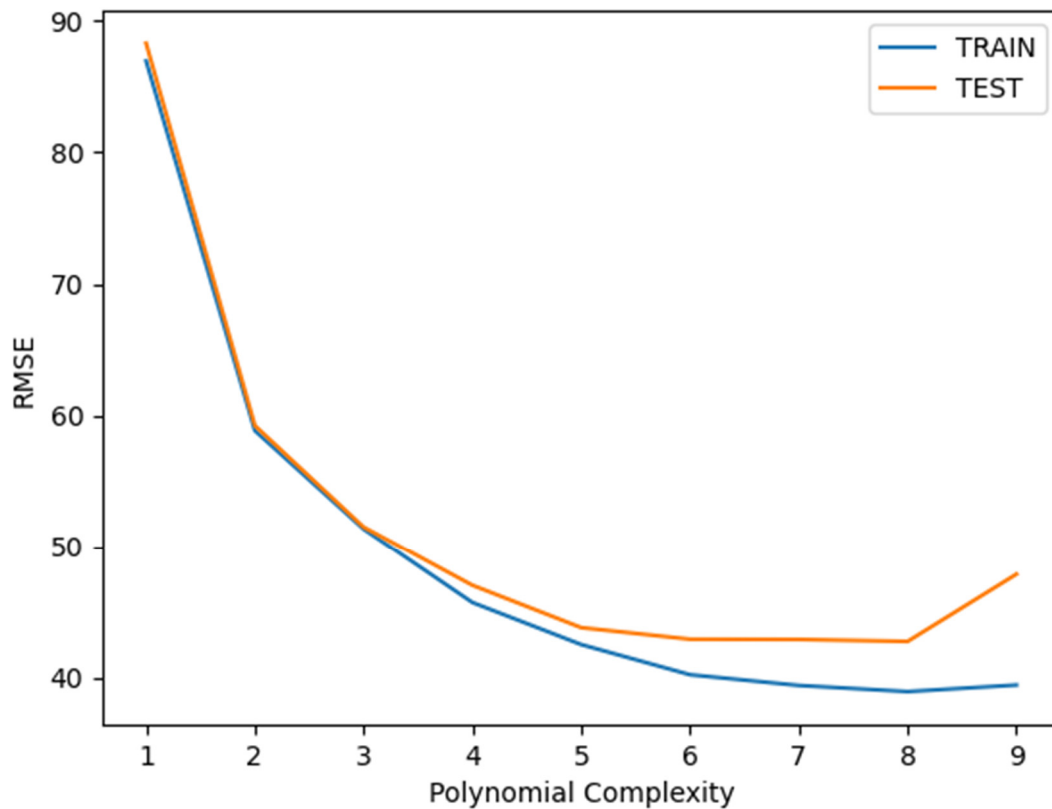


Figure 4. Monofacial power output RMSE vs. polynomial complexity.

3.2. Bifacial Models

The assessment of bifacial power output prediction models demonstrates a variance in their predictive performance. The linear regression, Ridge regression, and Elastic Net models provided similar performance levels within errors confined to the MAE of 62.1–62.3, an MSE of approximately 8777.8–8782.5, and an overall RMSE of 93.7. The R^2 values of stability was 86.7%, suggesting sufficient precision but limited capacity to capture the nonlinear dynamics of bifacial PV system behavior. The Lasso regression model exhibited slightly poorer results with an R^2 value of 83.4 along with comparably higher error metrics, leading to implications for rigid potential not accommodating any complex interactions in the data set.

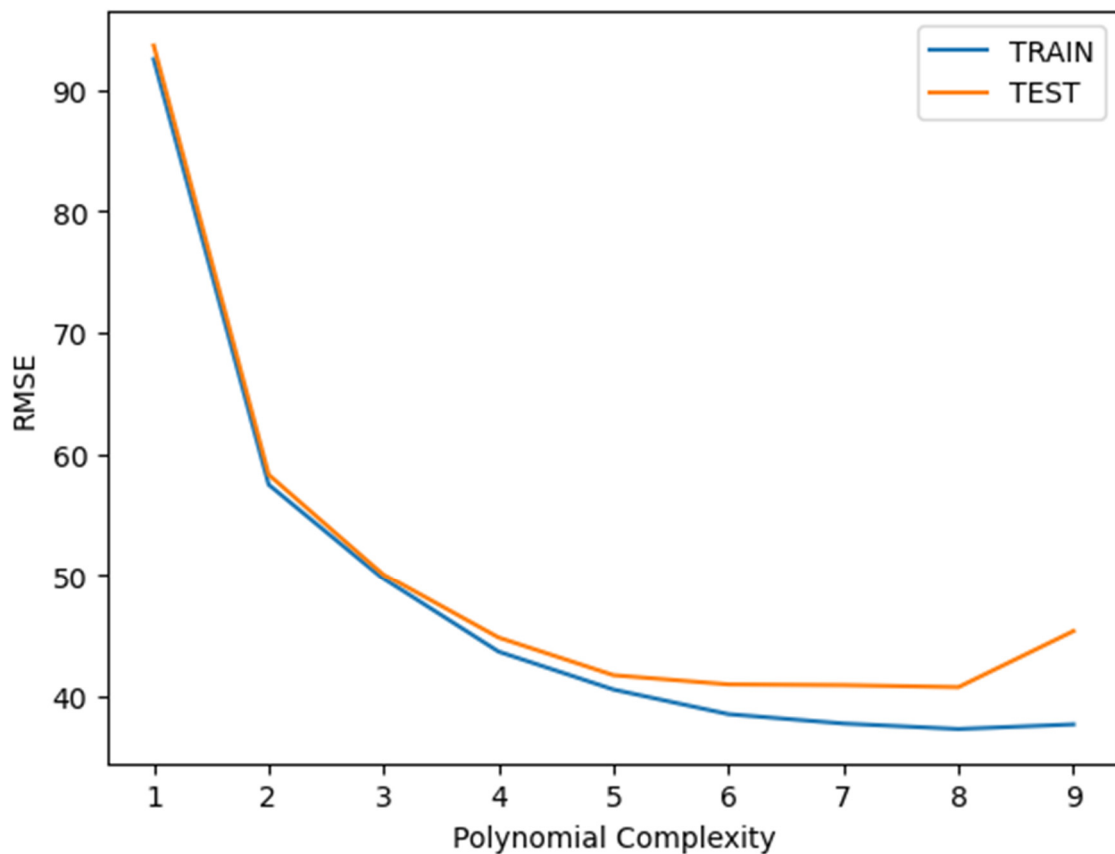
Polynomial regression models performed considerably better at predicting predictive accuracy in power output measures, and the magnitude of improvement significantly increased with a higher-order polynomial shape. The 2nd degree polynomial regression models yielded an MAE of 33 and an RMSE of 58.3, with an R^2 of 94.7%, proving adequate at modeling a nonlinear relationship. The 3rd degree polynomial regression model further reduced the MAE to 29.7 and the RMSE to 50.0 with an R^2 of 96.2%, qualifying it as one of the best-performing models tested. Also, it was recognized that increased polynomial complexity can result in overfitting that observe a lower predictive efficiency when tested on new data.

The best models are the optimized Decision Tree Regression, Random Forest and XGBoost models. When the Decision Tree is further optimized by using a maximum depth of 10 and the best splitting criterion, the MAE of 19.7, RMSE of 43.8, and R^2 of 97.1% indicates that this model is the best of the three, creating good predictions of the independent variable. With Random Forest also producing refined results, with an MAE of 16.8, RMSE of 37.3, and R^2 of 97.9% shows the effectiveness of this model to generalize, even more, offers the ability to obtain closer estimates values. Table 4 below shows results of all regression models for bifacial PV.

In Figure 5, which shows bifacial power output, similar tendencies are exhibited, with some nuanced differences. The RMSE is high at low polynomial degrees and continuously decreases as the complexity increases, much like the monofacial case. However, the RMSE decreases more slowly and the test RMSE remains higher than the training RMSE even at lower polynomial degrees. This suggests that bifacial power output data possess more variability or noise, likely due to more variables involved, such as Albedo. Figure 5 demonstrates that the RMSE in bifacial power output is at minimum between degrees 5 and 6 RMSE increases slightly and shows a sign of overfitting. While divergence between training and test RMSE has not significantly changed, it highlights the need for careful selection of polynomial complexity.

Table 4. Key performance indicators result for bifacial power output prediction models.

Regression Type	MAE	MSE	RMSE	R ² (%)
Linear	62.3	8777.8	93.7	86.7
Polynomial 2nd Degree	33	3397.8	58.3	94.7
Polynomial 3rd Degree	29.7	2501.1	50	96.2
Ridge	62.1	8782.5	93.7	86.7
Lasso	60.6	11001.1	104.9	83.4
Elastic Net	62.2	8778.5	93.7	86.7
Support Vector	40.2	8638.2	92.9	86.9
Decision Tree	21.2	2517.8	50.2	96.2
Decision Tree max depth: 10 min samples leaf: 4 min samples split: 12	19.7	1919.8	43.8	97.1
Random Forest	16.8	1387.9	37.3	97.9
XGBoost	18.03	1431.94	37.84	97.84
XGBoost Learning rate: 0.1 max depth: 7 number of estimators: 200 sub sample: 1.0	16.80	1301.37	36.07	98.0

**Figure 5.** Bifacial power output RMSE vs. polynomial complexity.

To assess sensitivity of bifacial-model performance to surface reflectance, we performed a dedicated albedo sensitivity experiment using model predictions for 2022 only. Table 5 reports RMSE (and R²) for each regression algorithm under fixed albedo values (0.2, 0.4, 0.6, 0.8) and under the time-varying weather-derived albedo.

Table 5. RMSE (R^2) results for bifacial power output prediction models under different albedo.

Regression Type	Albedo 0.2	Albedo 0.4	Albedo 0.6	Albedo 0.8	Weather Albedo
Linear	86.3 (88.8%)	86.8 (86.8%)	87.3 (90.6%)	87.8 (91.3%)	93.9 (87.0%)
Polynomial 2nd Degree	58.1 (94.9%)	56.5 (95.7%)	55.1 (96.3%)	53.7 (96.8%)	57.2 (95.2%)
Polynomial 3rd Degree	53.2 (95.7%)	51.8 (96.4%)	49.9 (96.9%)	48.5 (97.4%)	48.8 (96.5%)
Ridge	86.5 (88.7%)	87.1 (89.7%)	87.6 (90.6%)	88.1 (91.3%)	94.2 (86.9%)
Lasso	98.7 (85.3%)	100.1 (86.4%)	100.8 (87.5%)	101.2 (88.5%)	105.3 (83.7%)
Elastic Net	86.3 (88.8%)	86.8 (89.8%)	87.3 (90.6%)	87.8 (91.3%)	93.9 (87.0%)
Support Vector	133.1 (73.3%)	137.8 (74.3%)	142.4 (75.1%)	145.8 (76.1%)	152.5 (65.8%)
Decision Tree	70.6 (92.5%)	70.1 (93.4%)	67.1 (94.5%)	60.9 (95.8%)	51.5 (96.1%)
Decision Tree max depth: 6 min samples leaf: 4 min samples split: 12	64.6 (93.7%)	62.6 (94.7%)	60.7 (95.5%)	59.4 (96.1%)	45.4 (96.9%)
Random Forest	50.9 (96.1%)	49.5 (96.7%)	47.8 (97.2%)	46.2 (97.6%)	37.2 (97.9%)
XGBoost	49.8 (96.3%)	47.5 (96.9%)	47.2 (97.3%)	45.2 (97.7%)	36.8 (98.0%)
XGBoost Learning rate: 0.1 max depth: 5 number of estimators: 200 sub sample: 0.8	48.9 (96.4%)	46.8 (97.1%)	45.9 (97.4%)	44.1 (97.8%)	36.5 (98.1%)

The sensitivity results in Table 5 show consistent trends across models. Ensemble methods achieve the lowest RMSE values and highest R^2 for all albedo settings, with XGBoost giving the best performance (lowest RMSE = 36.5, $R^2 = 98.1\%$ for weather-derived albedo). Polynomial regressions (2nd–3rd degree) perform better than simple linear/regularized linear models, indicating modest nonlinearity is important for bifacial modeling. Performance of tree-based models generally improves with increasing albedo: peak bifacial gains (i.e., lowest RMSE) occur at higher fixed albedo (0.8) and under the weather-derived albedo time series. Support Vector Regression and Lasso show relatively poor results in this experiment, suggesting sensitivity to hyperparameter settings or regularization that underfits the bifacial albedo effects.

3.3. Cell Temperature Models

The evaluations of regression models for prediction of PV cell temperature provide excellent overall performance among all methods, but difference in accuracy and error metrics. The linear, Ridge and Elastic Net regression obtained similar results, a MAE of 1.77, a MSE of 6.08, a RMSE of 2.47, and R^2 as 97.1%. Although these models perform well, they are restricted in their performance to capture all nonlinear complexity within the data. The Lasso regression obtained larger MAE of 2.26 and lower R^2 of 95.3%, indicative of relatively weaker performance than the other models. Polynomial regression models improve performance significantly. A 2nd degree polynomial regression yielded MAE of 1.31, RMSE of 1.88 and R^2 of 98.3%, showing the ability to model nonlinear relationships.

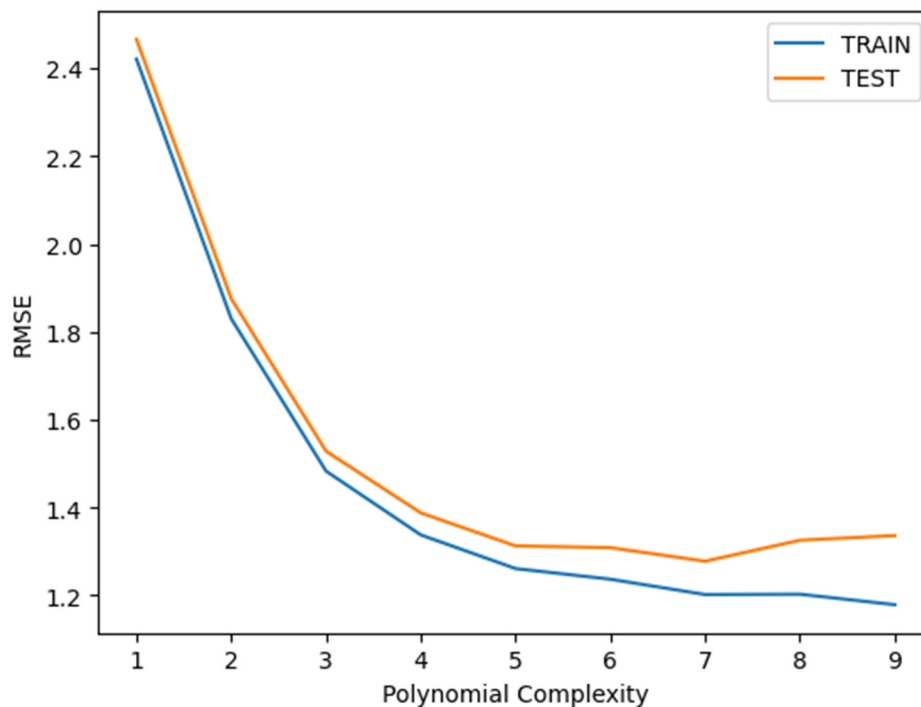
The 3rd degree polynomial regression further refines these results, achieving an MAE of 1.08, an RMSE of 1.53, and an R^2 of 98.9%, ranking it among the most accurate methods. However, polynomial models, particularly of higher degrees, risk overfitting, which could affect generalizability on new datasets.

The best results are achieved by ensemble-based models. The fine-tuned Decision Tree regression model, optimized with a maximum depth of 10 and specific splitting criteria, achieves an MAE of 1.33, an RMSE of 1.96, and an R^2 of 98.2%, demonstrating substantial predictive accuracy. Random Forest regression outperforms all other models, with an MAE of 0.93, an RMSE of 1.42, and an R^2 of 99.0%. Table 6 shows results of all regression models for cell temperature. However, the XGBoost models showed the best results among these models.

Figure 6 representing cell temperature prediction shows a different behavior. The initial RMSE is lower (~2.4) compared to the power output models, and it decreases more sharply as polynomial complexity increases. By degree 3, the RMSE for both training and test datasets becomes close, indicating that the model fits the data well. This stability continues up to degree 7, after which the test RMSE begins to increase slightly, suggesting minor overfitting. However, the increase is much less pronounced compared to the power output cases. This indicates that cell temperature data is less complex and more predictable, allowing for effective modeling with lower polynomial degrees. The optimal polynomial degree for cell temperature prediction lies between 4 and 5, where the model achieves high accuracy without overfitting.

Table 6. Key performance indicators result for cell temperature prediction models.

Regression Type	MAE	MSE	RMSE	R ² (%)
Linear	1.77	6.08	2.47	97.1
Polynomial 2nd Degree	1.31	3.52	1.88	98.3
Polynomial 3rd Degree	1.08	2.34	1.53	98.9
Ridge	1.77	6.08	2.47	97.1
Lasso	2.26	9.78	3.13	95.3
Elastic Net	1.76	6.08	2.47	97.1
Support Vector	1.08	3.11	1.76	98.5
Decision Tree	1.3	4.09	2.02	98.0
Decision Tree max depth: 10 min samples leaf: 4 min samples split: 10	1.33	3.83	1.96	98.2
Random Forest	0.93	2.02	1.42	99.0
XGBoost	0.91	1.75	1.32	99.16
XGBoost Learning rate: 0.1 max depth: 7 number of estimators: 200 sub sample: 0.5	0.83	1.50	1.23	99.28

**Figure 6.** Cell temperature RMSE vs. polynomial complexity.

In Figures 4–6, at low polynomial degree the model underfits because simple linear combinations of predictors fail to capture the principal nonlinear relationships in PV output: PV output is approximately proportional to incident plane-of-array irradiance but modulated by nonlinear effects such as cell temperature dependence, angular response, and spectral changes. Increasing polynomial degree up to a moderate level (typically degree 2–3) allows the regression to represent low-order interactions that mimic these physical dependencies, so RMSE falls.

However, further increasing polynomial degree produces fragile, high-variance predictors that tend to amplify measurement noise and multi-collinearity among generated features. Physically, the PV system does not exhibit arbitrarily high-order polynomial dependence on meteorological inputs—for example, the principal relation remains approximately linear in irradiance with corrections introduced by temperature and geometric factors. Thus, very high-degree polynomials often fit transient, non-physical fluctuations rather than improving the model of the underlying physical process; this leads to the increased RMSE observed for large degrees. This

effect is more pronounced when features such as albedo or wind are included, explaining why bifacial-module predictions (where albedo interactions matter more) may show a different optimal polynomial degree.

3.4. Interpretation of Feature Sensitivity Results

Both Tables 7 and 8, presented with the coefficients of different machine learning models (Linear, Ridge, Lasso, and Elastic Net) for predicting the power output of monofacial and bifacial solar panels based on various environmental and operational features. The coefficients represent the strength and direction of the relationship between each feature and the power output classification, with a positive coefficient indicating a direct relationship and a negative coefficient indicating an inverse relationship.

Table 7. Monofacial power output models coefficients.

Feature	Linear	Ridge	Lasso	Elastic Net
Temperature (°C)	-37.2346	-36.5488	-8.8464	-37.1395
GHI (W/m ²)	237.1639	235.226	200.567	236.6451
Relative Humidity	5.8361	4.6395	0	5.19718
Pressure (mbar)	11.2135	11.2026	0	10.83714
Surface Albedo	2.7006	2.9740	0	2.6560
Wind Speed (m/s)	15.8567	16.0294	1.12774	15.56685

Table 8. Bifacial power output models coefficients.

Feature	Linear	Ridge	Lasso	Elastic Net
Temperature (°C)	-36.8060	-36.1181	-8.1123	-36.7128
GHI (W/m ²)	250.705	248.6727	212.8182	250.145
Relative Humidity	5.121	3.8611	0	4.432
Pressure (mbar)	11.316	11.3014	0	10.911
Surface Albedo	8.941	9.1919	3.1134	8.891
Wind Speed (m/s)	17.591	17.7698	2.4278	17.281

Upon examining the coefficients for the monofacial power output models, the GHI feature stands out as the most prominent indicator because it exhibits the highest coefficients across each model (ranging from 200.57 to 237.16). This suggests GHI's role as the solar energy received per unit area makes it a strong indicator of monofacial panels' power output. Additionally, the coefficient for GHI varied little across each model, indicating the relationship of GHI to power output is consistently significant across all models but that the model and level of regularization applied in the Ridge, Lasso, and Elastic Net models resulted in slight variations in coefficient values. Similarly, temperature consistently displayed a negative relationship with power output across models reflected by negative coefficients (from -8.85 for Lasso to -37.23 for Linear). This indicates that as the temperature increases so does the decrease in power output of the monofacial panels, which is consistent with conventional justification for photovoltaic system reliability generally and with terms of photovoltaic system reliability improvements when temperature is lower. Relative Humidity shows a smaller impact on power output, with coefficients ranging from 0 to 5.84. While the effect of relative humidity is not very pronounced in the Lasso model (where the coefficient is 0), it still has a positive relationship with power output in the other models, albeit with a relatively small influence. This could reflect the fact that moisture in the air may affect the efficiency of the solar panels, but the effect is less significant compared to other factors like temperature or GHI.

Pressure has a small, positive influence on the power output, with coefficients ranging from 10.84 to 11.21 across models. This might suggest that atmospheric pressure has a relatively minor effect on the solar panel's performance, possibly due to its limited direct impact on irradiance or other environmental conditions that directly affect solar energy capture. Surface Albedo coefficients range from 2.66 to 2.97, indicating a modest positive relationship with power output. Albedo refers to the reflectivity of the surface surrounding the panel, and while this has some effect, it is not as significant as factors like GHI or temperature. Wind Speed also appears to have a positive effect, with coefficients ranging from 15.57 to 16.03. Wind helps cool the panels, potentially enhancing their efficiency, which aligns with the positive coefficients.

For the bifacial power output models, it can be observed similar trends in the coefficients for most features, though there are notable differences in the magnitude and significance of the coefficients.

GHI again shows the strongest relationship with power output, with coefficients ranging from 212.82 to 250.71. This suggests that similar to the monofacial panels, GHI is a dominant factor in predicting the power output of bifacial panels. However, the coefficients for GHI in the bifacial models are generally higher than those for monofacial panels,

which may be due to the additional contribution of reflected light captured by the rear side of bifacial panels, making GHI even more impactful for bifacial configurations. Temperature continues to show a negative relationship with power output across all models, with coefficients ranging from -8.11 (Lasso) to -36.81 (Linear). The temperature's effect is slightly less pronounced for bifacial panels compared to monofacial ones, but it still demonstrates the typical behavior of a negative correlation between temperature and efficiency.

Relative Humidity in the bifacial models shows a somewhat reduced influence compared to the monofacial models, with coefficients ranging from 3.86 to 5.12. Like the monofacial case, the impact is relatively small, and the relationship remains positive. Pressure is similar to the monofacial models, showing a small positive effect on power output, with coefficients ranging from 10.91 to 11.32. Surface Albedo holds significantly higher relevance in bifacial models, where coefficients ranged from 3.11 to 9.19. The implication of this is that surface reflectivity has a greater impact on bifacial panels, mostly in regard to rear-side light capture, where even the increased albedo of surrounding surface produces an increased overall energy yield. Wind Speed exhibits the same positive association with power output, with bifacial model coefficients ranging from 17.28 to 17.77. The positive association suggests wind cooling continues to be beneficial to both modeling types, although bifacial panels exhibited values that were slightly higher than monofacial models, most likely attributable to their greater overall energy capture potential.

When comparing the monofacial and bifacial models, the GHI remained a feature in both, with bifacial panel coefficients and estimates higher generally, indicating their increased energy capture capacity of sunlight and reflected light. The surface albedo has a more significant impact on bifacial panels, which makes sense, as these panels can utilize reflected light from the ground or other surfaces more effectively than monofacial panels. The temperature coefficient is slightly less negative for bifacial panels, suggesting that bifacial panels may be somewhat less sensitive to temperature increases, possibly due to their ability to capture additional light, which could offset some of the efficiency loss caused by higher temperatures. In general, the wind speed appears to have a slightly greater effect on bifacial panels, which may be due to their larger total area (front and rear) and increased efficiency, making them more sensitive to cooling effects.

Based on the results from Table 9, the cell's temperature is most strongly influenced by Temperature itself and GHI, as reflected by their larger positive coefficients. Both of these factors are central to understanding and predicting the temperature of the solar cell because they directly affect the amount of heat absorbed by the panel. Wind Speed also has a significant cooling effect, as indicated by its negative coefficient. Relative Humidity, Pressure, and Surface Albedo are other variables that have minimal influence on the cell temperature, and while both albedo and pressure have a slight positive effect, the effect of humidity is weak but negative.

Hence, it can be concluded that in order to successfully control or predict the temperature of a solar cell, the more predictive values to focus on would be ambient temperature, solar irradiance, and wind speed, and the relevance of other environmental parameters like humidity and pressure has a minimal significance.

Table 9. Cell's temperature models coefficients.

Feature	Linear	Ridge	Lasso	Elastic Net
Temperature ($^{\circ}\text{C}$)	9.9265	9.9265	9.1727	9.9686
GHI (W/m^2)	6.2495	6.2495	5.3953	6.2708
Relative Humidity	-0.1881	-0.1881	0	-0.1386
Pressure (mbar)	0.1004	0.1004	0	0.1184
Surface Albedo	0.3716	0.3716	0	0.3619
Wind Speed (m/s)	-1.6151	-1.6151	-0.5386	-1.5858

Figure 7a,b illustrate 3D cubic fits demonstrating the interaction between temperature, GHI, and power, with the first figure referring to monofacial modules. The monofacial modules show positive relationship between power and irradiance, and also a mild dependence on temperature, which indicates that temperature influenced performance, but irradiance was the primary influence on generating power in monofacial systems. This is contrary to the bifacial modules data that was more gradual in regard to power output, suggesting the modules would better tolerate temperature fluctuations from the additional energy generated from reflected light on the rear side of the module.

Figure 7c shows the expected cell temperature as a function of GHI and ambient temperatures and adds context to the thermal interaction and behavior at various environmental conditions. The results show that the cell temperature increased significantly as GHI increased for both module designs. While this relationship showed similarity to the power outputs in Figure 7a,b, this comparatively shows that bifacial modules better perform under higher temperature conditions, which implies that bifacial systems are not only maximizing power outputs as irradiance levels increase, but also improving operational stability in high temperature conditions.

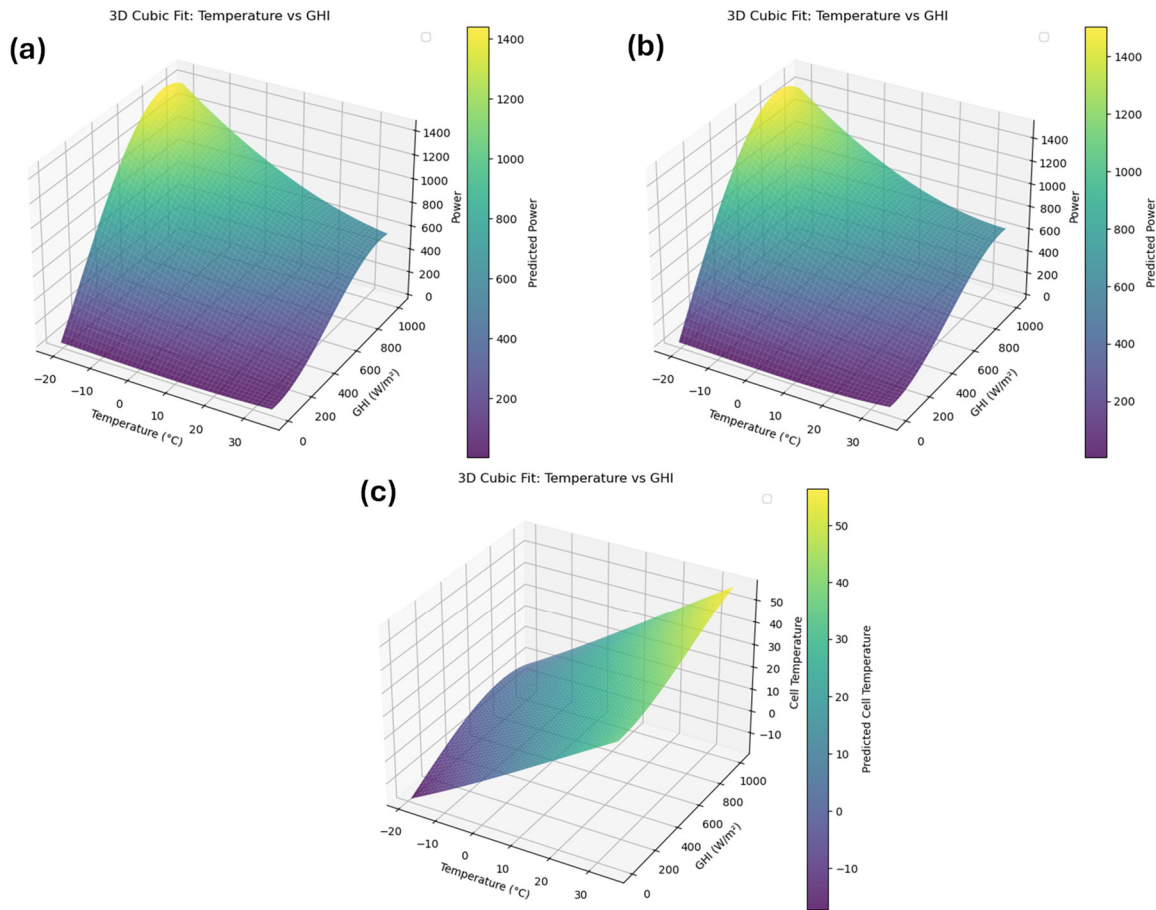


Figure 7. 3D cubic fit representations of the relationship between temperature, GHI, power output and cell temperature for monofacial and bifacial photovoltaic modules.

Figure 8 (difference surface) shows the bifacial minus monofacial predicted power across the sampled Temperature–GHI space. The difference is positive throughout the plotted domain, indicating a consistent bifacial advantage. The largest bifacial gains occur under low ambient temperatures combined with moderate-to-high GHI, a regime in which reduced cell temperatures increase module efficiency. These results are physically consistent with bifacial behavior, where the rear-side contribution becomes relatively more important when front-side irradiance is strong and ground reflectance is high.

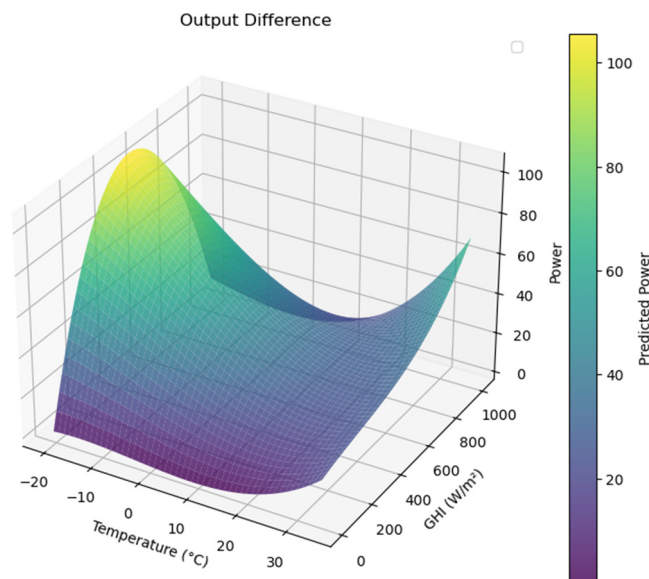


Figure 8. 3D cubic fit representations of the relationship between temperature, GHI, power output difference between bifacial and monofacial systems.

In Figure 9, at the root node, the tree splits on GHI, reflecting its significance in determining the target variable. Subsequent splits refine predictions by segmenting data into progressively homogeneous subsets. Each node provides quantitative details, including the squared error (a measure of model fit), the percent of samples, and the predicted value, which indicates how well the tree is able to explain the variance in the data. As the tree goes deeper the squared error usually goes down, indicating the model is isolating finer detail in the data. However, the nodes go deeper, usually, they have fewer samples making the predictions at risk for overfitting if the tree is too complex. The hierarchical structure helps to explain complex interactions among the features and how the totality of the features influences the target variable. For example, the first few splits are based on GHI indicating relative importance, and the following splits are based on temperature and Surface Albedo indicating the context in which these other factors are important in the prediction. The depth and branching of the tree indicate that there is a non-linear relationship among the inputs and the output, and they indicate the ability for machine learning models to reveal patterns is not obvious with traditional analysis.

This type of model is especially beneficial in renewable energy where it is important to understand and optimize variables such as the solar irradiance and albedo to better predict performance and optimize system efficiency. However, care must be taken to balance model complexity with interpretability because deeper trees that have splits that specify the targets may not generalize well to new data.

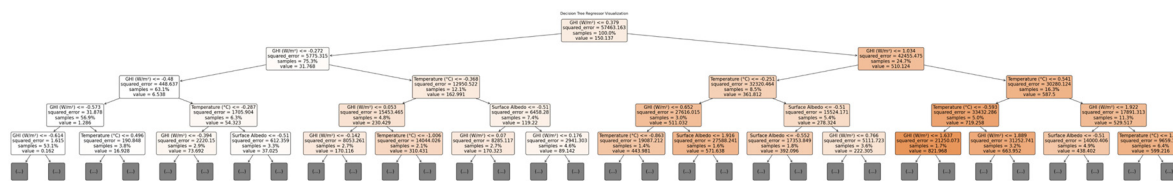


Figure 9. Decision tree visualization for predicting power output for the monofacial system.

In XGBoost Regression, the F-score indicates the importance of a feature used in the model, or how many times the feature was selected for splitting in a decision tree. The F-score for a feature is the number of times that feature was used to split the data based on all of the trees in the model. The higher the F-score, the more important the feature may be for predicting the target variable. However, the F-score only reflects the frequency of usage and does not take into account the magnitude of the feature or how much the model is influenced by the feature.

Figure 10 illustrates the feature importance for an XGBoost model using the F score. In terms of monofacial power output in Figure 10a, the most important feature is GHI (F1), which has the highest F score (1447), indicating its dominant impact on estimating monofacial power. This is logical, as GHI directly impacts the amount of solar isolation incident on the photovoltaic surface, and is critical for energy generation. The next important factor is ambient temperature (F0), which had a relatively high F score of (1150), indicating that it does impact for estimating monofacial power, as ambient temperature can influence the semiconductor properties of photovoltaic cells. Wind Speed (F5) and relative humidity (F2) are next at some level of contribution to the importance with F scores of (772) and (815), respectively, indicating that they had a secondary impact on estimating monofacial power output. Wind speed can cool the panels, while relative humidity could affect light transmission. Pressure (F3) and surface albedo (F4) were the least influential indicators, with F scores of (601) and (528), respectively, indicating that they have had relatively less impact with estimating monofacial systems under these conditions.

A similar trend is observed in the bifacial power output (Figure 10b) with GHI (F1) as the most significant feature, but with slightly less importance with an F score of 1389. Ambient temperature (F0) is still the second most significant feature with an F score of 1225, emphasizing its continued significance impacting system performance. Wind speed (F5) and relative humidity (F2) also have similar importance levels with scores of 765 and 809 respectively. Pressure (F3) and surface albedo (F4) again ranked the lowest, with F scores of 609 and 500. The lower importance of surface albedo than GHI in bifacial power output is surprising considering that bifacial systems are designed to utilize reflected light which is influenced by albedo.

In the case of cell temperature (Figure 10c), there is a slight variation in the importance of characteristics. GHI (F1), again, is the most important factor with an F value of 1301, followed by ambient temperature (F0) with a value of 1240, signifying that these factors play a very important role in the thermal properties of solar cells. Wind speed (F5) is given greater importance than in power output in predicting cell temperature ($F = 850$). This is a consequence of wind affecting surface temperatures. Relative humidity (F2), and pressure (F3), are moderately important characteristics with F values of 740, 586, respectively. Surface albedo (F4) is the least important parameter with a value of 448. This is primarily a result of the fact that temperature is not much affected by reflected sunlight.

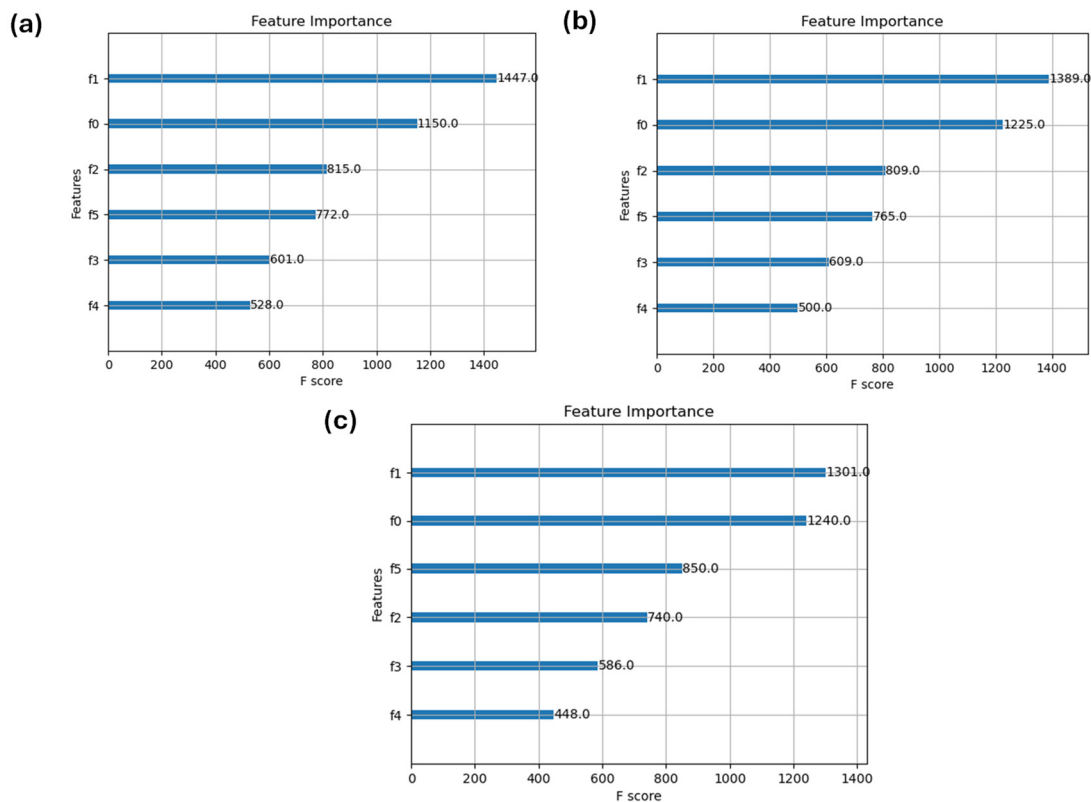


Figure 10. Feature importance (F score) for the XGBoost model predicting (a) monofacial power output, (b) bifacial power output, and (c) cell temperature.

Permutation-based feature sensitivity analysis reveals that PV power output is governed by the combined interaction of multiple meteorological variables rather than by a single dominant predictor. The baseline model achieved an RMSE of approximately 37 W with an R^2 of 0.98, and permutation of individual input features resulted in only marginal increases in RMSE ($\Delta\text{RMSE} < 0.02$), indicating a highly coupled multivariate response. Wind speed did not emerge as a dominant global predictor; however, this behavior is physically consistent with its indirect role in enhancing convective heat transfer and reducing cell operating temperature rather than directly contributing to energy generation. Consequently, the influence of wind speed becomes conditionally significant under high irradiance and elevated temperature regimes, where thermal losses are otherwise pronounced. The modest global importance values therefore confirm that wind speed functions as a secondary, temperature-mediated performance modifier rather than a primary energy driver.

From an engineering deployment perspective, the compared models also differ in computational efficiency. Linear and regularized linear regressors (Linear, Ridge, Lasso, ElasticNet) have the lowest training and inference costs because it uses closed-form or convex optimization structures. This makes them good for quick retraining or embedded implementations, but they are less accurate. Decision Tree models need modest training time and can make quick decisions. Random Forest, on the other hand, costs more to train as the number of trees increases, but it can still be used for practical plant-scale applications because it can be run in parallel and is easy to compute. XGBoost has the highest training cost because it uses sequential boosting and gradient-based optimization. However, this cost is usually incurred offline, so inference is still fast and can be done in near-real time. Since PV performance models are usually trained only a few times and used for continuous prediction, the higher training complexity of ensemble models is worth it for most operational and maintenance tasks because it makes them more accurate.

The incremental energy yield of bifacial modules has direct economic value but is strongly site dependent. Industry and peer-reviewed studies report typical annual bifacial energy gains in the range 5–15% depending on albedo, mounting and tracker use [38,39]. Capital cost for bifacial modules have narrowed in recent years; published analyses and project reports indicate small values in utility projects (1–5 $\text{¢}/\text{W}$) up to ~ 10 –12% in conservative academic assumptions, depending on glass/glass construction and Balance of System details [40]. In Ontario the standard techno-economic references place utility-scale PV capital and LCOE in well-established ranges; applying those references to bifacial economics shows that the net LCOE impact is the result of two competing effects; a modest increase in upfront cost, and an increase in annual energy production. Using industry

cost and LCOE ranges, bifacial designs that deliver more or equal to ~5–10% annual energy gain (typical for high-albedo or tracker sites) commonly reduce LCOE versus equivalent monofacial designs; at low albedo (near ground reflectance ≤ 0.2) the additional cost is unlikely to be fully recovered [41,42].

4. Conclusions

In this paper, the performance of various machine learning techniques for the prediction of power output and cell temperature for medium-scale photovoltaic systems was examined for monofacial and bifacial solar panels. With five years of meteorological data, starting from the solar global horizontal irradiance, ambient temperature, relative humidity, atmospheric pressure, surface albedo, and wind speed, the models could well indicate the effectiveness of machine learning algorithm-based techniques for the effective performance of solar power conversion systems for varied conditions. The models were tested for a data set comprising 43,800 h of observation for training (70%) and the rest for the process of testing (30%). The high degree of steadiness and generalization of ensemble models and their role in the process of optimizing hyperparameters indicate their suitability for mass applications for varied conditions.

- Power Output Prediction: The ensemble models (Random Forest and XGBoost) gave better results compared to the traditional regression models for both monofacial and bifacial solar panels. In monofacial power output prediction, the XGBoost model showed the best performance in terms of accuracy with an MAE of 17.58, RMSE of 37.79, and an R^2 of 97.56%. In the bifacial solar panel power output prediction similarly, the XGBoost model showed the best results in RMSE of 36.07 and an R^2 of 98.0%.
- Cell Temperature Prediction: The prediction capability for cell temperature was also equal, where both Random Forest and XGBoost gave an R^2 value of 99.0% and 99.16%, respectively for the cell temperature prediction task. The lowest RMSE value for the prediction of the cell temperature was achieved by XGBoost.
- Monofacial Modules: The performance of the models was more sensitive to the GHI and temperature, as revealed by the coefficients of 237.16 and -37.23 , respectively, from the linear regression analysis. The solar power generation was impacted negatively by the rise in temperature.
- Bifacial Modules: proved their superior efficiencies; they are capable of compensating for the effects of temperature changes effectively compared to monofacial modules. Feature importance showed the importance of the reflection albedo coefficient for bifacial modules (9.19), which was less prominent for monofacial modules (2.97).
- Feature importance analysis indicated that GHI was the dominant variable for both monofacial and bifacial solar power plant designs, having an F-score of 1447 and 1389 in the XGBoost models, respectively, followed by the temperature variable. Surface albedo and wind speed are very relevant for bifacial designs.
- The R^2 values of the polynomial regression models reached high levels (as high as 96.2% in bifacial solar panels), although the possibility of overfittings became apparent for higher levels of the polynomials above degree 5.
- Random Forest and XGBoost showed an impressive generalization capability, and their optimized hyperparameters avoided overfitting problems. In bifacial models, the Random Forest resulted in an MAE of 16.8 and an RMSE of 37.3, thus validating their prediction capability.
- Wind speed acts as an important secondary and conditional factor, influencing system performance indirectly through its effect on module operating temperature, particularly under high irradiance conditions.

While the comparative results show that tree-ensemble methods deliver strong pointwise predictions for the studied site, the study has important limitations; it relies primarily on SAM-simulated data for one location and a five-year NSRDB period with limited measured validation, uses a 70/30 random split. these constraints may limit generalizability across climates and operational plants.

For engineering practice it is recommended using the best-performing ensemble models as near-real-time performance estimators and anomaly detectors while coupling them to periodic local calibration against SCADA/plant measurements, deploying cell-temperature sensors and albedo monitoring for bifacial sites, retraining models regularly, and using residual/uncertainty estimates to inform targeted O&M (cleaning, soiling detection, and maintenance scheduling) before employing models for dispatch or financial decisions.

Future studies may utilize hybrid and deep learning algorithms to improve the predictability associated with power production, taking into consideration both environmental and operational variables in a more integrated manner. The scope of the analysis may be expanded to include physical layouts such as module height, tilt, spacing, and albedo to provide a deeper understanding of bifacial and monofacial module performance.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data will be made available upon request.

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author used Grammarly to rephrase sentences and enhance language. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Maka, A.O.M.; Alabid, J.M. Solar Energy Technology and Its Roles in Sustainable Development. *Clean Energy* **2022**, *6*, 476–483. <https://doi.org/10.1093/ce/zkac023>.
2. Anonymous. Clean Energy Can Fuel the Future—And Make the World Healthier. *Nature* **2023**, *620*, 245. <https://doi.org/10.1038/d41586-023-02510-y>.
3. Lazaroiu, A.C.; Gmal Osman, M.; Strejoiu, C.V.; et al. A Comprehensive Overview of Photovoltaic Technologies and Their Efficiency for Climate Neutrality. *Sustainability* **2023**, *15*, 16297. <https://doi.org/10.3390/su152316297>.
4. Sayed, E.T.; Olabi, A.G.; Alami, A.H.; et al. Renewable Energy and Energy Storage Systems. *Energies* **2023**, *16*, 1415. <https://doi.org/10.3390/en16031415>.
5. Garrod, A.; Ghosh, A. A Review of Bifacial Solar Photovoltaic Applications. *Front. Energy* **2023**, *17*, 704–726. <https://doi.org/10.1007/s11708-023-0903-7>.
6. Badran, G.; Dhimish, M. A Comparative Study of Bifacial versus Monofacial PV Systems at the UK’s Largest Solar Plant. *Clean Energy* **2024**, *8*, 248–260. <https://doi.org/10.1093/ce/zkae043>.
7. James, A. Solar Photovoltaic Energy: Advantages and Disadvantages. *Glob. J. Eng. Archit.* **2021**, *2*, 1–2.
8. Soomar, A.M.; Hakeem, A.; Messaoudi, M.; et al. Solar Photovoltaic Energy Optimization and Challenges. *Front. Energy Res.* **2022**, *10*, 879985. <https://doi.org/10.3389/fenrg.2022.879985>.
9. Mohammadi, F.; Neagoe, M. Emerging Issues and Challenges with the Integration of Solar Power Plants into Power Systems. In Proceedings of the 2020 Conference for Sustainable Energy (CSE), Brasov, Romania, 22–24 October 2020; pp. 157–173. https://doi.org/10.1007/978-3-030-55757-7_11.
10. Shaker, L.M.; Al-Amiery, A.A.; Hanoon, M.M.; et al. Examining the Influence of Thermal Effects on Solar Cells: A Comprehensive Review. *Sustain. Energy Res.* **2024**, *11*, 1–30. <https://doi.org/10.1186/s40807-024-00100-8>.
11. Sun, C.; Zou, Y.; Qin, C.; et al. Temperature Effect of Photovoltaic Cells: A Review. *Adv. Compos. Hybrid Mater.* **2022**, *5*, 2675–2699. <https://doi.org/10.1007/s42114-022-00533-z>.
12. Forootan, M.M.; Larki, I.; Zahedi, R.; et al. Machine Learning and Deep Learning in Energy Systems: A Review. *Sustainability* **2022**, *14*, 4832. <https://doi.org/10.3390/su14084832>.
13. Ukoba, K.; Onisuru, O.R.; Jen, T.-C. Harnessing Machine Learning for Sustainable Futures: Advancements in Renewable Energy and Climate Change Mitigation. *Bull. Natl. Res. Cent.* **2024**, *48*, 1–15. <https://doi.org/10.1186/s42269-024-01254-7>.
14. Yao, Z.; Lum, Y.; Johnston, A.; et al. Machine Learning for a Sustainable Energy Future. *Nat. Rev. Mater.* **2023**, *8*, 202–215. <https://doi.org/10.1038/s41578-022-00490-5>.
15. Al-Saban, O.; Alkadi, M.; Qaid, S.M.H.; et al. Machine Learning Algorithms in Photovoltaics: Evaluating Accuracy and Computational Cost Across Datasets of Different Generations, Sizes, and Complexities. *J. Electron. Mater.* **2024**, *53*, 1530–1538. <https://doi.org/10.1007/s11664-023-10897-7>.

16. Tina, G.M.; Ventura, C.; Ferlito, S.; et al. A State-of-Art-Review on Machine-Learning Based Methods for PV. *Appl. Sci.* **2021**, *11*, 7550. <https://doi.org/10.3390/app11167550>.
17. Mohamad Radzi, P.N.L.; Akhter, M.N.; Mekhilef, S.; et al. Review on the Application of Photovoltaic Forecasting Using Machine Learning for Very Short- to Long-Term Forecasting. *Sustainability* **2023**, *15*, 2942. <https://doi.org/10.3390/su15042942>.
18. Gaboitaolelwe, J.; Zungeru, A.M.; Yahya, A.; et al. Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison. *IEEE Access* **2023**, *11*, 40820–40845. <https://doi.org/10.1109/ACCESS.2023.3270041>.
19. Benitez, S.; Benitez, I.B.; Singh, J.G. A Comprehensive Review of Machine Learning Applications in Forecasting Solar PV and Wind Turbine Power Output. *J. Electr. Syst. Inf. Technol.* **2025**, *12*, 54. <https://doi.org/10.1186/s43067-025-00239-4>.
20. Hong, D.; Ma, J.; Wang, K.; et al. Real-Time Power Prediction for Bifacial PV Systems in Varied Shading Conditions: A Circuit-LSTM Approach Within a Digital Twin Framework. *IEEE J. Photovolt.* **2024**, *14*, 652–660. <https://doi.org/10.1109/JPHOTOV.2024.3393001>.
21. Grisanti, M.; Mannino, G.; Tina, G.M.; et al. Thermal Models of Monofacial and Bifacial PV Modules: Machine Learning and Physical Estimation Models Comparison. In Proceedings of the 2023 IEEE 50th Photovoltaic Specialists Conference (PVSC), San Juan, PR, USA, 11–16 June 2023. <https://doi.org/10.1109/PVSC48320.2023.10360013>.
22. Shi, Y.; Zhang, L.; Wang, S.; et al. A Short-Term Photovoltaic Power Interval Forecasting Method Based on Fuzzy Granular Computing and CNN-BiGRU. *Int. J. Low-Carbon Technol.* **2024**, *19*, 306–314. <https://doi.org/10.1093/ijlct/ctad131>.
23. Said, N.M.; Suleiman, R.F.R.; Rahim, N.H.A.; et al. Short-Term Photovoltaic (PV) Energy Prediction Using Machine Learning Approach. In *Springer Briefs in Applied Sciences and Technology*; Springer Nature Switzerland, 2024; pp. 111–118. https://doi.org/10.1007/978-3-031-63326-3_14.
24. Souhe, F.G.Y.; Mbey, C.F.; Kakeu, V.J.F.; et al. Optimized Forecasting of Photovoltaic Power Generation Using Hybrid Deep Learning Model Based on GRU and SVM. *Electr. Eng.* **2024**, *106*, 7879–7898. <https://doi.org/10.1007/s00202-024-02492-8>.
25. Miraftebzadeh, S.M.; Colombo, C.G.; Longo, M.; et al. A Day-Ahead Photovoltaic Power Prediction via Transfer Learning and Deep Neural Networks. *Forecasting* **2023**, *5*, 213–228. <https://doi.org/10.3390/forecast5010012>.
26. Zhang, C.; Fu, X.; Li, Z.; et al. Unified Fourier Graph-Based Spatiotemporal Learning and Corrected NWP for Multi-Site Ultra-Short Term Photovoltaic Power Forecasting. *IEEE Trans. Smart Grid* **2025**, *17*, 1639–1652. <https://doi.org/10.1109/TSG.2025.3628129>.
27. Fu, X.; Chang, F.; Li, Z.; et al. Redesigning the Decoder and Loss Function of Diffusion Transformer for PV Temporal Simulation. *IEEE Trans. Smart Grid* **2025**, *17*, 1629–1638. <https://doi.org/10.1109/TSG.2025.3627923>.
28. Psomopoulos, C.S.; Ioannidis, G.C.; Kaminaris, S.D.; et al. A Comparative Evaluation of Photovoltaic Electricity Production Assessment Software (PVGIS, PVWatts and RETScreen). *Environ. Process.* **2015**, *2*, S175–S189. <https://doi.org/10.1007/s40710-015-0092-4>.
29. Maulud, D.; Abdulazeez, A.M. A Review on Linear Regression Comprehensive in Machine Learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147. <https://doi.org/10.38094/jastt1457>.
30. Belany, P.; Hrabovsky, P.; Sedivy, S.; et al. A Comparative Analysis of Polynomial Regression and Artificial Neural Networks for Prediction of Lighting Consumption. *Buildings* **2024**, *14*, 1712. <https://doi.org/10.3390/buildings14061712>.
31. De Vlaming, R.; Groenen, P.J.F. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. *BioMed Res. Int.* **2015**, *2015*, 143712. <https://doi.org/10.1155/2015/143712>.
32. Li, X.; Wang, Y.; Ruiz, R. A Survey on Sparse Learning Models for Feature Selection. *IEEE Trans. Cybern.* **2022**, *52*, 1642–1660. <https://doi.org/10.1109/TCYB.2020.2982445>.
33. Hajihosseini, M.; Maghsoudi, A.; Ghezlbash, R. Regularization in Machine Learning Models for MVT Pb-Zn Prospectivity Mapping: Applying Lasso and Elastic-Net Algorithms. *Earth Sci. Inform.* **2024**, *17*, 4859–4873. <https://doi.org/10.1007/s12145-024-01404-5>.
34. Harati, S.; Gomari, S.R.; Rahman, M.A.; et al. Performance Analysis of Various Machine Learning Algorithms for CO₂ Leak Prediction and Characterization in Geo-Sequestration Injection Wells. *Process Saf. Environ. Prot.* **2024**, *183*, 99–110. <https://doi.org/10.1016/j.psep.2024.01.007>.
35. Singh Kushwah, J.; Kumar, A.; Patel, S.; et al. Comparative Study of Regressor and Classifier with Decision Tree Using Modern Tools. *Mater. Today Proc.* **2022**, *56*, 3571–3576. <https://doi.org/10.1016/j.matpr.2021.11.635>.
36. Strobl, C.; Malley, J.; Tutz, G. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychol. Methods* **2009**, *14*, 323–348. <https://doi.org/10.1037/a0016973>.
37. Sagi, O.; Rokach, L. Approximating XGBoost with an Interpretable Decision Tree. *Inf. Sci.* **2021**, *572*, 522–542. <https://doi.org/10.1016/j.ins.2021.05.055>.
38. Deline, C.; Ayala Peláez, S.; Marion, B.; et al. Bifacial PV System Performance: Separating Fact from Fiction; National Renewable Energy Laboratory: Chicago, IL, USA, 2019.

39. Alam, M.; Gul, M.S.; Muneer, T. Performance Analysis and Comparison Between Bifacial and Monofacial Solar Photovoltaic at Various Ground Albedo Conditions. *Renew. Energy Focus* **2023**, *44*, 295–316. <https://doi.org/10.1016/j.ref.2023.01.005>.
40. Rodríguez-Gallegos, C.D.; Bieri, M.; Gandhi, O.; et al. Monofacial vs. Bifacial Si-Based PV Modules: Which One Is More Cost-Effective? *Sol. Energy* **2018**, *176*, 412–438. <https://doi.org/10.1016/j.solener.2018.10.012>.
41. Lazard 2023 Levelized Cost of Energy+. Available online: https://www.lazard.com/research-insights/2023-levelized-cost-of-energyplus/?utm_source=chatgpt.com (accessed on 5 February 2026).
42. Annual Planning Outlook. Available online: <https://www.ieso.ca/Sector-Participants/Engagement-Initiatives/Engagements/Annual-Planning-Outlook> (accessed on 5 February 2026).