Critical Feature Selection and Machine Learning-based Models for Monofacial and Bifacial Photovoltaics

Ansari Aadil Shahzad, Prajowal Manandhar, O. A. Qureshi(✉), Ahmer A. B. Baloch, Edwin Rodriguez-Ubinas, Vivian Alberts, and Sgouris Sgouridis

Dubai Electricity and Water Authority, Dubai, UAE
{ansari.shahzad,prajowal.manandhar,omer.qureshi,ahmer.baloch,
edwin.ubinas,vivian.alberts,sgouris.sgouridis}@dewa.gov.ae

Abstract. Prediction accuracy has paramount importance in reliable PV solar plant performance. This helps with optimal plant design, economic assessment, smooth grid integration, and plant operations. Machine learning (ML) models help with faster, reliable, and accurate prediction of annual energy yield that is valid over a wide range of climatic conditions, module specifications, and site conditions. In this study, an ensemble-learning algorithm with regression trees is used to predict the performance of both monofacial and bifacial modules. Training data is prepared from parametric simulation results obtained using System Advisor Model (SAM) with an energy yield range of 120–584 kWh/m^2 for monofacial modules and 134–706 kWh/m^2 for bifacial modules. The results showed that ensemble-learning based ML algorithm can predict the energy yield of monofacial and bifacial modules with RMSE of 2.89 kWh/m^2 and 4.65kWh/m^2.

Keywords: Generic Model · Machine Learning · Photovoltaics · Monofacial · Bifacial

1 Introduction

Photovoltaic (PV) cell technology aggressively competes with other renewable technologies based on the lower installation cost and faster return on investment [1]. Monofacial technology has been dominant in the past; however, a new technology that can generate power on both sides of the PV module shows great promise. The market share of bifacial PV cells, modules, and systems offers a pathway to significantly decrease the Levelized cost of energy (LCOE) compared with conventional monofacial PV modules. As a result, bifacial PV technologies are expected to reach 80% of the market share by 2031 [2]. One major barrier to the broader use of bifacial PV modules and systems is a lack of knowledge and experience with a system design that takes advantage of the specific features of bifacial cells. Bifacial system performance cannot be predicted with confidence using current PV performance modeling applications because of difficulties in accurately estimating the reflected irradiance to the back-side. It is therefore crucial

Y.-B. Lin et al. (Eds.): SGIoT 2021, LNICST 447, pp. 138–150, 2022.
https://doi.org/10.1007/978-3-031-20398-5_12

to develop new improve methods to forecast the energy yield (EY) of PV modules to determine the best economic design for a PV solar plant.

Simulation software such as System Advisor Model (SAM), PVSyst, and HelioScope can predict the performance of both monofacial and bifacial modules for different configurations and climates. Fixed modules with a tilt the same as the location's latitude are the most commonly adopted configuration for PV modules in large to small-scale applications. The mentioned simulation software can predict energy yield for this configuration at multiple locations, range of module specifications, and site conditions. However, it's a time-consuming affair and requires careful setup of the software parameters. A trained Machine Learning (ML) based model can be used to predict a PV plant's energy yield with a high accuracy level. These ML models provide the advantage of fast setup and accuracy of the prediction. Accurate and reliable prediction of the power output of other renewable energy sources such as wind turbines using ML models was demonstrated by the previous research [3]. Similar efforts were made to predict the daily mean solar power using ML techniques [4]. However, the prediction of these ML models depends on the training data [5]. Hence, the simulation environment must be set up carefully to obtain the training data for the ML model. Khandakar et al.demonstrated the performance of multiple ML algorithms for predicting the power output of the PV modules operating in Qatar [6]. They could predict the instantaneous output power of a PV module with a minimal RMSE of 2.14 W..

In this work, SAM software is utilized to obtain the annual energy yield of both monofacial and bifacial modules for a wide range of site conditions, module specifications, and climatic conditions. It is observed that an ensemble-learning algorithm with bagged trees can predict the annual energy yield of both PV technologies faster and more accurately. These fast results obtained from ML models can help to perform quick assessments and obtain the near-optimal design of PV plants according to site conditions [4].Also, it can be helpful to estimate the impact of selecting the PV module specifications on plant EY based. Critical feature selection criterion has been determined in this work as well. The selection of the independent variables is performed based on the ranking of the correlation of each independent variable to the dependent variable, which determines their significance to dependent variables such as DC Energy. The correlation of each independent variable to the dependent variable is obtained, and features are selected based on the higher ranks of their correlations. This work aims to incorporate a machine learning approach with simulated data to provide a fast and accurate global estimation of PV module performance.

2 Methodology

This section describes the approach adopted to generate training data, feature selection criteria and development of the ML model. Mainly there are two stages in this process.The training data is generated using the SAM software in the first stage. The range of parametric study is defined to cover a wide range of site conditions and module specifications, followed by a selection of locations to provide global coverage of climatic conditions. This data provides the performance of a fixed-mounted PV module, tilted at an angle the same as the location's latitude, over a wide range of climatic conditions, cell efficiency (η),albedo(α_{sur}), bifaciality (φ_{bi}), and temperature coefficient (γ).

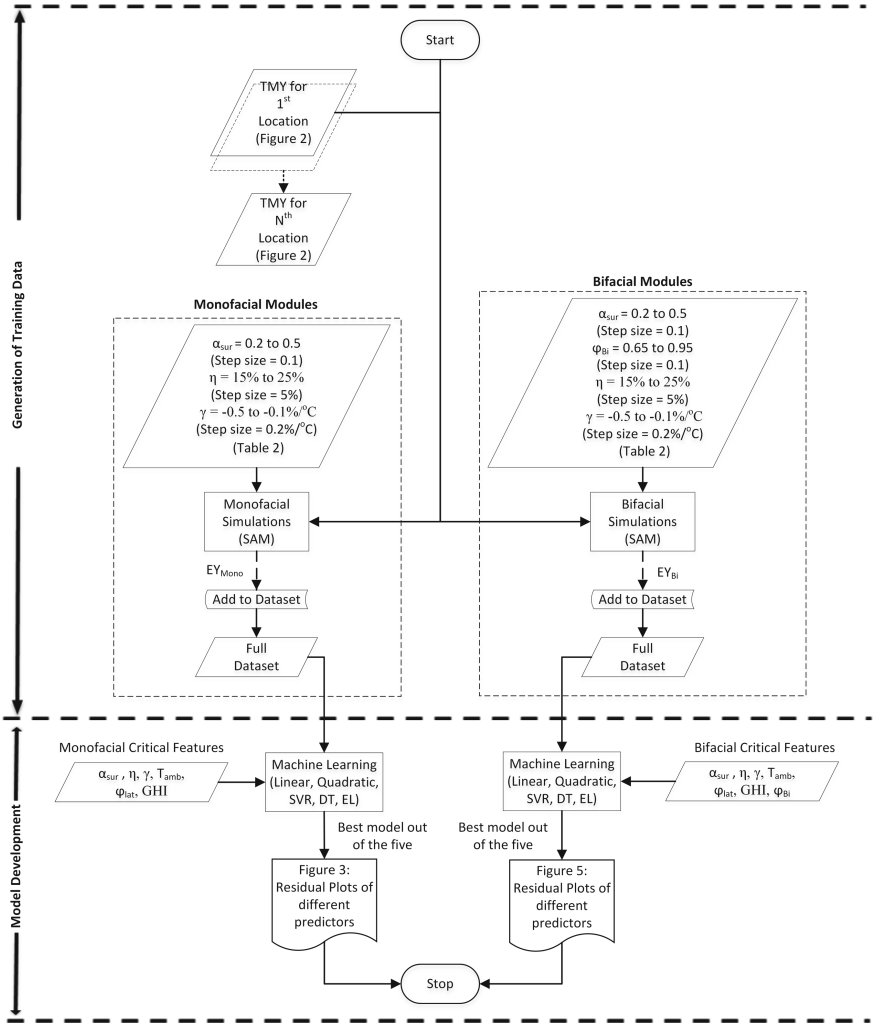

Fig. 1. Pictorial representation of the workflow for the SAM simulations and Machine Learning modeling.

The second stage starts with feature selection by checking the correlation between the parametric study variables and annual energy yield. Once the critical features are selected, multiple ML algorithms are tested and the best-performing algorithm is reported along with the ranking of modeling features. These ML algorithms include Robust linear, Quadratic, Support Vector Regression, Decision Trees and Ensemble Learning. Figure 1 shows the workflow from training data generated from the parametric study to the developing ML models.

2.1 Simulations Model Details

Simple efficiency model is setup in SAM to perform the parametric simulation runs over a wide range of conditions. A standard PV plant with two rows of seven modules, each with a total module area of 10.4 m^2 was set up for the simulations. The modules are oriented in the south direction for locations in the northern hemisphere and vice versa. Modules are tilted at an angle equal to the latitude of that location. Table 1 shows the fixed parameter values of both monofacial and bifacial simulations.

Table 1. Selected values of the parameter in the Simple Efficiency PV model in SAM.

Parameter	Monofacial module	Bifacial module
Module area (A_m)	0.7474 m^2	
Maximum power voltage (V_{mp})	30 V_{dc}	
Open circuit voltage (V_{oc})	36 V_{dc}	
Module structure/Mounting	Glass-cell-glass/Ground mounted	
Module elevation	1.5 m	
Ground coverage ratio	0.3	
Shading	No Shading	
Transmission factor	–	0.013

2.2 Range of Parameters

Table 2 provides the range of parametric studies covered for both monofacial and bifacial simulations. The range of module specifications, such as power temperature coefficient, cell efficiency, and bifaciality, are fully covered in the parametric study, along with a range of possible site albedo. Other site conditions such as GHI and ambient temperature are covered by simulating a wide range of climatic conditions.

Table 2. Summary of parametric study with a range of critical variables

Parameter range	Monofacial simulation	Bifacial simulation
Power temperature coefficient (γ)	−0.5 to −0.1%/°C (Step size = 0.2%/°C)	
Cell efficiency (η)	10%–20% (Step Size = 5%)	
Albedo (α_{sur})	0.2–0.5 [7] (Step size = 0.1)	0.2–0.5 [7] (Step size = 0.1)
Bifaciality (φ_{Bi})	–	0.65–0.95 [8] (Step size = 0.1)

2.3 Selection of Locations

A selected weather station from a grid of 10° × 10° (Latitude × Longitude) is used as a representative weather station that provides a total of 218 weather stations [9]. Polar or near-polar locations have a very cold climate and low GHI. Hence locations above the latitude of 60° and below −60° were removed before random selection. The filtered locations are plotted on the world map with climate classification by Koppen-Geiger [10], as shown in Fig. 2.

Average sun-hours ambient temperature (T_{ASHAT}) was calculated for each location using the subset of hourly temperatures observed during sun-hours. This averaged value should provide a better sense of change in the performance of PV modules compared to the average ambient temperature estimated for all hours throughout the day.

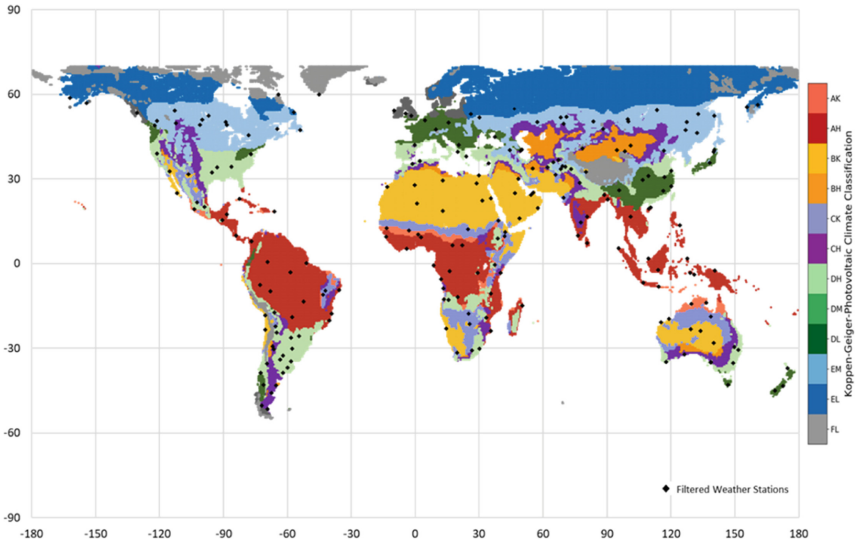

Fig. 2. Koppen-Geiger PV classification w.r.t Irradiation and Temperature. The first letter shows the Temperature-Precipitation zones (A- Tropical, B- Desert, C- Steppe, D- Temperate, E- Cold & F-Polar). The second letter shows the Irradiation zones (K- Very High, H- High, M- Medium & L- Low) [9, 10]

2.4 Feature Selection

The feature selection process allows the identification of significant features constituting the given data. Thus, it facilitates identifying and isolating prominent features to ensure quality in the underlying information. In addition, feature selection reduces the dimensionality of available data, making mining tasks much simpler. In this study, Correlation Ranker-based feature selection algorithm was used to evaluate the features and a rank-list was generated based on their score. These scores were used to select the most relevant features among cell efficiency, albedo, temperature coefficient, bifaciality, irradiance, Average sun-hours ambient temperature, and latitude.

2.5 Machine Learning Model

In this study different ML models have been used to provide an improved prediction of the significant impact of the predictors on the energy yield between bifacial and monofacial modules. These different ML models are compared with robust linear and quadratic models to determine which one of the two PV technologies is affected more by different predictors [11].

Data is key to any black-box model, and there is never enough data to train these model. The full dataset is divided into training and testing datasets in 7:3 ratios in random order. Therefore, removing a part of training data from an already reduced dataset for validation poses a problem of under-fitting. Since the holdout approach reduces the training data size, there is a risk of losing important patterns/trends in the data set, increasing errors induced by bias. Therefore, using the K-fold cross-validation approach in the training dataset that provides ample data for training the model and leaves a subset for validation. In K-fold cross-validation, the data is divided into K-subsets. In the holdout method, the '1/K' part of the data is repeated 'K' times, such that each time, one of the 'K' subsets is used as the validation set and the other 'K-1' subsets are put together to form a training set. The error estimation is averaged over all 'K' trials to get the total effectiveness of the model. In the case of every data, the point gets to be in a validation set exactly once and gets to be in a training set 'K-1' times. This allows to significantly reduce a bias as most of the data is used for fitting, and it also significantly reduces variances as most of the data is also being used for invalidation. And interchanging the training and validation sets also adds to the effectiveness of this method. As a general rule or empirical evidence, $K = 5$ is used in ML modeling.

2.5.1 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a supervised learning algorithm that uses the same principle as the Support Vector Machines (SVM) to model and predicts discrete values [12]. The basic idea behind SVR is to find the best fit line, which is to have the hyperplane that contains the maximum number of points. In SVR, the kernel is used to represent the function to approximate the data. Amongst various kernels available, Gaussian functions (radial basis [15] perform well for the data that shows Gaussian or normal distribution.

2.5.2 Decision Trees (DT)

Decision Trees (DT) is a supervised learning algorithm to predict discrete values of the parameters used to train the model [13]. DT starts at the root and traverses slowly towards the nodes by partitioning the predictors using the divide and conquer strategy as per that predictor that has the highest impact on uniformity in the result after a supervised learning technique used to combine the multiple regression as per the split.

2.5.3 Ensemble Learning (EL)

Ensemble Learning (EL) is a supervised learning algorithm that combines the outputs from various ML approaches, such as decision trees (with various bagging and/or boosting) or various ML models into one [14]. It is primarily used to improve the performance of the regressive prediction or function approximation as ensemble methods use multiple learning approaches to obtain better predictive performance than that could be obtained from any of its constituent learning algorithms alone. In the ensemble approach, 200 bagged trees have been considered and it consist of varying sample of training data.

3 Results and Discussion

Different machine learning models are used to understand the importance of other predictors such as efficiency (η), albedo (α_{sur}), temperature coefficient (γ), bifaciality (φ_{bi}), GHI_T, T_{ASHAT}, and latitude (φ_{lat}) on the predictors' energy yield of monofacial and bifacial modules.

Table 3 shows that efficiency, GHI, and T_{ASHAT} are the most significant attributes to the dependent variable DC energy yield compared to other attributes for both Monofacial and Bifacial modules. Similarly, latitude has an inverse effect on the EY of the monofacial modules, which means that the higher the absolute value of latitude lower will be the EY. In the case of monofacial modules, the albedo has the least significant attribute. This is expected as the site albedo has very little effect on the EY of monofacial modules compared to bifacial modules, which play an imperative role in predicting performance. For the bifacial module, bifaciality has the least significant attribute. Generally, bifaciality has a higher effect on the performance of vertically oriented bifacial modules [16]. Since, the tilt of the modules was set according to the latitude, the overall effect of change in bifaciality might be diminished. Hence, these two attributes have been removed on separate occasions to see the impact on the prediction performance of each model (see Table 4).

Table 3. Correlation of features to predict EY of Monofacial and Bifacial modules.

Attributes	Monofacial modules	Bifacial modules
Efficiency (η)	0.7716	0.7616
GHI	0.5603	0.5561
T_{ASHAT}	0.3325	0.3402
Temperature Coefficient (γ)	0.0912	0.1047
Albedo (α_{sur})	**0.0254**	0.1563
Latitude (φ_{lat})	−0.2302	−0.2336
Bifaciality (φ_{bi})	–	**0.0645**

Table 4 compares RMSE for different machine learning approaches: Support vector regression, decision trees, and ensemble learning along with the robust linear and quadratic model. The R-squared values are given only for the best model out of the above-mentioned models. "Table 4 also shows the change in RMSE when feature with least correlation is removed from their respective model." E.g., removing albedo from the monofacial model results in an increase in RMSE by minimal (<10%) for each of the ML modelling approaches. However, removing bifaciality from bifacial modules can increase the RMSE by as high as 70% compared to the model that includes all the features. Also, there is a notable change in the model correlation due to the exclusion of bifaciality. Physical interpretation, however, supports the inclusion of all these features, so it is recommended to use the model with all features included.

SVR in comparison to statistical models (linear/quadratic) performs better as it uses feature space to transform and fit the data into a higher dimension by using its kernels.

Table 4. Comparison of Root Mean Square Error (RMSE) between different machine learning techniques.

Technology	RMSE (kWh/m^2)					R-squared
	Robust linear	Quadratic	SVR	Decision trees	Ensemble model	
Monofacial model inclusive of all predictors	18.58	7.12	6.63	5.39	2.89	0.998
Monofacial model without albedo	20.06	7.81	6.55	5.59	3.18	0.997
% change in RMSE	*8%*	*10%*	*−1%*	*4%*	*10%*	
Bifacial model inclusive of all predictors	20.61	7.56	6.38	9.85	4.64	0.997
Bifacial model without Bifaciality	22.73	10.65	9.05	10.45	7.89	0.992
% change in RMSE	*10%*	*41%*	*42%*	*6%*	*70%*	

Figure 3 provides details about residual plots for different predictors for monofacial modules. It can be observed that there is no biasness occurring and no specific structure can be observed in the plots.

Figure 4 shows the simulated vs the ensemble modeled response for the DC energy for monofacial modules. It is clear the the response falls between the 5% error for the complete range throughout the prediction and most of the time, the data points are very

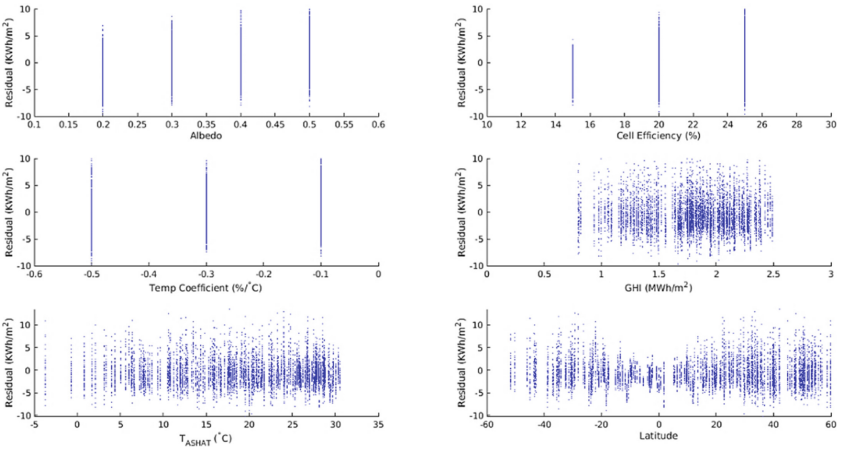

Fig. 3. Residual plots for monofacial modules for all parametric variables to test for any biased error trend (a) w.r.t albedo (b) w.r.t cell efficiency (c) w.r.t. power temperature coefficient (d) w.r.t. aggregate irradiance (e) w.r.t T_{ASHAT} and (f) w.r.t. latitude

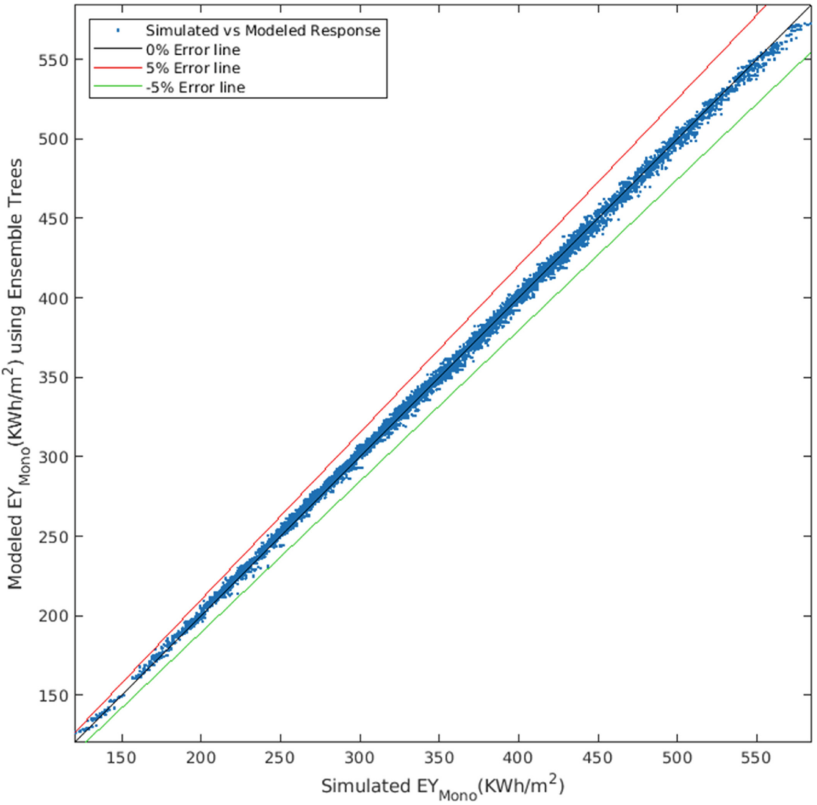

Fig. 4. Prediction performance of ensemble model for monofacial modules

close to the 0% error line. This is because the predicted values of EY are very close to the actual EY generated from the SAM simulations. This shows that the developed ML model model works very accurately.

Figure 5 provides details about residual plots for different predictors for bifacial modules. We can observe that there is no biasness occurring and no specific structure can be observed in the plots.

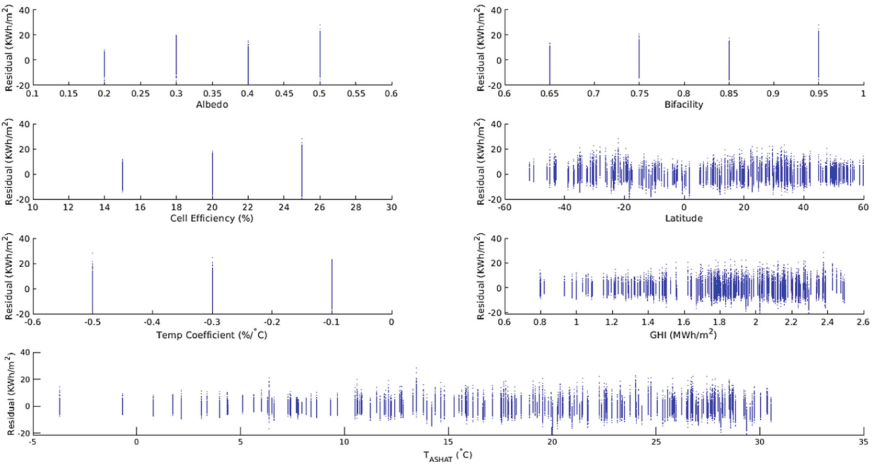

Fig. 5. Residual plots for bifacial modules for all parametric variables to test for any biased error trend (a) w.r.t albedo (b) w.r.t bifaciality (c) w.r.t cell efficiency (d) w.r.t. latitude (e) w.r.t. power temperature coefficient (f) Aggregate Irradiance and (g) w.r.t T_{ASHAT}

Figure 6 shows the simulated vs the ensemble modeled response for the DC energy for bifacial modules. It can be observed that the reposne falls between the 5% error for the complete range throughout the prediction. Since the factors affecting the EY of bifacial modules are comparatively higher than that of monofacial modules, the deviation of data points from the 0% error line increases compared to the monofacial module response.

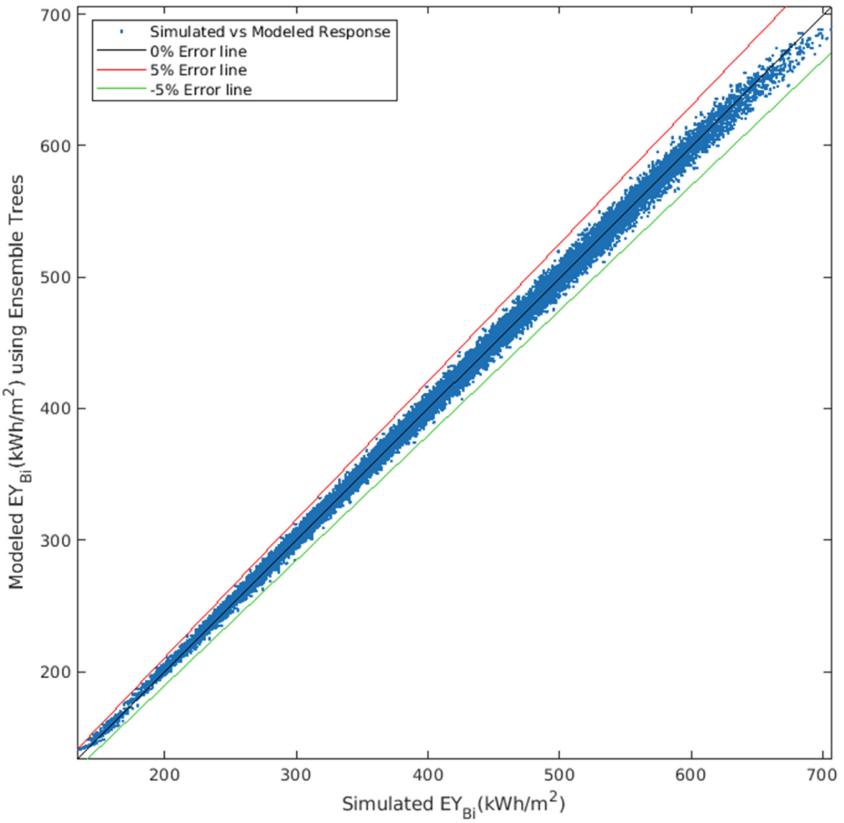

Fig. 6. Prediction performance of ensemble model for bifacial modules.

4 Conclusion

A simulation-based parametric study for both monofacial and bifacial modules was carried out for global climatic conditions. The generated simulation data is used to train and comapre different ML algorithms. Feature selection was done using a correlation ranking filter that allowed choosing significant predictor attributes for DC Energy. Among the analyzed algorithms, ensemble-learning algorithm with bagged trees performed best ($RMSE_{Mono}$= 2.89 kWh/m^2, R^2_{Mono} = 0.9984; $RMSE_{Bi}$ = 4.64 kWh/m^2, R^2_{Bi} = 0.9967) in predicting the output power for both PV technologies.

Nomenclature

Abbreviations		Subscripts	
A	Tropical Regions	amb	Ambient
ASHAT	Average Sun-Hours Ambient Temperature	bi	Bifacial

(*continued*)

(*continued*)

B	Desert Regions	dc	Direct Current
C	Steppe Regions	lat	Latitude
c-Si	Crystalline Silicon	m	Module
D	Temperate Regions	mono	Monofacial
DT	Decision Trees	mp	Maximum Power
E	Cold Regions	oc	Open Circuit
EL	Ensemble Learning	sur	Surface
EY	Energy Yield (kWh)	t	Total
F	Polar Regions	Symbols	
GCR	Ground Coverage Ratio	α_{sur}	Albedo
GHI	Global Horizontal Irradiation (MWh/m^2)	φ_{bi}	Bifaciality (%)
H	Regions with High Irradiance	φ_{lat}	Latitude
I	Current (Amps)	η_{ι}	Cell efficiency of PV Module (%)
K	Regions with very High Irradiance	γ	Power Temperature Coefficient (%/°C)
L	Regions with Low Irradiance		
M	Regions with Medium Irradiance		
ML	Machine Learning		
PV	Photovoltaic		
SAM	System Advisor Model		
STC	Standard Test Conditions		
SVR	Support Vector Regression		
T	Temperature (°C)		
TMY	Typical Meteorological Year		
V	Voltage (V)		

References

1. Pv, U.S. Arabia, S.: Saudi Arabia's second PV tender draws a world record low bid of $0.0104/kWh, pp. 6–13 (2021)
2. International, T., Roadmap, T.: Bifacial solar cells, and modules. https://pv-manufacturing.org/bifacial-solar-cells-and-modules/ (2020)
3. Mishra, S.P., Dash, P.K.: Short term wind power forecasting using Chebyshev polynomial trained by ridge extreme learning machine. In: 2015 IEEE Power, Communication and Information Technology Conference (PCITC), pp. 173–177 (2015)
4. Jawaid, F., NazirJunejo, K.: Predicting daily mean solar power using machine learning regression techniques. In: 2016 Sixth International Conference on Innovative Computing Technology (INTECH), pp. 355–360 (2016)

5. Sheng, H., Xiao, J., Cheng, Y., Ni, Q., Wang, S.: Short-term solar power forecasting based on weighted gaussian process regression. IEEE Trans. Industr. Electron. **65**(1), 300–308 (2018)
6. Khandakar, A., et al.: Machine learning based photovoltaics (PV) power prediction using different environmental parameters of Qatar. Energies **12**(14), 2782 (2019)
7. Product-updates, P. U. B. Surface Albedo – most frequent questions, pp. 1–12 (2019)
8. Deline, C., et al.: Evaluation and field assessment of bifacial photovoltaic module power rating methodologies. In: 2017 IEEE 44th Photovoltaics Special Conference PVSC 2017, pp. 1–6 (2017)
9. Lawrie, L.K., Crawley, D.B.: Development of global typical meteorological years (TMYx) (2019). https://climate.onebuilding.org/ (Accessed 10 Oct 2021)
10. Ascencio-Vásquez, J., Brecl, K., Topič, M.: Methodology of Köppen-Geiger-Photovoltaic climate classification and implications to a worldwide mapping of PV system performance. Sol. Energy **191**, 672–685 (2019)
11. Su, D., Batzelis, E., Pal, B.: Machine learning algorithms in forecasting of photovoltaic power generation. In: 2019 International Conference on Smart Energy Systems and Technologies (SEST), pp. 1–6 (2019)
12. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. In: Proceedings of the 9th International Conference on Neural Information Processing Systems (NIPS'96), pp. 155–161. MIT Press, Cambridge, MA, USA (1996)
13. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. **21**(3), 660–674 (1991)
14. Arbib, M.A.: The Handbook of Brain Theory And Neural Networks (1998)
15. Wu, Y., Wang, H., Zhang, B., Du, K.-L.: Using radial basis function networks for function approximation and classification. Int. Sch. Res. Not. **2012**, 324194 (2012)
16. Baloch, A..A..B.., Hammat, S., Figgis, B., Alharbi, F.H., Tabet, N.: In-field characterization of key performance parameters for bifacial photovoltaic installation in a desert climate. Renew. Energy **159**, 50–63 (2020)