International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

A Comprehensive Analysis of Machine Learning for Forecasting PV System

Surya Prakash Chaudhary, Dr. Dinesh Kumar Rao, Kriti Srivastava
Department of Mechanical Engineering
Institute of Engineering and Technology, DRMLAU, Ayodhya, U.P., India
ajayatsky@gmail.com

Abstract: Due to the growing demand, assessing performance has become obligatory for photovoltaic (PV) energy harvesting systems. Performance assessment involves estimating different PV system parameters. Traditional ways, such as calculating solar radiation using satellite data and the IV characteristics approach as assessment methods, are no longer reliable enough to provide a reasonable projection of PV system parameters. Estimating system parameters using machine learning (ML) approaches has become a reliable and popular method because of its speed and accuracy. This paper systematically reviewed ML-based PV parameter estimation studies published in the last three years (2020 – 2022). Studies were analyzed using several criteria, including ML algorithm, outcome, experimental setup, sample data size, and error metric. The analysis revealed several interesting factors. The neural network was the most popular ML method (32.55%), followed by random vector functional link (13.95%) and support vector machine (9.30%). Dataset was sourced from hardware tests and computer-based simulations: 66% of the studies used data from only computer simulation, 18% used data from only hardware setup, and the 16% experiments used data from both hardware and simulations to evaluate different system parameters. The top three most commonly used error metrics were root mean square error (29.1%), mean absolute error (17.5%), and coefficient of determination (15.9%). Our systematic review will help researchers assess ML algorithms' projection in PV system parameters estimation. Consequently, scopes shall be created to establish more robust governmental frameworks, expand private financing in the PV industry, and optimize PV system parameters.

Keywords: Photovoltaics, System parameter estimation, Machine learning, Systematic review.

1. Introduction

Energy harvesting systems, such as coal-fired and nuclear power plants, are among the most polluting elements of our environment. Scientists and researchers are working hard to create a clean energy-harvesting environment. Hydroelectric power plants, wind turbines, and photovoltaics are some of the best-known alternatives to those non-renewable energy sources. In photovoltaic systems, electricity is produced utilizing the light and the heat generated from the sun. Solar energy is so common because it has no resource cost and is known as an illimitable energy source [1]. As a result, energy consumers are rapidly being adapted to solar-based power systems. For example, in 2020, almost 12.30% of total renewable energy came from solar energy in South Korea, whereas in 2012, it was only about 3.04% [2]. Due to the solar orbital motion, the sun cannot provide the same irradiance throughout the whole world at the same time [1,3]. Sometimes it provides the peak irradiance and sometimes averages throughout the day. In our modern industrial world, producing maximum power in the shortest amount of time is crucial. To harvest maximum power by forecasting and analyzing photovoltaics (PV) performance, reliable solar cell modeling is a critical factor to consider [4]. Complex machine learning (ML) models can predict a PV system's output current-voltage (I-V) and power-voltage (P-V) parameters with very high accuracy [5]. Because of its low-cost setup and reliability, its usage in grid distribution networks is also increasing day by day. Due to these facilities, the government subsidizes and provides frameworks that ultimately accelerate the implementation and scalability of PV systems [6,7]. However, for all of this to happen, there is always a concern about their return on investment, whether it is worth their time and effort, which ultimately necessitates predicting PV performance through system parameters. IV curve comparison, various diode models, and thermal models are some of the traditional ways to predict the PV system parameters. However, these models are very complex to implement or need better accuracy [8]. To overcome the limitations of these traditional models, ML has become one of the best alternatives because of its speed and accuracy. Also, collecting recent works and analyzing their findings would be a matter of toil and trouble for anyone interested in working and contributing to the photovoltaic industry. Considering all these factors, we have compared and analyzed the literature on PV system performance prediction based on different ML algorithms. Mainly there are two types of prediction for PV system characterization: direct and indirect. A direct prediction system predicts PV power by training an ML model using existing PV power data. On the other hand, indirect prediction estimates the performance depending on system parameters like solar irradiance and temperature, which is not directly related to PV power parameter. In our review paper, we considered both prediction types to keep the comparison fair and balanced. ML performance is generally quantified in terms of error metrics,

such as root mean square error (RMSE), mean absolute error (MAE), and mean relative error (MRE). Apart from these metrics, while reviewing the articles, we have analyzed the performance based on several other metrics like mean absolute percentage error (MAPE) and root relative squared error (RRSE). Our work was further motivated by the absence of a comprehensive review of the latest advancements in predicting parameters for the PV systems. The latest SLR on PV performance using ML was published in early 2021 [9]. They cited only four reference works from 2020, which is logical as they completed the review in 2020. Sobri et al. [10] conducted a comprehensive analysis of various techniques including statistical methods based on time-series data, physical models, and ensemble approaches to forecast parameters of photovoltaic (PV) systems. Their findings indicated that the utilization of artificial intelligence models had the potential to significantly reduce errors in comparison to conventional approaches. Antonanzas et al. [11] stated that the uncertainty in available solar resources challenges the reliable prediction system. In the case of economic analyses, researchers mainly focused on probabilistic prediction rather than regression analysis. Even statistical techniques were found to outperform traditional parametric techniques. Because of the convenience of the modeling process, the most recent articles in their work preferred to adopt the ML technique, which enhanced the prediction performance. Similarly, it has also been shown that convolutional neural network (CNN) performs the best in collaboration with other ML methods when they are used for short-term forecasting purposes [12]. Therefore, these works show that researchers have been using ML methods to predict PV performance for a long time. However, literature review papers on these works are either not recent or did not follow any systematic approach. We, therefore, present a comprehensive review paper on this topic for the articles published from 2020 to 2022. We present the status of current progress in PV research in terms of ML-based algorithms, how much more efficient they are compared to the old PV parameter prediction methods, and what is left unexplored. In terms of its contribution, this paper not only saves substantial time in the search for papers on machine learning-based prediction of PV system parameters but also serves as an initial reference for individuals embarking on their exploration of PV systems.

2. Necessity of ML-based PV parameter prediction

A microgrid is a modern approach that combines electricity from distributed power generation sources and diverse modules for storing energy, serving local electricity demands. An energy-storing facility creates a symbiotic relationship between conventional and renewable energy sources. Because of its independent functioning in grid-connected mode, the microgrid system is favored over conventional electricity distribution approaches. The efficiency of renewable energy systems plays a vital role in the functioning of a microgrid. However, intermittent and unpredictable sources, such as PV modules, provide difficulties in demand, supply, and

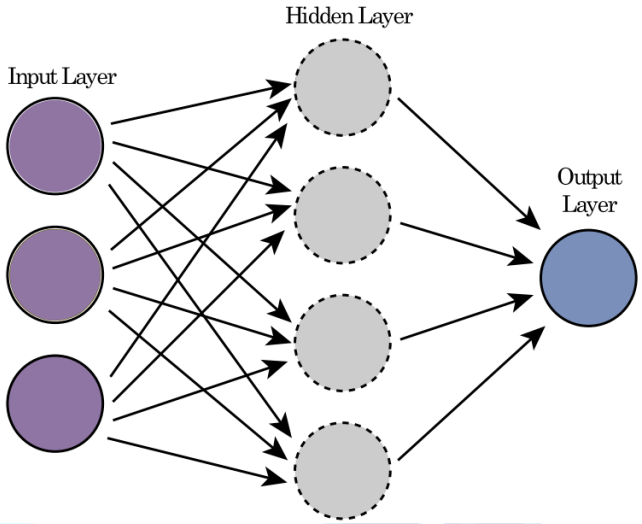

Fig. 1. Neural Network block diagram

operation. Environmental elements like solar irradiance depend on other meteorological factors, like weather systems, which include humidity, air, and other parameters. It eventually affects the functioning of the microgrid as a whole. However, to fully leverage the benefits of decentralized electricity generation, it is crucial to maintain a balanced equilibrium between electricity generation and demand. In this case, ML models can work as reliable sources to forecast solar irradiance performance, allowing system operators to plan improved scheduling, energy distribution, proper maintenance, an uninterrupted power system, and other operational features [13]. The ESS is one of the most reliable and consistent energy sources during peak hours while saving energy during off-peak hours. Photovoltaic systems are a common source of energy for the ESS system. However, because of the volatility of these photovoltaic sources induced by the varying solar irradiation, it is critical to anticipate output power to ensure an accurate estimate of loads and power generation. Rajamand et al. [14] optimizes ESS size and location by combining ANN with MLP and GNN. It resulted in up to 31% cost savings because of PV power prediction with an optional ESS deployment. So, another important aspect of accurate PV power prediction is minimizing the microgrid system's production and distribution costs. Several solutions existed before the advent of machine learning technologies to justify PV performance. Methods like the ideal model, the four-parameter model, the single-diode or five-parameter model, and the double-diode or seven-parameter model are examples of non-machine learning approaches [15]. These models rely on a few distinct PV module-related characteristics and a few meteorological attributes. However, in most cases, a PV module's performance and quality are determined by four factors that are influenced by the intensity of solar irradiation and the module's temperature [16]. Generally speaking, there are two ways to assess the performance of these models. The first technique is computing the instantaneous peak power under certain circumstances. The second method involved regression analysis, which used long-term data supplied by existing PV modules. The first approach

is tested under standard test conditions, whereas the second is tested under utility-scale and normal cell temperature test conditions. The standard tests are commonly carried out under specific parameters, including an air mass equivalent to 1.5, a cell temperature maintained at 25 degrees Celsius, and an irradiation spectrum of 1000 $W/m2$. On the other hand, the second testing method is carried out at 20 degrees Celsius under 1000 $W/m2$ of illumination [17]. However, maintaining these standards in real life is quite challenging. As a result, the test findings differ and are falsified. Another frequent issue is the second test's requirement for a significant amount of data before the regression analysis [18]. Also, the limited computation capacity hampers the testing procedure. Therefore, it might result in inaccurate output due to data unavailability and a lack of processing resources. Various models concentrate on the thermal characteristics of the PV modules in addition to these electrical equivalent models. While some models are based on thermal capacitance [19], others are based on the total heat loss coefficient [20,19]. However, as the manufacturers do not offer adequate details about these features, these models are not realistic to use. Haouari-Merbah et al. [21] proposed a novel model that effectively captures the IV curve by delineating it into two distinct regions, facilitating the extraction of the physical parameters. But rather than offering comprehensive statistics, this model provides data on three PV parameters.

3. Machine learning models

3.1. Neural network

A neural network (NN) replicates the functioning of neurons in the human brain and lies at the heart of pattern recognition. It consists of three essential elements: an input layer, one or more hidden layers, and an output layer (Fig. 1). These layers consist of neurons having network parameters (weights and biases) [22]. All the data is fed into the input layer. The network can incorporate one or multiple hidden layers followed by activation functions to handle and analyze the input data. Finally, the output layer generates predictions based on the input data provided.

3.1.1. Advantages

NN can quickly handle problems with uncertain behavior or structure using its non-linear activation functions. It is well recognized for its adoptable mechanism; that is, it can change its structure depending on the purpose of its usage. It is possible by taking full advantage of the cognitive abilities that lie within its algorithm. The input data that pass through the network of a NN determines how it modifies its pattern. Because of its non-linear activation function, it can work with data of any dimension as long as the input is a continuously differentiable function.

3.1.2. Disadvantages

The primary drawback of NN is that it demands a lot of computer resources due to the vast amount of input data requirements. To improve prediction performance, a large amount of training data is needed. Another issue is that it is susceptible to the initial randomization of network parameters. Furthermore, the rate of processing time also increases exponentially as the number of hidden layers increases.

3.2. Decision Tree

As shown in Fig. 2, a decision tree (DT) initiates by establishing a root node that branches out into multiple child nodes. These child nodes encompass both leaf nodes and decision nodes. While the root node itself is a decision node, the algorithm restricts the presence of only one root node throughout its entire operation. However, in the case of a child decision node, it can have more child nodes, including leaf nodes and other decision nodes. Leaf nodes are not divisible further. They are the output for a particular decision case. Also, the whole evaluation consists of all sorts of nodes and is called a tree. However, within a tree, it is possible to have multiple sub-trees consisting of leaf and decision nodes.

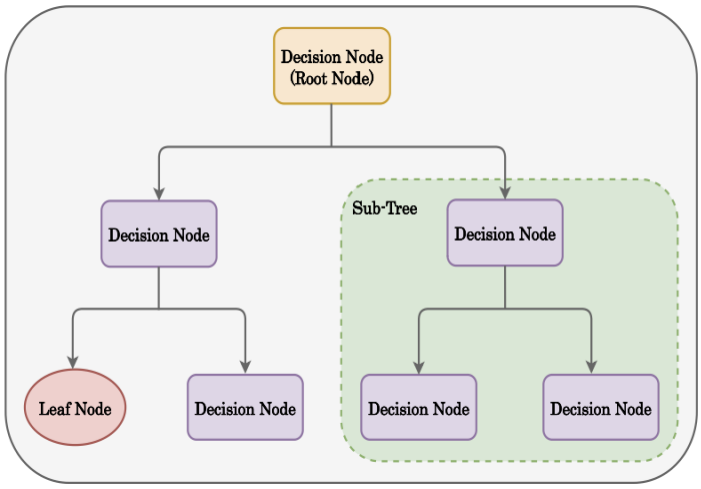

Fig. 2. Decision Tree block diagram.

3.2.1. Advantages

The first benefit of a decision tree is that it can solve both regression and classification tasks. Next, neither standardization nor normalization is necessary for it. It is built on a rule-based methodology as opposed to computing data distance. Additionally, scaling of data features is not necessary. Contrary to curve-based algorithms, the DT method is unaffected by non-linear parameters. Other benefits include: (1) non-parametric behavior, (2) excellent efficiency due to the tree traversal technique, and (3) the ability to fill in missing values with the best suitable one.

3.2.2. Disadvantages

The primary issue with decision trees is overfitting. The outcome is anticipated inaccurately as a result. Sometimes, when the algorithm tries to fit the data, it keeps producing new nodes after each iteration, which makes the method harder to comprehend. Additionally, because of the volatility of the testing data, data overfitting might result in a substantial degree

of inaccuracy. Furthermore, data containing many features may delay the prediction and make the system less effective.

3.3. Support vector machine

The support vector machine (SVM) algorithm is used to classify data based on different features available for a particular dataset. It generates a better output in the case of classification problems when most of the features are categorical. It also aids in determining the best fit line between various classes. These lines are called hyperplanes, as shown in Fig. 3. Hyperplanes are created with the vectors of a particular class located at the edge. For these reasons, these are called support vectors. The maximum margin between two hyperplanes is called the optimal hyperplane. In SVM, the target remains to maximize this optimal hyperplane. It ensures that all the classes are differentiable enough to make a reliable prediction.

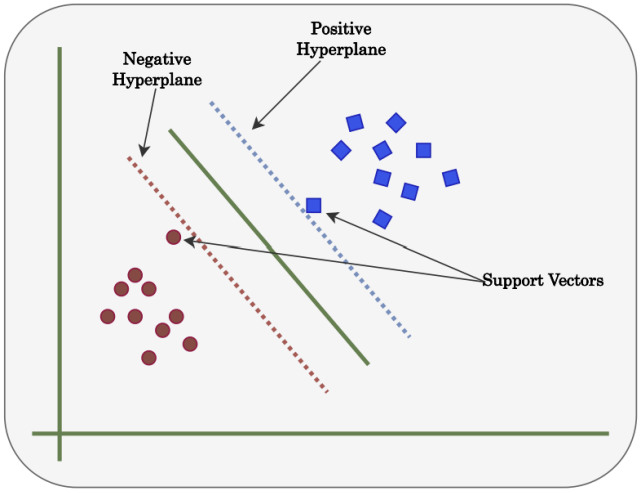

Fig. 3. Support Vector Machine block diagram.

3.3.1. Advantages

The SVM provides several advantages. Firstly, it is capable of addressing both classification and regression tasks, allowing it to effectively handle diverse types of data, including structured and semi-structured datasets. Kernel functions are a very important aspect of SVM. These are a bunch of mathematical functions defined by different SVM algorithms. Multiple varieties of kernel functions exist, including linear, nonlinear, and sigmoid options. If the kernel function is appropriately developed, it can anticipate extremely complicated data thanks to its ability to use that function as a kernel. When the number of samples is less than the number of dimensions, it demonstrates greater efficacy when operating within high-dimensional spaces. Additionally, it has a lower chance of overfitting due to its universality. Thus, it avoids becoming entangled in local optima.

3.3.2. Disadvantages

When the data collection is quite extensive, one of the disadvantages is that the performance diminishes with time. Additionally, it has inconsistent performance with noisy data. Following that, selecting a suitable kernel function is a challenging and timeconsuming procedure. Unlike DT, the SVM algorithm is very complex. For this reason, for datasets with lots of features, sometimes it becomes challenging to interpret the outcome. Lastly, the performance of SVM deteriorates when the number of features surpasses the number of training instances.

3.4. Long short-term memory

The recurrent neural network (RNN) family of models involve feedback mechanisms to process sequential data. A type of RNN is long short-term memory (LSTM) that tackles the challenge of vanishing gradients, specifically when dealing with larger datasets [23]. A RNN can predict with higher accuracy for the recently processed information. However, in the case of LSTM, it can retain information for a longer time passed through this model's memory cell. Because of its capacity to remember prior knowledge, LSTM can predict significantly quicker with higher precision [24]. The LSTM memory block is small in size but retains the memory for a more extended period. There are three gates that control the memory block: the forget gate, input gate, and output gate (Fig. 4). The forget gate is responsible for deciding which information will be retained and processed further. If it is not required, that particular information is not fed into the input gate. If it is required, the important pieces of information are fed into the input gate. In this gate, relevant information is introduced and combined with the cell state. In the output gate, the regulated data from the input gate is multiplied with the cell's *tanh* generated vector to extract useful information.

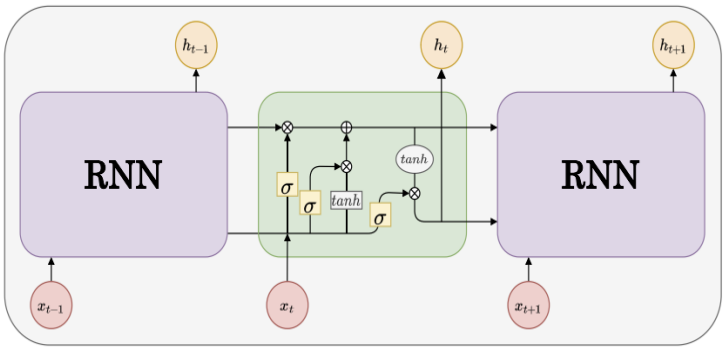

Fig. 4. Long-Short Term Memory block diagram

3.4.1. Advantages

The primary benefit of LSTM is its capacity for longer-term knowledge retention. The LSTM may even filter redundant data that is no longer needed with its forget gate. It fills in the gaps left by its parent RNN model by offering extra variables, including learning rates, input biases, and output biases. The capacity to manage noise, the need for zero-fine adjustments, and distributed representations with four interacting layers are additional characteristics of LSTM.

3.4.2. Disadvantages

Due to the linear layers included in its cells, the first drawback of LSTM is the need for significant computational resources to prepare it for usage in production. Similar to feedforward NN, LSTM also has problems with random weight initialization. Over time, it tends to get overfitted for a more considerable volume of data. Additionally, each of its cells becomes more complicated due to the so-called forget gates, and the model itself cannot eliminate redundant gradients between its previous and latter cells.

5. Review:

It is evident from the NN constitutes the largest parent group (32.55%), comprising several extensively employed methods for ML among the papers we examined. The multilayer perceptron (MLP) is used more often than the other NN methods in applications like solar system prediction. Our findings show that MLP holds the highest usage percentage among all the NN models. It is highly prevalent because of its pattern extraction capability. Unlike other ML algorithms, MLP does not need many feature extraction strategies. It automatically learns the extraction strategies from the supplied data. The second most frequently used ML parent method is random vector functional link (RVFL) (13.95%). Unlike in MLP, where all input parameters are optimized during the training process, in RVFL, only the output weights are adapted. The rest of the parameters, like input nodes, hidden weights, and thresholds, are constrained and predicted randomly in advance. This feature minimizes the complexity of dimensionality while predicting. RVFL randomly calculates weights and biases of the input data, which is incredibly quick compared to older approaches like IV curve comparison and different diode models [29]. RVFL is the second most popular ML method because of different researchers' use of different optimizers. The PV system has many linear and non-linear characteristics. Due to these characteristics, SVM is a popular ML model for predicting these system parameters. For this reason, it occupies the third most frequently used ML parent method (9.30%), covering least squares SVM (LSSVM), support vector regression (SVR), SVR-radial basis function (SVR-RBF), and SVM itself. As the fourth widely used ML algorithm, LSTM incorporates three subsidiary algorithms, namely attention LSTM, bidirectional (Bi) LSTM, and LSTM itself. As LSTM can retain information for longer, it has been used to predict PV system parameters. It is because the PV system usually consists of several vital parameters that must be retained for further performance prediction. Similarly, adaptive neuro-fuzzy inference system (ANFIS) and DT have two child ML models, each with almost similar popularity. The rest of the ML models and their usage frequency. They presented the top ten ML methods resulting from a narrower analysis. They may seem to present similar information. The former describes their usage frequency based on both the parent methods and child methods, whereas the latter depicts the usage frequency of the top ten methods regardless of their parent category. BPNNs and GNNs are part of ANNs; the difference is whether they are trained with backpropagation algorithm (BPNN) or genetic algorithm (GNN). They followed the same naming conventions used by the respective authors and showed them separately because the difference was unclear from the articles. Artificial neural network (ANN) holds 27.8% of the use cases while evaluating the performance, followed by back propagation neural network (BPNN) with 13.9%, genetic neural network (GNN) with 9.72%, and LSTM with 8.33%. Other ML models like auto regressive moving average (ARMA) and random forest (RF) have also been used on a small scale. For example, Shahsavar et al. [30] found RF to perform better than other major ML models, including MLP and multiple linear regression (MLR).

The sample data size of the chosen publications spans from days to years, from a few hundred to several thousand samples. For example, Hashemi et al. [31] produced a massive dataset comprising more than 3.6 million data points with a rooftop installation in Denver, Colorado. However, instead of the actual number, a time interval has been presented as the sample series of measures in the more significant part of the instances. This is due to the fact that the majority of a PV panel's performance relies on the solar irradiance, and solar irradiance itself is highly dependable during daytime. For instance, Onal [1] created a PV system of dimensions 2 × 2 by interconnecting two panels in a series configuration and two panels in a parallel configuration, and then proceeded to simulate its performance by collecting solar irradiation data for an entire day. Along with this short-term trial, some experiments extend from one year to more than thirty years [32,33]. In some situations, researchers have collected data at different time intervals to enhance their understanding. They have categorized data collection into long-term intervals, where observations span one day; medium-term intervals, ranging from six hours to one day; short-term intervals, spanning from thirty minutes to six hours; and ultra-short-term intervals, covering time periods from a few seconds to thirty minutes [34].

It is crucial to know the procedure of an experiment to evaluate its quality and other characteristics since it strongly influences the result. It is more prominent in solar experiments since researchers can postulate solar panels' performance and durability using computational tools in either simulation or practical experiments. During our screening, we found that some authors performed their studies practically in real time [35,36,58], and many authors adopted simulation tools with curated data sets to obtain their findings [5,67,4]. In another situation, simulation and practical approaches have been applied in the same experiments for a better comparative analysis [50,61,41]. They observe a clear difference between the simulation and the other two (hardware and both) operations. The simulation approach has been deployed in 66% of the instances, while in the case of the hardware method, it is 18%, and in the case of both, it is 16%. One of the main reasons for simulation being the highest-picked approach might be the immediacy of the outputs. If the datasets provided in the simulation program are solid and reliable enough, it produces accurate results with minimum errors. Simulation-based studies are the most common because of their cost-effectiveness and

simplicity of data manipulation. Also, several authors blended simulation and hardware methodologies to elevate the experiment's performance and dependability. This way, minimizing the experimental cost with a significant performance boost is possible.

Among different metrics, RMSE indicates how a specific range of data has deviated from their best fit line. It corresponds to the measure of dispersion or spread exhibited by the residuals within a dataset. It is beneficial when we need to know how the predicted values have deviated from the original or the expected values. When the RMSE value is higher, it signifies a substantial deviation between the predicted values and their corresponding original values. Conversely, a lower RMSE value indicates a close alignment between the predicted and original values. In the case of PV systems, features—such as temperatures, irradiance, and power efficiency—greatly vary due to external environmental factors. RMSE and similar error metrics provide a visual representation of the deviation between the predicted performance and the actual performance. That is why RMSE was the most used error metric utilized by the authors to evaluate their works. When presenting the RMSE, it is assumed that the errors are unbiased and conform to a normal distribution. The purpose of RMSE is to offer a comprehensive depiction of the error distribution [69]. Hence, RMSE is a suitable option for evaluating PV performance. MAE determines the precision of continuous variables and provides the mean difference between measured and actual values. Since it offers absolute values, it provides information on the size of the divergence but not its direction. A prediction's accuracy increases with decreasing MAE score. An MAE has a minimum value of zero, indicating no difference between the actual and anticipated data. Compared to the RMSE metric, this error index in the PV sector is more helpful in evaluating a number of parameters. It is because RMSE squares the error, occasionally distorting the intended meaning and introducing further problems. The coefficient of determination ($R2$) is a statistical approach for determining how well a model predicts. It is often referred to as the goodness of fit. It describes how a dependent variable responds to changes in an independent variable. $R2$ can have a value as low as zero and as high as one. Values near one indicate that a model is better at predicting, whereas near zero indicates a model cannot accurately predict. This statistical technique is helpful in the context of PV systems since PV parameters, such as power efficiency and DC to AC conversion, are influenced by several independent factors. $R2$ helps researchers relate those metrics to the PV systems. the complete comparison among different error metrics based on their usage. Just 11.6% behind the most popular error metric (RMSE), the second position was secured by MAE. The third and fourth most popular error metrics are $R2$ and MSE; they differ by a negligible percentage (17.5% and 15.9%, respectively). However, the fifth error metric, MAPE, differs significantly from its immediate most used error metric. The last three least used error metrics—absolute error (AE), relative absolute error (RAE), and RRSE—have the same use cases of only 0.529%. The AE is not commonly employed as an error metric due to its limited ability to offer meaningful insights. It merely represents the absolute magnitude of the difference between the exact and measured values, without providing further informative details. RAE is another metric used by researchers to justify solar performance [30]. Solar irradiance is a complex system that provides luminance differently depending on the daytime and weather conditions. For a simple error metric like RAE, it is almost impossible to predict such a complex system, which is why it is not so popular in the case of solar performance evaluation.

ML methods have performed better in different experiments depending on their experimental setup, criteria, location, weather, and, most importantly, the simulation data. For instance, Onal [1] conducted a study across 12 cities located in a representative climate zone in China. Their findings revealed that the cold, hot summer, and cold winter zones were the most favorable environments for their designed system. SVR was found to be the best performing ML method in their experiments producing MSE of 0.000011548 and RMSE of 0.0034. Likewise, solar PV performs very well in hotter regions, as shown in [38,31,49], which is not very surprising as the higher the temperature, the higher the voltage we get from our solar cells [70]. In hotter areas, the simulation data are more stable and do not vary much, which helps predict more accurate results.

Authors varied time duration to yield a better outcome. Theocharides et al. [32] conducted experiments using a one-year sample dataset and determined that ANN was the most suitable model for the given scenario. Duchaud et al. [55] ran the experiment during the peak hours of the day. They ran it in France in two locations (Ajaccio and Odeillo). They conducted the experiment in different intervals, namely, ultra-short-term, short-term, and long-term. First, they conducted it at varying times from 2 minutes to 1 hour (short-term), then varying times from 10 minutes to 6 hours (long-term) at maximum. The experiment was based on those two locations' global horizontal and tilted irradiance data. In their experiment, ARMA was the top performing ML algorithm among RF, MLP, SVM, and ARMA. The normalized RMSE (nRMSE) score was 0.8% in the short-term and 1.6% in the long-term scenario. Table 3 reveals that ANN wins in terms of best performing ML methods [38,32,50,54,58,61]. Jung et al. [54] achieved 97% $R2$ score in their testing dataset using ANN, whereas Kim and Kim [36] achieved 96.42% and Zhao et al. [50] achieved 82.90% $R2$ score. Along with ANN, other ML approaches, such as BPNN, ANFIS, RVFL, and RF, were extensively applied in the PV performance tests. Prediction of ANN and BPNN becomes very accurate with increasing data. Aljanad et al. [47], Wang et al. [5] have shown different performances using different optimizers even after using the same ML algorithms. Aljanad et al. [47] have developed a solid and persistent BPNN model to predict the global solar irradiance for tropical countries like Malaysia for an extremely short time interval, showing remarkable improvement in terms of system parameter estimation. In addition, Wang et al. [5] have proposed an enhanced equilibrium optimizer that is employed in conjunction

with BPNN. This optimizer demonstrates enhanced efficiency in optimization and produces fitness values that are more reasonable. They stated that their model has the capability to enhance both the precision and reliability of optimizer when it comes to estimating photovoltaic cell parameters. We have encountered several other cases where the same ML methods with different optimizers yielded different results [41,4]. Therefore, it is evident that the same ML methods can perform differently in different optimizers, and finding the right optimizer is an essential step in extracting the best performance. Along with the use of different optimizers, we have seen another interesting result: the order of the combinations of ML methods used in the experiments influences the results [43]. The naive forecast was the baseline in [43]'s experiment. To check the accuracy, several locations in Europe were considered. Also, a maximum of 24 hours of forecasts were recorded and then compared using different goodness of fit for a better comparison. They experimented four different models: only LSTM, only CNN, LSTM-CNN, and CNN-LSTM. Interestingly, the last two combinations showed different results. In Ulm, Germany, LSTM-CNN outperformed the other three combinations with an RMSE of 95.25, where its nearest competitor was CNN-LSTM with an RMSE of 97.35. In Almería, Spain—the sunniest location among other experimental locations—LSTM-CNN outperformed all three other ML methods in terms of MAE and root mean absolute error (RMAE) with values of 51.89 and 76.87%, respectively. Though it is evident that LSTM and CNN combinations performed around 2% better than individual models. Yet, according to the author, these minor differences matter because they enhance the reliability and robustness of the model. In the context of grid connection systems, Bukar et al. [71] proposed a novel approach for managing energy in microgrid systems. The author's approach utilizes a technique called the grasshopper optimization algorithm to optimize the energy management process. This optimization resulted in a reduction of fuel consumption by 92.4% and carbon dioxide emission by 92.3%. Therefore, it is evident that energy loss, power optimization, and advanced system parameters prediction is getting improved with the help of different ML algorithms.

6. Conclusion

In this review paper, we investigated the patterns in PV performance evolution over the last three years, how its performance varied over time, and how they were assessed with the help of different ML algorithms. According to our findings, the way PV performance is assessed has changed significantly. Sophisticated ML algorithms outperform traditional approaches. We classified and evaluated the articles based on five research questions. Ranking the ML methods is difficult due to the differences in locations, materials, and environmental factors, such as irradiance and humidity. Overall, it was evident after the evaluation that NN was the most popular ML parent category, followed by RVFL and SVM. Within the category of individual machine learning methods, ANN emerged as the most widely used approach, followed by BPNN, GNN, and LSTM, respectively. Our work also shows that NN is meaningful for its capacity to anticipate PV performance accurately. In terms of error quantification, a large number of researchers deployed RMSE in conjunction with additional metrics, such as MAE, *R*2, MSE, and MAPE. Limitations of this study include the number of research papers and the variety of source databases. Due to a lack of available ML-based PV parameters prediction works, we had to stick with a limited number of articles. Scopus, well-known for having top-notch papers, is the only database from which we sourced the articles; however, other databases might help get new insights. Despite limitations, this paper reveals the trend and open issues in PV parameter estimations, which would help future researchers further improve the parameters estimation performance.

References

[1] Y. Onal, Gaussian kernel based SVR model for short-term photovoltaic MPP power prediction, Comput. Syst. Sci. Eng. 41 (2022) 141–156, https://doi.org/10. 32604/csse.2022.020367.
[2] Korea Energy Agency, Renewable Energy Supply Status, e-Country Indicators Index Inquiry Details, https://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1171, 2022. (Accessed 17 January 2022).
[3] T.N. Woods, G.J. Rottman, Solar Ultraviolet Variability over Time Periods of Aeronomic Interest, Geophysical Monograph-American Geophysical Union, vol. 130, 2002, pp. 221–234.
[4] L. Wang, Z. Chen, Y. Guo, W. Hu, X. Chang, P. Wu, C. Han, J. Li, Accurate solar cell modeling via genetic neural network-based meta-heuristic algorithms, Front. Energy Res. 9 (2021), https://doi.org/10.3389/fenrg.2021.696204.
[5] J. Wang, B. Yang, D. Li, C. Zeng, Y. Chen, Z. Guo, X. Zhang, T. Tan, H. Shu, T. Yu, Photovoltaic cell parameter estimation based on improved equilibrium optimizer algorithm, Energy Convers. Manag. 236 (2021) 114051, https://doi.org/10.1016/j.enconman.2021.114051.
[6] A. Allouhi, R. Saadani, T. Kousksou, R. Saidur, A. Jamil, M. Rahmoune, Grid-connected PV systems installed on institutional buildings: technology comparison, energy analysis and economic performance, Energy Build. 130 (2016) 188–201.
[7] M.A. Eltawil, Z. Zhao, Grid-connected photovoltaic power systems: technical and potential problems—a review, Renew. Sustain. Energy Rev. 14 (2010) 112–129.
[8] T. Ma, H. Yang, L. Lu, Solar photovoltaic system modeling and performance prediction, Renew. Sustain. Energy Rev. 36 (2014) 304–315.
[9] K. Ba¸saran, F. Bozyigit, ˘ P. Siano, P. Yıldırım Ta¸ser, D. Kılınç, Systematic literature review of photovoltaic output power forecasting, IET Renew. Power Gener. 14 (2020) 3961–3973, https://doi.org/10.1049/iet-rpg.2020.0351.
[10] S. Sobri, S. Koohi-Kamali, N.A. Rahim, Solar photovoltaic generation forecasting methods: a review, Energy Convers. Manag. 156 (2018) 459–497, https://doi.org/10.1016/j.enconman.2017.11.019.
[11] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.M. de Pison, F. Antonanzas-Torres, Review of photovoltaic power

forecasting, Sol. Energy 136 (2016) 78–111, https://doi.org/10.1016/j.solener.2016.06.069.
[12] R. Ahmed, V. Sreeram, Y. Mishra, M. Arif, A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization, Renew. Sustain. Energy Rev. 124 (2020) 109792, https://doi.org/10.1016/j.rser.2020.109792.
[13] Y.K. Semero, D. Zheng, J. Zhang, A PSO-ANFIS based hybrid approach for short term PV power prediction in microgrids, Electr. Power Compon. Syst. 46 (2018) 95–103, https://doi.org/10.1080/15325008.2018.1433733.
[14] S. Rajamand, M. Shafie-khah, J.P. Catalão, Energy storage systems implementation and photovoltaic output prediction for cost minimization of a microgrid, Electr. Power Syst. Res. 202 (2022) 107596, https://doi.org/10.1016/j.epsr.2021.107596.
[15] M. Mittal, B. Bora, S. Saxena, A.M. Gaur, Performance prediction of PV module using electrical equivalent model and artificial neural network, Sol. Energy 176 (2018) 104–117, https://doi.org/10.1016/j.solener.2018.10.018.
[16] W. Zhou, H. Yang, Z. Fang, A novel model for photovoltaic array performance prediction, Appl. Energy 84 (2007) 1187–1198, https://doi.org/10.1016/j. apenergy.2007.04.006.
[17] G.G. Kim, J.H. Choi, S.Y. Park, B.G. Bhang, W.J. Nam, H.L. Cha, N. Park, H.-K. Ahn, Prediction model for PV performance with correlation analysis of environmental variables, IEEE J. Photovolt. 9 (2019) 832–841, https://doi.org/10.1109/JPHOTOV.2019.2898521.
[18] R. Dows, E. Gough, PVUSA Model Technical Specification for a Turnkey Photovoltaic Power System, Technical Report, Pacific Gas and Electric Co., San Ramon, CA (United States) 1995.
[19] O. Ulleberg, Stand-alone power systems for the future: Optimal design, operation and control of solar-hydrogen energy systems, Ph.D. dissertation, Norwegian University of Science and Technology, 1998, https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/DE99751232.xhtml. (Accessed 3 June 2023).
[20] J.A. Duffie, W.A. Beckman, W.M. Worek, Solar engineering of thermal processes, 2nd ed., J. Sol. Energy Eng. 116 (1994) 67–68, https://doi.org/10.1115/1. 2930068.
[21] M. Haouari-Merbah, M. Belhamel, I. Tobías, J. Ruiz, Extraction and analysis of solar cell parameters from the illuminated current–voltage curve, Sol. Energy Mater. Sol. Cells 87 (2005) 225–233, https://doi.org/10.1016/j.solmat.2004.07.019.
[22] M.R. Hasan, M.M. Hasan, M.Z. Hossain, Effect of vocal tract dynamics on neural network-based speech recognition: a Bengali language-based study, Expert Syst. 39 (2022) e13045, https://doi.org/10.1111/exsy.13045.
[23] Y. Wang, A New Concept Using Lstm Neural Networks for Dynamic System Identification, in: 2017 American control conference (ACC), IEEE, 2017, pp. 5324–5329.
[24] A. Shewalkar, D. Nyavanandi, S.A. Ludwig, Performance evaluation of deep neural networks applied to speech recognition: Rnn, LSTM and GRU, J. Artif. Intell. Soft Comput. Res. 9 (2019) 235–245.
[25] E.T. Rother, Systematic literature review X narrative review, Acta Paul. Enferm. 20 (2007) vii–viii, https://doi.org/10.1590/S0103-21002007000200001.
[26] R.W. Wright, R.A. Brand, W. Dunn, K.P. Spindler, How to write a systematic review, Clin. Orthop. Relat. Res. 455 (2007) 23–29.
[27] D.J. Cook, D.L. Sackett, W.O. Spitzer, Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation on meta-analysis, J. Clin. Epidemiol. 48 (1995) 167–171.
[28] B. Gillen, E. Snowberg, L. Yariv, Experimenting with measurement error: techniques with applications to the caltech cohort study, J. Polit. Econ. 127 (2019) 1826–1863.
[29] Y.-H. Pao, G.-H. Park, D.J. Sobajic, Learning and generalization characteristics of the random vector functional-link net, Neurocomputing 6 (1994) 163–180.
[30] A. Shahsavar, H. Moayedi, A.H.A. Al-Waeli, K. Sopian, P. Chelvanathan, Machine learning predictive models for optimal design of building-integrated photovoltaic-thermal collectors, Int. J. Energy Res. 44 (2020) 5675–5695, https://doi.org/10.1002/er.5323.
[31] B. Hashemi, S. Taheri, A.-M. Cretu, E. Pouresmaeil, Systematic photovoltaic system power losses calculation and modeling using computational intelligence techniques, Appl. Energy 284 (2021) 116396, https://doi.org/10.1016/j.apenergy.2020.116396.
[32] S. Theocharides, G. Tziolis, J. Lopez-Lorente, G. Makrides, G.E. Georghiou, Impact of data quality on day-ahead photovoltaic power production forecasting, in: 2021 IEEE 48th Photovoltaic Specialists Conference (PVSC), 2021, pp. 0918–0922.
[33] B. Brahma, R. Wadhvani, Solar irradiance forecasting based on deep learning methodologies and multi-site data, Symmetry 12 (2020), https://doi.org/10.3390/ sym12111830.
[34] I.A. Ibrahim, M.J. Hossain, B.C. Duck, An optimized offline random forests-based model for ultra-short-term prediction of PV characteristics, IEEE Trans. Ind. Inform. 16 (2020) 202–214, https://doi.org/10.1109/TII.2019.2916566.
[35] M.E. Zayed, J. Zhao, W. Li, A.H. Elsheikh, M.A. Elaziz, A hybrid adaptive neuro-fuzzy inference system integrated with equilibrium optimizer algorithm for predicting the energetic performance of solar dish collector, Energy 235 (2021) 121289, https://doi.org/10.1016/j.energy.2021.121289.
[36] S. Kim, S. Kim, Performance estimation modeling via machine learning of an agrophotovoltaic system in South Korea, Energies 14 (2021) 6724, https://doi.org/10.3390/en14206724.
[37] S. Yao, Q. Kang, M. Zhou, A. Abusorrah, Y. Al-Turki, Intelligent and data-driven fault detection of photovoltaic plants, Processes 9 (2021), https://doi.org/10. 3390/pr9101711.
[38] Y. Chaibi, M. Malvoni, T. El Rhafiki, T. Kousksou, Y. Zeraouli, Artificial neural-network based model to forecast the electrical and thermal efficiencies of PVT air collector systems, Clean. Eng. Technol. 4 (2021) 100132, https://doi.org/10.1016/j.clet.2021.100132.

[39] D. Mazzeo, M.S. Herdem, N. Matera, M. Bonini, J.Z. Wen, J. Nathwani, G. Oliveti, Artificial intelligence application for the performance prediction of a clean energy community, Energy 232 (2021) 120999, https://doi.org/10.1016/j.energy.2021.120999.
[40] M. Abd Elaziz, S. Senthilraja, M.E. Zayed, A.H. Elsheikh, R.R. Mostafa, S. Lu, A new random vector functional link integrated with mayfly optimization algorithm for performance prediction of solar photovoltaic thermal collector combined with electrolytic hydrogen production system, Appl. Therm. Eng. 193 (2021) 117055, https://doi.org/10.1016/j.applthermaleng.2021.117055.
[41] M.E. Zayed, J. Zhao, W. Li, A.H. Elsheikh, M.A. Elaziz, D. Yousri, S. Zhong, Z. Mingxi, Predicting the performance of solar dish Stirling power plant using a hybrid random vector functional link/chimp optimization model, Sol. Energy 222 (2021) 1–17, https://doi.org/10.1016/j.solener.2021.03.087.
[42] S. Gbémou, J. Eynard, S. Thil, E. Guillot, S. Grieu, A comparative study of machine learning-based methods for global horizontal irradiance forecasting, Energies 14 (2021), https://doi.org/10.3390/en14113192.
[43] S. Liebermann, J.-S. Um, Y. Hwang, S. Schlüter, Performance evaluation of neural network-based short-term solar irradiation forecasts, Energies 14 (2021), https://doi.org/10.3390/en14113030.
[44] S. Jung, W.J. Yun, M. Shin, J. Kim, J.-H. Kim, Orchestrated scheduling and multi-agent deep reinforcement learning for cloud-assisted multi-UAV charging systems, IEEE Trans. Veh. Technol. 70 (2021) 5362–5377, https://doi.org/10.1109/TVT.2021.3062418.
[45] O. Ulucak, E. Kocak, O. Bayer, U. Beldek, E.Ö. Yapıcı, E. Aylı, Developing and implementation of an optimization technique for solar chimney power plant with machine learning, J. Energy Resour. Technol. 143 (2021) 052109, https://doi.org/10.1115/1.4050049.
[46] B. Du, P.D. Lund, J. Wang, M. Kolhe, E. Hu, Comparative study of modelling the thermal efficiency of a novel straight through evacuated tube collector with MLR, SVR, BP and RBF methods, Sustain. Energy Technol. Assess. 44 (2021) 101029, https://doi.org/10.1016/j.seta.2021.101029.
[47] A. Aljanad, N.M.L. Tan, V.G. Agelidis, H. Shareef, Neural network approach for global solar irradiance prediction at extremely short-time-intervals using particle swarm optimization algorithm, Energies 14 (2021), https://doi.org/10.3390/en14041213.
[48] E.E. Looney, Z. Liu, A. Classen, H. Liu, N. Riedel, M. Braga, P. Balaji, A. Augusto, T. Buonassisi, I. Marius Peters, Representative identification of spectra and environments (RISE) using k-means, Prog. Photovolt., Res. Appl. 29 (2021) 200–211, https://doi.org/10.1002/pip.3358.
[49] I. Arora, J. Gambhir, T. Kaur, Data normalisation-based solar irradiance forecasting using artificial neural networks, Arab. J. Sci. Eng. 46 (2021) 1333–1343, https://doi.org/10.1007/s13369-020-05140-y.
[50] X. Zhao, Y. Han, L. Shen, Multi-objective optimisation of a free-form building shape to improve the solar energy utilisation potential using artificial neural networks, in: PROJECTIONS, vol. 1, Chinese University of Hong Kong, 2021, pp. 221–230 and Online.
[51] S. Prajapati, E. Fernandez, Performance evaluation of membership function on fuzzy logic model for solar PV array, in: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), 2020, pp. 609–613.
[52] T.-P. Chu, J.-H. Jhou, Y.-G. Leu, Image-based solar irradiance forecasting using recurrent neural networks, in: 2020 International Conference on System Science and Engineering (ICSSE), 2020, pp. 1–4.
[53] G. Guariso, G. Nunnari, M. Sangiorgio, Multi-step solar irradiance forecasting and domain adaptation of deep neural networks, Energies 13 (2020), https://doi.org/10.3390/en13153987.
[54] D.E. Jung, C. Lee, K.H. Kim, S.L. Do, Development of a predictive model for a photovoltaic module's surface temperature, Energies 13 (2020), https://doi.org/10.3390/en13154005.
[55] J.-L. Duchaud, C. Voyant, A. Fouilloy, G. Notton, M.-L. Nivet, Trade-off between precision and resolution of a solar power forecasting algorithm for micro-grid optimal control, Energies 13 (2020), https://doi.org/10.3390/en13143565.
[56] T. David, G. Amorim, D. Bagnis, N. Bristow, S. Selbach, J. Kettle, Forecasting OPV outdoor performance, degradation rates and diurnal performances via machine learning, in: 2020 47th IEEE Photovoltaic Specialists Conference (PVSC), 2020, pp. 0412–0418.
[57] R. Shah, Solar cell parameters extraction using multi-target regression methods, in: 2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I CPS Europe), 2020, pp. 1–6.
[58] S.K. Jung, Y. Kim, J.W. Moon, Performance evaluation of control methods for PV-integrated shading devices, Energies 13 (2020), https://doi.org/10.3390/ en13123171.
[59] S. Wang, Y. Wang, Y. Cheng, S. Sun, N. Wang, P. Yu, S. Wang, An improved model for power prediction of PV system based on Elman neural networks, in: 2020 Asia Energy and Electrical Engineering Symposium (AEEES), 2020, pp. 902–907.
[60] A.M. Karimi, J.S. Fada, N.A. Parrilla, B.G. Pierce, M. Koyutürk, R.H. French, J.L. Braid, Generalized and mechanistic PV module performance prediction from computer vision and machine learning on electroluminescence images, IEEE J. Photovolt. 10 (2020) 878–887, https://doi.org/10.1109/JPHOTOV.2020. 2973448.
[61] C. Qiu, Y.K. Yi, M. Wang, H. Yang, Coupling an artificial neuron network daylighting model and building energy simulation for vacuum photovoltaic glazing, Appl. Energy 263 (2020) 114624, https://doi.org/10.1016/j.apenergy.2020.114624.
[62] J. Alsarraf, H. Moayedi, A.S.A. Rashid, M.A. Muazu, A. Shahsavar, Application of PSO–ANN modelling for predicting the exergetic performance of a building integrated

photovoltaic/thermal system, Eng. Comput. 36 (2020) 633–646, https://doi.org/10.1007/s00366-019-00721-4.
[63] K. Pahwa, M. Sharma, M.S. Saggu, A. Kumar Mandpura, Performance evaluation of machine learning techniques for fault detection and classification in PV array systems, in: 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), 2020, pp. 791–796.
[64] P. Li, K. Zhou, X. Lu, S. Yang, A hybrid deep learning model for short-term PV power forecasting, Appl. Energy 259 (2020) 114216, https://doi.org/10.1016/j.apenergy.2019.114216.
[65] M. Sridharan, Application of generalized regression neural network in predicting the performance of solar photovoltaic thermal water collector, Ann. Data Sci. (2020), https://doi.org/10.1007/s40745-020-00273-1.
[66] Y. Uzun, M. Özcan, Rule extraction and performance estimation by using variable neighborhood search for solar power plant in Konya, Turk. J. Electr. Eng. Comput. Sci. 28 (2020) 635–645.
[67] C. Zhang, Y. Zhang, J. Su, T. Gu, M. Yang, Performance prediction of PV modules based on artificial neural network and explicit analytical model, J. Renew. Sustain. Energy 12 (2020) 013501, https://doi.org/10.1063/1.5131432.
[68] C. Correa-Jullian, J.M. Cardemil, E. López Droguett, M. Behzad, Assessment of deep learning techniques for prognosis of solar thermal systems, Renew. Energy 145 (2020) 2178–2191, https://doi.org/10.1016/j.renene.2019.07.100.
[69] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature, Geosci. Model Dev. 7 (2014) 1247–1250, https://doi.org/10.5194/gmdd-7-1525-2014.
[70] P. Singh, S. Singh, M. Lal, M. Husain, Temperature dependence of I–V characteristics and performance parameters of silicon solar cell, Sol. Energy Mater. Sol. Cells 92 (2008) 1611–1616, https://doi.org/10.1016/j.solmat.2008.07.010.
[71] A.L. Bukar, C.W. Tan, L.K. Yiew, R. Ayop, W.-S. Tan, A rule-based energy management scheme for long-term optimal capacity planning of grid-independent microgrid optimized by multi-objective grasshopper optimization algorithm, Energy Convers. Manag. 221 (2020) 113161.
[72] M. Palanivel, U. Kaithamalai, P. Parthsarathi, Performance assessment of IC and ANFIS based MPPT for PV system using Super Lift Boost Converter, in: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 6–11.