PV Output Power Prediction Using Regression Methods

Abdulhameed Aboumadi[1] and Hilal Arslan[2(✉)]

[1] Department of Electrical and Electronic Engineering, Ankara Yıldırım Beyazıt Univesity, Ankara, Turkey
[2] Department of Software Engineering, Ankara Yıldırım Beyazıt University, Ankara, Turkey
hilalarslan@aybu.edu.tr

Abstract. Over the past few years, the general public has become increasingly aware of climate change and the role of greenhouse gas emissions, especially carbon dioxide, in contributing to it. Therefore, individuals, businesses, and governments around the world have taken steps to reduce their emissions. One of these steps is to increase adoption of renewable energy sources, such as solar power which provides clean energy, in addition to low building and operation costs and minimal maintenance requirements. Accurate estimation of solar energy production is crucial to ensure the stability of electrical networks as the transition to renewable energy sources such as solar power increases. In this study, machine learning regression algorithms including artificial neural networks, support vector regression, regression trees, and k-nearest neighbor are performed to estimate hourly solar energy production of one month using historical production data and various meteorological parameters. The models are optimized using grid search and validated using K-fold cross validation method. The performance of the models is evaluated using the RMSE, MAE, and R2 evaluation metrics. The results showed that the k-nearest neighbor regression model achieves the highest performance with an R^2 score of 0.9715.

Keywords: Power prediction · Machine learning regression · Photovoltaic

1 Introduction

In recent years, renewable energy sources such as solar power have become more popular globally due to advances in photovoltaic (PV) technology. However, there are concerns about the reliability of these energy sources as they are affected by variables such as weather, seasonality, and production patterns. In order to ensure a stable solar power sector, it is important to accurately forecast solar energy production. This is important because as the world transitions to renewable energy sources, such as solar power, the accurate estimation of solar energy production becomes critical for ensuring the stability of electrical networks. Inaccurate predictions of solar energy production can lead to overloading of the electrical grid or, conversely, underutilization of the solar energy that is available, which can result in higher energy costs and grid instability.

A. Ortis et al. (Eds.): ICAETA 2023, CCIS 1983, pp. 204–213, 2024.
https://doi.org/10.1007/978-3-031-50920-9_16

The use of machine learning regression algorithms provides a reliable and accurate way to estimate solar energy production to increase the adoption of renewable energy sources, which is critical to reducing greenhouse gas emissions and mitigating the effects of climate change. In addition, accurate estimation of solar energy production can help grid operators, distributed energy resource (DER) aggregators, and PV power plant owners make informed decisions about energy storage and distribution, improving the efficiency of the electrical grid.

Li et al. [1] performed Artificial Neural Network (ANN) and Support Vector Regression (SVR) models to predict PV power production for 15 min, 1 h, and 24 h in advance using historical production data from online meteorological services. They used one year of data and converted the historical production data to 15 min and 1-h average values. In their approach, a hierarchical methodology was followed, in which forecasts were made for each inverter separately based on the historical data for that inverter, and a forecast for the entire plant was also made. They found that forecasting production for each inverter individually resulted in more accurate results. Theocharides et al. [2] applied ANN, SVR, and Random Tree (RT) models to predict day-ahead hourly power production for PV systems from historical PV production data, incident global irradiance (GI), and ambient temperature (Tamb) data. Their experimental results showed that the ANN model performed better than the other model. In another study published by Theocharides et al. [3], ANN, K-means clustering, and linear regressive correction models are applied to predict day-ahead hourly power production for PV systems using historical power production data, wind direction (Wa), ambient temperature (Tamb), incident global irradiance (GI), wind speed (Ws), relative humidity (RH), solar azimuth (φs) and elevation (α) angles data. Their method achieved a MAPE of 4.7%. Leone et al. [4] applied to an SVR model to predict day-ahead production at 15 min interval using solar irradiance, ambient temperature, and historical production data. Their model achieved an R2 value exceeding 90%. In a study published by Khandakar et al. [5], historical production data, ambient temperature, dust accumulation, wind speed, solar irradiance, relative humidity, and panel temperature data were applied to an ANN model, linear regression, M5P tree model, and gaussian process regression to predict hourly PV power output. The interval of collected data was not specified. They stated that the ANN model outperformed the other methods and achieved an RMSE of 2.1436. Qu et al. [10] proposed a prediction model called ALSM that uses a combination of CNNs, LSTMs, and an attention mechanism to forecast solar power output over multiple relevant and target variables. This model takes into account a variety of inputs, including historic PV output power, latitude, longitude, array rating, and other geographic data, to capture both short-term and long-term temporal patterns and provide hourly forecasts for the coming day. The model is designed to operate under the multiple relevant and target variables prediction pattern (MRTPP). Visser et al. [11] used historical weather and PV output power data to assess the efficacy of 12 alternative approaches that forecast day-ahead power production based on market circumstances. SVR, deep learning, physical-based techniques, and ensemble learning were among the models used. They also evaluated the effect of aggregating numerous PV systems with different inter-system distances on the forecasting models' efficacy. The models were assessed on their technical and economic performance. Eniola et al. [12] developed a model validation method using

more recent input datasets, including temperature, mod temperature, historical production data, wind speed, and solar irradiance, based on an existing prediction model built on a genetic algorithm (GA)-optimized hidden Markov model (HMM). Normalized root mean square error was considered as an evaluation method for the models (nRMSE). Mahmud et al. [13] used various machine learning algorithms to perform short-term and long-term PV output power prediction. They found that random forest regression model outperformed other machine learning algorithms on their dataset that was collected from Alice Springs, which is one of the areas of high PV power generation in Australia. In [14], Mellit et al. predicted short-term PV output power using several kinds of deep learning neural networks. The data used in [14] was gathered from a microgrid in a university in Italy. They found that the case of 1-min with one-step ahead achieved highest accuracy scores, but up to 8 steps ahead gives acceptable results.

In this study, accurate prediction of PV power production predicted for 1 month with 1-h resolution using RT, KNN, SVR, and ANN regression methods. A grid search algorithm was applied to find optimal hyperparameters for the machine learning methods used in this work. In the literature, day-ahead PV production forecast is common (e.g. [3, 10], and [11]). Although day-ahead prediction is important, accurate monthly prediction gives a wider view and better insight for monthly dispatch planning and can be considered a step forward towards managing PV plants as conventional dispatchable power plants. This will be possible because accurate and reliable production values will be used for solving the unit commitment and economic dispatch problems. The significance of this prediction is that it allows long-term (1 month) optimal dispatch of generation and storage assets which helps in maintaining grid resiliency for systems with high PV penetration. Without accurate predictions, it is very hard to optimally dispatch generation units and maintain grid stability in the case of high PV penetration.

This paper is organized as follows: Sect. 2 describes the machine learning methods used in this work, Sect. 3 introduces the dataset and presents the results and discussion, and Sect. 4 concludes the paper.

2 Machine Learning Regression Methods

In this study, support vector regression, k-nearest neighbor, decision tree, and artificial neural networks methods were applied to predict PV output power.

2.1 Support Vector Regression

Support vector machines (SVMs) are statistical learning techniques that are frequently applied to solve regression and classification problems. In SVR, a dataset is first transformed into a high-dimensional space, and a curve is fitted to the data using a "cylinder" that is defined by support vectors, which are the points that determine borders of the cylinder. SVR model estimates the relationships between the inputs and outputs utilizing Eq. 1:

$$f(x) = \omega\varphi(x) + b \tag{1}$$

where $\varphi(x)$ is the transfer function that maps the input data to high-dimensional feature spaces. The regularized risk function is minimized to estimate the parameters ω and b:

$$\begin{gathered} \min \tfrac{1}{2}\omega^T * \omega + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ s.t. \\ y_i - \omega^T \varphi(x_i) - b \leq \varepsilon + \xi_i \\ \omega^T \varphi(x_i) - b - y_j \leq \varepsilon + \xi_i^* \end{gathered} \tag{2}$$

In Eq. 2, n is the number of samples used for training, ξ represents the error slacks that ensure the results are within particular tolerances, C denotes a regularization penalty, and ε is the tube's target tolerance range. $\omega^T * \omega$, the first term in the equation, is a regularization term that aids in flattening the curve. The second term is a determined empirical error with ε-insensitive loss function. The loss function being described here measures the difference between expected values and the radius of a cylinder. If the anticipated values are within the cylinder, the loss is 0. If the anticipated values are outside of the cylinder, the loss is equal to the absolute difference between the expected values and the radius of the cylinder ε. This loss function may be used to evaluate the accuracy of predictions made by a machine learning model. The model's goal would be to minimize the loss by making more accurate predictions. The Lagrange multiplier is used to optimize both ε and C, and the corresponding Lagrangian structure of Eq. (2) can be stated as the following equation, where $K(x_i, x_j)$ represents a kernel function:

$$f(x) = \sum\nolimits_{i=1}^{n} (a_i - a_i^*) K(x_i, x_j) + b \tag{3}$$

Equation 3 defines the Lagrange multipliers a_i and a_i*, which are obtained by solving the dual version of Eq. (2) in Lagrange structure. The use of a kernel function has the advantage of allowing us to work with feature spaces of any size without having to manually construct the map $\varphi(x)$. Any function that meets Mercer's criteria, such as a polynomial or radial basis function (RBF) kernel [1, 7], can be employed as a kernel function. SVR model with RBF kernel was used in this study.

2.2 Regression Trees

The regression tree approach is a method for constructing a predictive model by dividing a dataset into smaller divisions and fitting a simple prediction model to each partition. The Analysis of Variance (ANOVA) method is used to assess differences or variations between the partitions. To build a numeric prediction regression tree (RT), the dataset is first partitioned at the root node using a decision tree induction algorithm based on the feature that maximizes the gain in homogeneity in the outcome after the split. The tree-growing method assesses homogeneity, which is often measured using statistics such as absolute deviation, variance, and standard deviation from the mean. The standard deviation reduction (SDR) is a common criterion for determining the split, and it is defined as follows:

$$SDR = sd(T) - \sum\nolimits_i \frac{|T_i|}{|T|} * sd(T_i) \tag{4}$$

The $sd(T)$ function in this equation representing standard deviation of the value of samples in a dataset T, sets of values obtained from a feature split are represented as T_i. The number of observations in the dataset T is represented by the $|T|$ symbol. The splitting criterion is used to measure the decrease in standard deviation from the original value to the weighted standard deviation after the data has been split [2].

2.3 k-Nearest Neighbor

The k-nearest neighbor (KNN) is an approach that uses a predictor variable X to estimate the conditional distribution of a response variable Y and allocates Y to the class with the highest estimated probability. To categorize a new test observation x_0, the KNN method finds the K points in the training data that are closest to x_0 (using the Euclidean distance) and are represented by N_0. The conditional empirical distribution for class j is then calculated as the ratio of the K nearest points categorized as j:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum\nolimits_{i \in N_0} I(y_i = j) \tag{5}$$

Finally, the class j with the highest estimated probability receives x_0. . It is vital to notice that the value of k influences the KNN classifier significantly. When K is set to one, the decision boundary will overfit the training data, producing a classifier with low bias but large variance. The decision boundary gets more linear as k grows (i.e., low variance but high bias). The bias-variance trade-off is affected by k, which should be considered [9].

KNN can be used to solve regression problems as well. The KNN algorithm in this situation selects the K-nearest neighbors based on some distance metric and assigns the average value of those neighbors as the forecast. The forecast is expressed as follows:

$$\widehat{Y} = \frac{1}{K} \sum\nolimits_{i \in N_0} y_i \tag{6}$$

2.4 Artificial Neural Networks

Artificial neural networks (ANNs) are computer systems that are designed to mimic the way the human brain works. They are composed of interconnected "neurons" that can process and transmit information. ANNs are commonly used in machine learning and artificial intelligence applications, and they can be trained to perform a variety of tasks by being exposed to large amounts of data and adjusting the strengths of the connections between neurons. ANN can be used to represent complex functions and solve real-world issues. To solve challenging nonlinear problems, ANNs employ a network of artificial neurons or nodes. A typical artificial neuron can be represented using a function that processes n input values (also known as "dendrites") to produce a single output value (also known as the "axon"). This function typically combines linear and non-linear operations to weight, sum, and transform the input values in some way. The specific form of the function will depend on the design of the neural network, but it may involve matrix

multiplications, activation functions, convolutional filters, and/or pooling operations. A typical artificial neuron with n dendrites can be represented as follows:

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{7}$$

The weights w_i in this equation allow each of the n input variables x to contribute to the total input signals. The activation function f(x) takes the net sum of these input signals and generates the output signal y(x), which is represented by the output axon.

3 Results

In this section, we evaluate results of the machine learning regression methods to predict PV output power. First, we explain our dataset, second, we give performance metrics used in this study and finally we discuss the results.

3.1 Dataset

The dataset was collected from a solar power plant in Konya province in Türkiye and includes hourly measurements of solar production and various meteorological parameters such as incident global irradiance (G_I), ambient temperature (T_{amb}), wind speed (W_s), and mod temperature (T_m) for the period of January 1, 2021, to December 31, 2021. The purpose of the study is to use these features to train machine learning models to accurately forecast solar energy production on an hourly basis, with the goal of improving the reliability of solar power as a renewable energy source. The dataset used in this work is original and has not been used in any previous research.

To evaluate and optimize the performance of the machine learning models, the dataset was split into 11 months of training data and 1 month of test data, and k-fold cross-validation was applied during the hyperparameter optimization process. The value of k was chosen as 12, meaning that the data was split into 12 folds and the model was trained and tested 12 times, each time using a different fold as the test set. The performance of the models was then evaluated using the RMSE, MAE, and R^2 evaluation metrics, and the results showed that the k-nearest neighbor regression model achieved the highest performance with an R^2 score of 0.9715.

3.2 Performance Metrics

In this study, the prediction models' performance was assessed using the following metrics:

- Mean absolute error (MAE) (given in Eq. 8): This is a measurement of the average difference between actual and forecasted data.

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |x_i - y_i| \tag{8}$$

- Root mean square error (RMSE) (given in Eq. 9): The standard deviation of the prediction errors is described by the root mean square error.

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (x_i - y_i)} \tag{9}$$

- Coefficient of determination (R2) (given in Eq. 10): A measure of the fraction of data variability described by the model, ranging from 0 to 1. A number of 0 indicates that the model does not describe the data at all, whereas a value of 1 show that the model explains the data correctly.

$$R^2 = 1 - \frac{RSS}{TSS} \tag{10}$$

3.3 Experimental Setup

The computer used for this work has an Intel Core i5-7200U CPU @ 2.50 GHz 2.70 GHz processor, 8 GB RAM, and 64-bit operating system. Version 5.2.2 of Spyder python development environment is used with Python version 3.9.15. All machine learning models used are from scikit-learn library version 1.0.2.

3.4 Experimental Results

In this work, the results of RT, KNN, SVR, and ANN to predict PV output power were evaluated. We performed grid search method with k-fold cross validation to determine the best hyperparameters of the machine learning methods. In the grid search approach, for the KNN regression model, K values from 1 to 29 were chosen. For the ANN model, 3 options were considered for hidden layers number and sizes. These options are (50,50,50), (50,100,50), and (100), which means 3 layers each containing 50 neurons, 3 layers with 50, 100, and 50 neurons, and a single layer with 100 neurons. Three values were also considered for learning rate, and alpha hyperparameters which are: 0.1, 0.01, and 0.001. Finally, for the SVR model, RBF kernel was used. C and epsilon hyperparameters were chosen as 1, 10, 100, 1000, and 0.01, 0.1, 1, 10 respectively. We note that hyperparameter optimization was applied to all models except RT model, which shows promising results even without optimization.

The dataset was split into 11 and 1 months, 11 months were used for training while 1 month was used to test the model. Results are shown in Table 1. In RT method, minimum samples split is set to 2, minimum samples leaf is set to 1. The RT method achieves an RMSE of 157.13, a MAE of 58.78, and R^2 of 0.96. Prediction results of the RT are shown in Fig. 1. In SVR method, RBF kernel is used. C parameter is set to 100, and epsilon parameter is set 10. The SVR method achieves an RMSE 164.64, MAE of 63.99, and R^2 of 0.95. The prediction results of SVR are shown in Fig. 2. In the KNN method, k is chosen as 29 and Euclidean metric is used. The KNN achieves an RMSE of 137.77, a MAE of 53.95, and an R' of 0.9715. The KNN prediction results are shown in Fig. 3. Finally, in the ANN method, the number of hidden layers is set to 50 and 50 neurons are used in each hidden layer. The learning rate is set to 0.001, and alpha parameter is set to 0.1. The ANN method achieves RMSE of 148.82, MAE

of 68.34, and R^2 of 0.96. The prediction results of the ANN are shown in Fig. 4. It is also worth mentioning that hyperparameter optimization was not performed for the RT model because of the high computation complexity which causes very long convergence time. Values of performance metrics shown in Table 1 indicates the high precision of the prediction curves observed in Figs. 1, 2, 3 and 4 quantitatively. It can be clearly seen that reliable predictions are generated, which can be used by grid operators, DER aggregators, and PV power plant owners to optimally dispatch assets, since PV output power is predicted with acceptable tolerance.

Table 1. Results of machine learning methods

Heading level	RMSE	MAE	R^2
RT	157.13	58.78	0.9629
SVR	164.64	63.99	0.9593
KNN	137.77	53.95	0.9715
ANN	148.82	68.34	0.9668

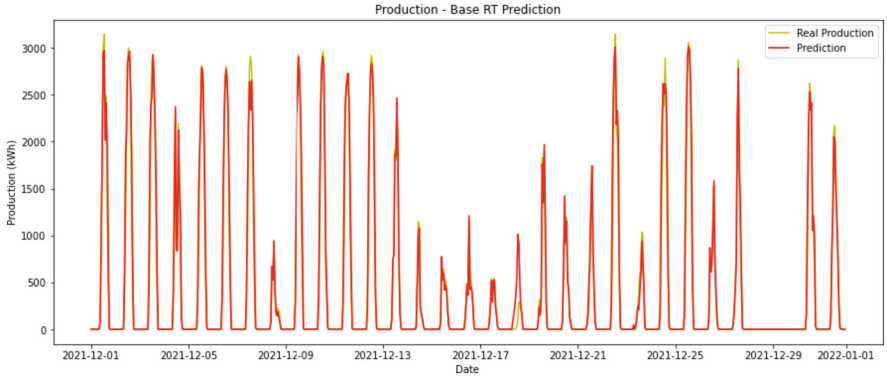

Fig. 1. Prediction results of the RT method

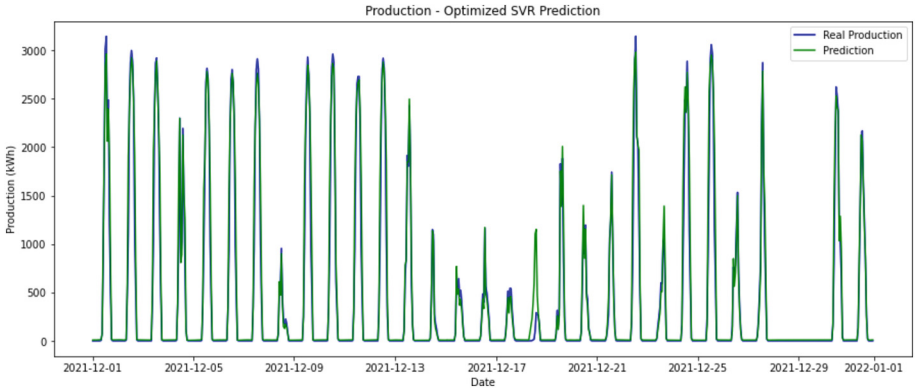

Fig. 2. Prediction results of the SVR method

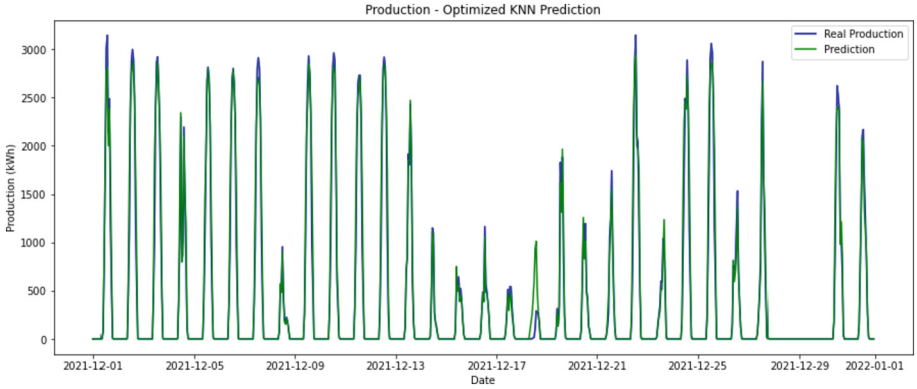

Fig. 3. Prediction results of the KNN method

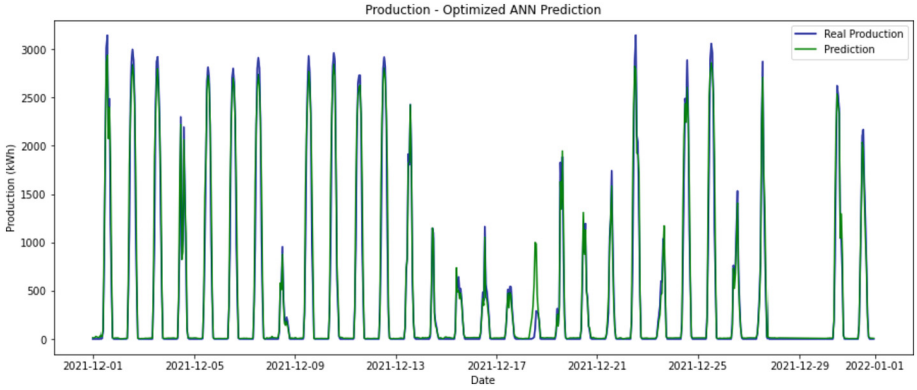

Fig. 4. Prediction results of the ANN method

4 Conclusion

The problem addressed in this study was the need for accurate forecasting of solar power production in order to ensure the stability of electrical networks as the adoption of renewable energy sources such as solar power increases. The solution proposed was the use of machine learning algorithms, including ANN, SVR, RT, and KNN regression, to forecast solar energy production using historical production data and various meteorological parameters. The models were optimized using grid search and validated using the K-fold cross validation method. The results of the study present that the KNN regression model was the most accurate results with an R^2 score of 0.97. These results suggest that machine learning techniques can be effectively used to predict solar energy production and contribute to the stability of electrical networks as the use of renewable energy sources increases. In future studies, deep learning algorithms can also be applied in addition to parallel computing in order to improve the computation speed for the suggested models.

References

1. Li, Z., et al.: A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. Energies **9**(1), 55 (2016)
2. Theocharides, S., et al.: Machine learning algorithms for photovoltaic system power output prediction. In: 2018 IEEE International Energy Conference (ENERGYCON). IEEE (2018)
3. Theocharides, S., et al.: Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. Appl. Energy **268**, 115023 (2020)
4. De Leone, R., Pietrini, M., Giovannelli, A.: Photovoltaic energy production forecast using support vector regression. Neural Comput. Appl.. **26**(8), 1955–1962 (2015)
5. Khandakar, A., et al.: Machine learning based photovoltaics (PV) power prediction using different environmental parameters of Qatar. Energies **12**(14), 2782 (2019)
6. Deng, F., et al.: Prediction of solar radiation resources in China using the LS-SVM algorithms. In: 2010 the 2nd International Conference on Computer and Automation Engineering (ICCAE), vol. 5. IEEE (2010)
7. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1999)
8. Almeida, M.P., Perpinan, O., Narvarte, L.: PV power forecast using a nonparametric PV model. Sol. Energy **115**, 354–368 (2015)
9. Isaksson, E., Karpe Conde, M.: Solar power forecasting with machine learning techniques (2018)
10. Qu, J., Qian, Z., Pei, Y.: Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. Energy **232**, 120996 (2021)
11. Visser, L., AlSkaif, T., van Sark, W.: Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution. Renewable Energy **183**, 267–282 (2022)
12. Eniola, V., et al.: Validation of genetic algorithm optimized hidden Markov model for short-term photovoltaic power prediction. Int. J. Renewable Energy Resour. **11**(2), 796–807 (2021)
13. Karim, A., et al.: Machine Learning Based PV Power Generation Forecasting in Alice Springs
14. Mellit, A., Pavan, A.M., Lughi, V.: Deep learning neural networks for short-term photovoltaic power forecasting. Renewable Energy **172**, 276–288 (2021)