

Article

A Comparison of the Performance of Supervised Learning Algorithms for Solar Power Prediction

Leidy Gutiérrez^{1,*}, Julian Patiño^{1,*}  and Eduardo Duque-Grisales^{1,2} 

¹ Facultad de Ingeniería, Institución Universitaria Pascual Bravo, 050034 Medellín, Colombia; leidy.gutierrez@pascualbravo.edu.co (L.G.); eduardo.duque@esumer.edu.co (E.D.-G.)

² Facultad de Estudios Empresariales y de Mercadeo, Institución Universitaria Esumer, 050035 Medellín, Colombia

* Correspondence: julian.patino@pascualbravo.edu.co; Tel.: +57-4-448-0520

Abstract: Science seeks strategies to mitigate global warming and reduce the negative impacts of the long-term use of fossil fuels for power generation. In this sense, implementing and promoting renewable energy in different ways becomes one of the most effective solutions. The inaccuracy in the prediction of power generation from photovoltaic (PV) systems is a significant concern for the planning and operational stages of interconnected electric networks and the promotion of large-scale PV installations. This study proposes the use of Machine Learning techniques to model the photovoltaic power production for a system in Medellín, Colombia. Four forecasting models were generated from techniques compatible with Machine Learning and Artificial Intelligence methods: K-Nearest Neighbors (KNN), Linear Regression (LR), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The results obtained indicate that the four methods produced adequate estimations of photovoltaic energy generation. However, the best estimate according to RMSE and MAE is the ANN forecasting model. The proposed Machine Learning-based models were demonstrated to be practical and effective solutions to forecast PV power generation in Medellín.

Keywords: photovoltaic systems; machine learning; supervised learning; prediction; artificial neural networks; k-nearest neighbors; linear regression; support vector machine



Citation: Gutierrez, L.; Patiño, J.; Duque-Grisales, E. A Comparison of the Performance of Supervised Learning Algorithms for Solar Power Prediction. *Energies* **2021**, *14*, 4424. <https://doi.org/10.3390/en14154424>

Received: 19 May 2021

Accepted: 30 June 2021

Published: 22 July 2021

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



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The increase in the world's energy demand is evident, creating a threat of a global energy crisis, which causes adverse environmental effects on our habitat [1]. The efficient use of energy is a problem that has caused great interest in the world because the raw material (fossil fuels) has a significant drop in international reserves, causing severe economic, political and social problems [2].

The constant technological and social development of humanity implies a progressive demand for electrical energy. However, the primary energy generation methods used in the world come from fossil fuels, reaching an annual rate of consumption in oil, gas and carbon of 3.1 million tons (Mt) [3], representing more than 80% of world consumption [4]. These sources represent a higher demand due to their low cost, but they negatively affect the environment, considering that they increase carbon dioxide (CO₂) and greenhouse gas emissions [5], contributing to global warming. Figure 1 shows a record and prediction of the global demand for different types of energy, where the use of fossil fuels undoubtedly continues to lead.

Non-conventional renewable energies emerged at the end of the 1990s as an alternative to mitigate the impacts of greenhouse gases. These energy systems have sources that reside in natural phenomena: processes or materials that can be transformed into energy and regenerate naturally, so they are available continuously or periodically [6].

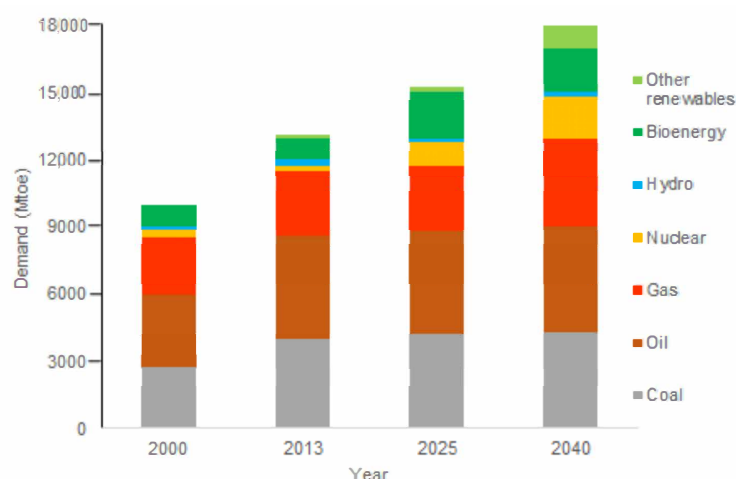


Figure 1. Projected world energy demand.

The importance of renewable energies has been increasing; this can be seen from their integration into public energy networks [7], to the application of technologies such as neural networks in the prediction of energy production, as well the application of different techniques to obtain an optimization in the prediction process [8,9]. The development of energy prediction models is an important task, allowing optimization and extracting the most energy production of the system.

Photovoltaic (PV) solar energy systems constitute one of the primary sources of renewable energy generation. The PV effect generates electricity from the energy transported by photons of light when they affect semiconductor materials [10]. The construction of solar power generation systems depends on the incident radiation (solar irradiance) and climatic variables in the selected location, such as temperature, relative humidity and wind speed [11]. The inherent variability of these environmental factors makes the power generated from a PV system a dynamical variable changing with time. The difficulties in predicting PV power production cause adverse effects in electric grid aspects such as reliability, stability, planning and scheduling tasks and market operations [12,13]. Consequently, one of the main topics of interest for research in PV systems lies in forecasting power generation.

PV power predictions are mainly based on reviewing statistical data over time and long-term meteorological data [11], providing essential information to determine the expected behavior in generation systems by different methods. Many reported studies focus on forecasting solar irradiance by image-based approaches, statistical properties and numerical weather simulations [8,14–16]. The predicted solar radiation and other information are employed as input data for PV commercial simulation software tools [17] to calculate PV power output. In addition, various forecasting models are obtained from historical data, including techniques supported by Machine Learning and Artificial Intelligence methods [13,15,18–20].

The availability of computational models for practical and effective PV power forecasting can decrease the effects of PV uncertainty in the power grid and increase the deployment of PV systems [19]. Despite the abundant literature on the matter, few studies report the application of Machine Learning techniques to predict PV power generation from systems located in the city of Medellín, Colombia. In [16], the authors described a Markov chain approach for the day-ahead forecasting of the hourly solar irradiance in Medellín. An Artificial Neural Network was used to evaluate the electrical performance of two different photovoltaic technologies in Medellín [21]. In this sense, there is a need to obtain PV power production models to explore the potential of penetration of PV systems in a city such as Medellín, Colombia.

This work presents a performance comparison of supervised learning algorithms for PV solar power prediction in Medellín, Colombia. Four different forecasting models are

generated using the techniques of K-Nearest Neighbors (KNN), Linear Regression (LR), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The performance of these methods is evaluated using data from a PV energy system located in the university campus of the I. U. Pascual Bravo in Medellin, Colombia. The work is organized as follows. First, there is a summary of PV solar forecasting methods found in the literature. The following section describes the PV system and the data collection process. Then, the selected methods for PV power prediction are presented, with a description of the chosen performance metrics. Following the methodology description, the proposed algorithms are evaluated using the collected data. Finally, after the corresponding discussion and analysis of the results, some conclusions are presented.

2. A Literature Review on PV Power Estimation

Until 2010, research and development of photovoltaic generation prediction models was minimal. Most of the models were based on predicting the radiation incident on the photovoltaic solar park, and the electrical power produced was calculated from these values. The data sources were the curves provided by the PV solar panel manufacturers or a series of equations or known empirical relationships [22]. However, in the last ten years, the publication of new prediction models has grown considerably due to the exponential increase of PV systems worldwide and the studies on the characteristics of this energy source.

Prediction models usually rely on reviewing statistical data of production over time and long-term meteorological data [11], providing essential information to determine the expected behavior in generation systems by a wide variety of methods. There is an active interest in forecasting energy production in systems with multiple sources assessing the available power output of each component [23,24]. These predictions make it possible to identify the amount of energy generated, according to the climatic and operating conditions of the system, with adequate modeling and analytical treatment [25].

Different methodologies for prediction in photovoltaic energy systems are identified in the literature. In some studies [15,26,27], the energy generated by photovoltaic systems is predicted using neural network methods. This type of analysis has also been applied to predicting the temperature of photovoltaic modules [28]. Estimates of solar radiation have been determined with statistical tests for percentage errors, mean absolute bias and squared error. In the works published by Halabi et al. [29] and Yaniktepe and Genc [30], several methods are proposed to know the global solar radiation by months, based on historical data from meteorological services [12].

Today, thanks to the development achieved by the different models used for prediction, several classifications can be made depending on the criteria taken into account [13,18]. Some criteria consider the linearity of the model and classify them as linear and nonlinear. Others consider the method used for the mathematical development of the model and classify them into models based on Artificial Intelligence techniques or regressive models [12]. Figure 2 presents a classification of PV Prediction models (adapted from [12,13,18]), with two main approaches: models based on past values and atmospheric models.

2.1. Models Based on Past Values

These models only use past values as input, which can only be the variable to be predicted or the variable to be predicted complemented with other variables that may influence it. These variables can include not only those corresponding to the instant of time in which they occurred, but they can also be meteorological variables measured locally in those past instants. As shown in Figure 2, these models can be broadly classified as described in the next subsections.

2.1.1. Persistence Models

Based solely on historical records, the estimation of the energy production of the PV system equates to the registered power production around the same time in a previously

measured day of operation. This prediction technique is used mainly for comparison or performance benchmarking of other modeling approaches [12].

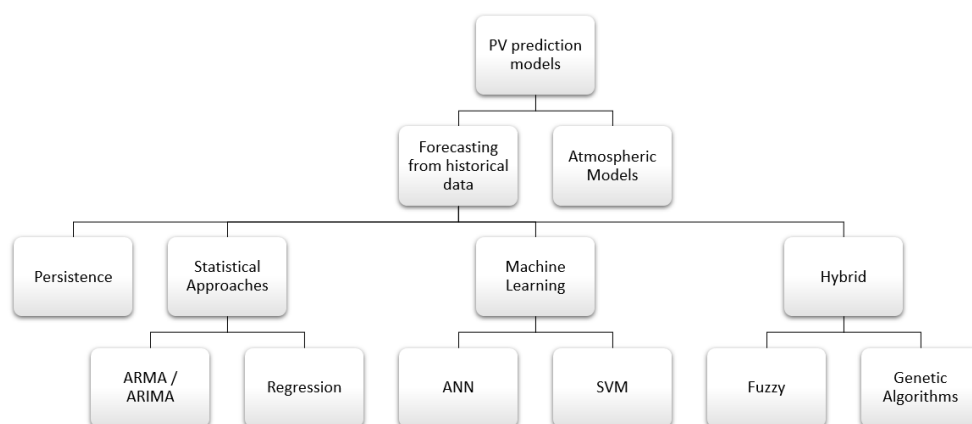


Figure 2. Classification of PV Prediction models and techniques (adapted from [12,13,18]).

2.1.2. Statistical Approaches

In these methods of PV prediction, time series analysis can be used to understand the behavior of an observed data series or to predict future values of these series. These methods are very useful for short-term estimation of PV power production. The following are some of the techniques employed for the statistical approaches:

- Regression models: PV power output is considered a dependent variable explained by the meteorological variables [31]. They usually require mathematical models and the consideration of explanatory variables.
- Auto-regressive models: ARMA (Auto Regressive Moving Average) and ARIMA (Auto-Regressive Integrated Moving Average) are commonly employed techniques for PV prediction using time series. These techniques assume that the past values of the series, called the history of the series, influence the future of the series through a combination of Auto-Regressive (AR) and Moving Average (MA) elements. In a pure auto-regressive process, the future values of the series only depend on past values. In the process of moving averages, the future values of the series depend on random variables, independent of each other and which are modeled as white noise [32].

2.1.3. Machine Learning Techniques

These models are based on Artificial Intelligence approaches. Often, these methods require a large volume of data to offer an accurate estimation of PV energy production. The following are some of the techniques employed for the Machine Learning approaches:

- Artificial Neural Networks (ANN): Artificial neural networks consist of a mathematical model based on the biological nervous system. The vast majority of the studies are carried out with networks of the Multilayer Perceptron type (MLP). An MLP can approximate nonlinear relationships between input and output data. There is considerable interest in ANN-based approaches for solar power prediction [15].
- Support Vector Machines (SVM): SVM consists of supervised learning algorithms related to classification and regression problems. They are employed for PV power estimation using a time series analysis approach, and the interest in these methods is growing [13].

2.1.4. Hybrid Models

These models combine physical and statistical models, looking to enhance the advantages of both approaches to raise accuracy in PV power estimation. For example, neuro-fuzzy systems combine the supervised learning capacity of a neural network with the knowledge representation of a fuzzy inference system. A prevalent name for this type

of system is Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and it has been applied to PV power estimation [10].

Other cases of hybrid models are the use of neural networks optimized utilizing genetic algorithms, the use of ARMA models with neural networks, the union of several types of neural networks and the combination of atmospheric models such as MM5 for the prediction of radiation with fuzzy logic or neural networks for power prediction [8].

2.2. Atmospheric Models

These models incorporate the prediction values of meteorological variables obtained by the numerical prediction programs existing in different meteorological institutes. In addition, these inputs may be complemented by those indicated in the previous group. In this category, the most widely used models are MM5 (from Pennsylvania University and National Center for Atmospheric Research) and WRF-NMM (from National Oceanic and Atmospheric/National Centers for Environmental Prediction) [33].

3. Methodology

As shown in Section 2, there are many strategies based on historical data for PV power prediction. This paper proposes the modeling of PV power production by computational methods based on historical data from a generation system located in Medellín, Colombia. Machine Learning is a wide field of computer science that provides suitable techniques for making predictions. This work aims to study different Machine Learning techniques and supervised learning models to identify which one provides the best estimation of power produced by photovoltaic plants. The performance of the proposed methods was evaluated from experimental data. The proposed models could be of interest for the simulation and future implementation of similar PV systems in the region, contributing to the satisfaction of energy demand.

Usually, Machine Learning algorithms are divided into two main techniques:

- Unsupervised Learning: These models group and interpret data based only on input data. Clustering techniques are applied to find “natural” groups or patterns in data.
- Supervised Learning: These models develop a predictive model based on both input and output data, using techniques such as classification and regression.

For this work, four algorithms of supervised learning are employed to estimate PV power output: K-Nearest Neighbors (KNN), Linear Regression (LR), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Figure 3 describes the framework of the employed methodology. At first, the data collection and preprocessing stage involves exploring, correcting and normalizing the database and the division into training and validation sets. The modeling stage intends to train the selected algorithms with the training data until a suitable model is obtained. The last stage involves evaluating the models with the testing data, calculating the estimation error and analyzing the results.

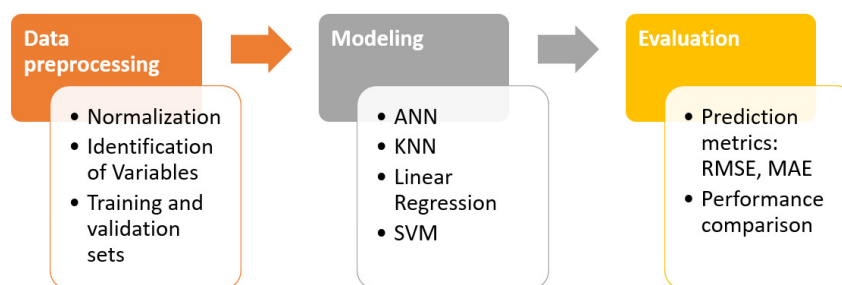


Figure 3. Framework of the proposed methodology.

The algorithms were developed using Matlab R2021a, a robust technical calculation and programming language developed by The MathWorks Inc. (Natick, MA, USA) for

algorithm development, data analysis, visualization, and numerical computation. In addition, Matlab counts with special customized packages or toolboxes for specialized topics. The Neural Network Toolbox supported the design, training and validation of the ANN. The remaining methods were implemented with the Statistics and Machine Learning Toolbox. The following subsections describe the four methods and the error metrics.

3.1. Artificial Neural Networks (ANN)

Artificial Neural Networks consist of a mathematical model based on the biological nervous system, made up of many simple elements that process information through their dynamic state in response to external inputs [34]. The basic units of the model are neurons. Each of these neurons interconnects with the inputs and the different elements of the model with an associated weight. The main stages of neural network-based modeling are: choice of input variables, network type and the number of layers, dataset preparation, neural network creation, neural network training and validation. The time series of electric power are usually nonlinear functions of external variables. Therefore, due to this non-linearity, Artificial Neural Networks receive significant attention in solving problems of this type [15].

Figure 4 shows a standard representation of an ANN. The mathematical expressions for the neural network are shown in Equations (1) and (2), with x_j being the input variables to the neuron k weighted by the synaptic coefficients w_{kj} . The weighted sum of inputs is added to b_k , a bias factor (negative or positive). This total sum u_k is applied to the activation function φ to produce an output y_k . This output can be an input for another neuron or the output of the full neural network.

$$u_k = \sum_{j=1}^m w_{kj} x_j. \quad (1)$$

$$y_k = \varphi(u_k + b_k). \quad (2)$$

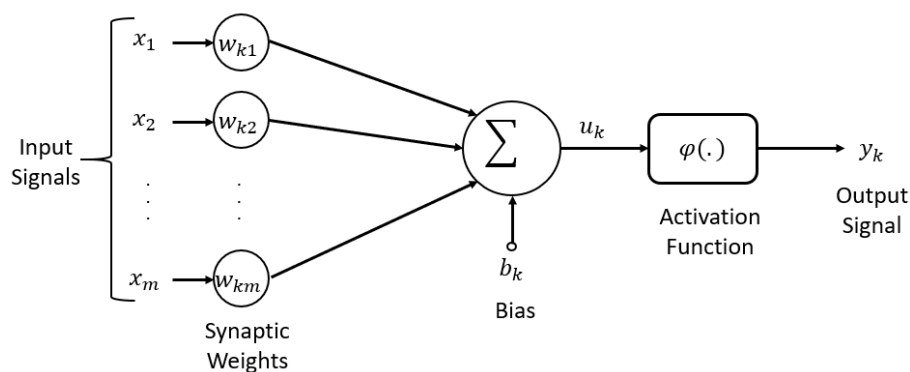


Figure 4. A standard schematic representation of an Artificial Neural Network.

Supervised ANN learning requires a training dataset with both vectors of inputs and the corresponding outputs. In the training stage, synaptic weights of the neurons are adjusted to minimize the error signal: the difference between model predicted values and observed data. Usually, error derivatives back-propagate to each neuron layer to change the parameters. A widely employed method for ANN training is the Levenberg–Marquardt algorithm, which calculates weights according to the following rule:

$$w_{k+1} = w_k - (J_k^T J_k + \mu I)^{-1} J_k e_k. \quad (3)$$

In Equation (3), e_k is the error, J denotes the Jacobian matrix of e_k and μ is a parameter increasing or decreasing with each step. The iteration process finishes when a stop condition is reached, such as a given number of cycles (epochs) or predefined error value. When

designing an ANN, both the number of hidden layers and neurons in each layer must be selected. As the number of layers and neurons grows, the ability of the ANN to adjust any function also grows. However, the training time increases, and there is a greater risk of overtraining the network. Often, these parameters are defined by a heuristic process of trial and error.

3.2. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a non-parametric approximation method, which allows solving classification and regression problems. KNN is based on the assumption that an object corresponds to the same class as its closest neighbors. At first, the method requires the specification of a positive integer k . The algorithm identifies the k points on the dataset with a similar pattern to the sample (the so-called K-Nearest Neighbors) for any sample [20]. This selection process requires knowing the distance between all the samples in the database and the newly analyzed sample.

When used in forecasting applications, the KNN technique finds the neighbors: the elements from the training set matching the reference conditions according to some pre-determined features [35]. In this work, the features are the historical data of input variables and PV power. These data are assembled in a matrix X_{ij} , where each row denotes a feature vector for a particular time in the estimation. The nearest neighbor for a new data point at time t , characterized by the feature vector y_j , is compared with all the rows in X_{ij} and the value is stored in the vector of Euclidean distances d_i :

$$d_i = \sqrt{\sum_i (X_{ij} - y_j)^2}. \quad (4)$$

The distance values are sorted in ascending order, and the first k matches are identified. The numerical value for y_j is the average of all the variable numerical values of the K-Nearest Neighbors.

3.3. Linear Regression (LR)

Multiple LR constitutes one of the most widely employed algorithms in supervised learning. This statistical technique looks for the weighted linear combination of input variables that better fits an output variable [31]. The formula for the calculation of multiple LR is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon \quad (5)$$

In Equation (5), y_i is the estimated output, x_i are the input variables, β_0 is a constant term, β_i are the slope coefficients for each input variable and ϵ denotes the model's error term (residuals).

3.4. Support Vector Machines (SVM)

They are a set of supervised learning algorithms developed by Vladimir Vapnik and his team at the AT&T Labs. These methods are appropriately related to classification and regression problems. Given a set of training patterns, a set of classes can be labeled, and an SVM can be trained to build a model that predicts the class from a new sample [31]. Intuitively, an SVM is a model that represents the sample points in space and separates the classes by as wide a space as possible. When the new samples are in correspondence with the model, they can be classified depending on their class proximity.

SVM can also be used as a regression method without changing the main characteristics of the algorithm, performing a regression from the classifier [19]. Consider a set of data $\{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$, where $x_i \in R^n$ represent the input variables and $y_i \in R^1$ is the PV power output. The prediction function $g(x)$ is presented in Equation (6):

$$y_i = g(x) = w \times \psi(x) + b \quad (6)$$

where $\psi(x)$ is the nonlinear mapping function of input x , $w \in R^n$ denotes a weight vector and b represents the bias value [19]. Regression is done through a nonlinear mapping of the training data to a higher-dimensional space over a kernel, where linear regression can be applied. The efficiency of the model depends on the kernel selection [31].

3.5. Evaluation Metrics

Performance measurement approaches, such as the Square Root of the Mean Square Error (RMSE) and the Mean Absolute Error (MAE), were applied to evaluate the ability of the proposed models to predict the PV power production in the system. Both MAE and RMSE are commonly used metrics to calculate the differences between measured and estimated values. The RMSE is calculated as the sum of the individual squared errors, while the MAE involves the sum of the magnitudes (absolute values) of the errors to obtain the 'total error' and then dividing by the number of errors [36]. Equations (7) and (8) describe the formulas for the selected performance measurements, where y_i is the observed experimental data, $\hat{y}_i$ is the estimated data, N is the total number of data and i denotes an index from 1 to N .

$$RMSE = \sqrt{\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N}}. \quad (7)$$

$$MAE = \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{N}. \quad (8)$$

4. Description of the PV Power Plant and Data Recollection

This work presents a performance comparison of data-based forecasting models for PV solar power prediction in Medellín, Colombia. As shown in Section 2, these approaches require historical data for applying a computational or statistical technique to develop a proper estimation of PV power production. To achieve this, the model that best represents the power generation system is evaluated, starting from the physical principles that provide the basis of studied phenomena and the historical meteorological and power generation data over a particular time. According to the literature [9,12,34], the study and analysis of the input data constitute a necessary previous stage to recognize the variables and correctly identify inputs and outputs of the process to model [15].

In this case, the output variable is the power generated from the PV array. Production of PV power generation systems is closely linked to incident solar radiation and climatic variables in the region where the system is implemented, such as temperature, relative humidity and wind speed [14,17]. Thus, these are the input variables considered for the system:

- **Solar Irradiance:** It is the energy emitted by the sun propagated by electromagnetic waves reaching the Earth's surface with different intensities classified according to its wavelength. This energy is measured per unit area of radiation that strikes a terrestrial surface for one unit of time [14]. This radiation is classified as direct, diffuse, reflected and global, the latter being the sum of the initial three. Direct radiation comes directly from the sun and only occurs when the sun is visible, while diffuse radiation occurs when the solar radiation disperses when going through the Earth's atmosphere. The reflected radiation is the part of solar radiation reflected by the Earth's surface.
- **Temperature:** It is a variable defined as the internal energy of a body. In the context of solar generation, the temperature has an essential effect on the value of the generated voltage [14]. For the analyzed system, temperature is measured in degrees Celsius (°C).
- **Relative Humidity:** It is defined as the amount of water vapor present in the air and is directly related to temperature. This variable is highly sensitive to changes, and its measurement is a percentage value from 0 and 100.

- Wind speed: Wind is the directional volumetric movement of the air, with an energy content dependent on its speed. For this system, wind speed is measured in meters per second.

4.1. PV System Description

For this analysis, statistical data were registered for six months of operation from a solar energy generation system connected to the electrical network (Grid Tied) in Medellin (Colombia). The plant consists of 52 poly-crystalline solar panels, each of 250 Watts of nominal power with a total nominal capacity of 13 kWp. Figure 5 shows an actual picture of the system.

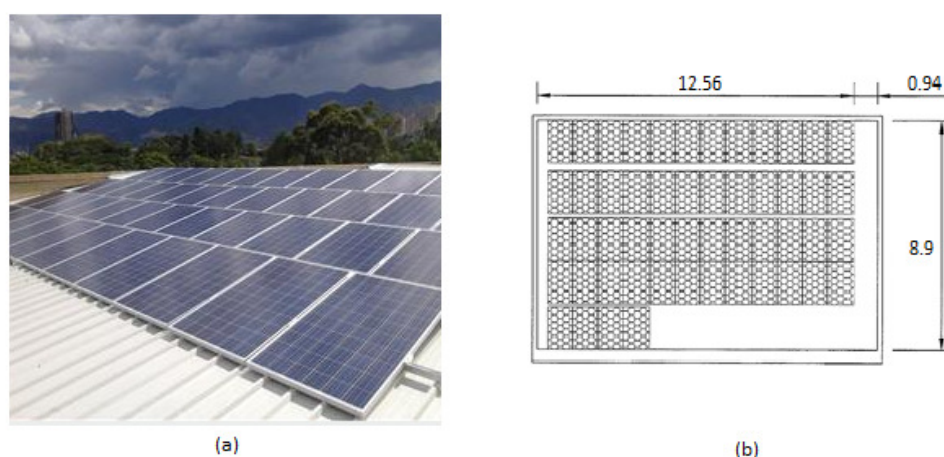


Figure 5. Photovoltaic (PV) plant located in Institución Universitaria Pascual Bravo at Medellin, Colombia: a picture of the PV panels (a); and the panel array on the library rooftop (b), with dimensions in meters.

4.2. Meteorological Variable Measurement

Another data collection stage involved capturing the information of solar irradiance, temperature, relative humidity and wind speed. For this task, we used the measurement stations of the environmental monitoring and attention system of Medellin (SIATA, the Spanish acronym). SIATA disposes of several monitoring stations dispersed around the city, equipped with multiparametric sensors providing minute-by-minute information of temperature, relative humidity, precipitation, atmospheric pressure, wind speed and wind direction [37]. Figure 6 depicts the SIATA Tower Meteorological Station, located approximately 3.2 km away from the solar plant.

4.3. Data Collection and Pre-Processing

The power generation database of the solar plant consists of the PV power plant output measured every 5 min for a six-month window between 1 January 2020 and 30 June 2020. Then, inconsistent values due to failures in the system's measurement devices were filtered and removed from the database. Figure 7 presents the PV power measurement for a one-week operation. The power profile is pretty stable, as Medellin weather tends to be relatively fixed and consistent, a behavior attributed to Colombia's closeness to the equator.



Figure 6. A picture of the Siata Tower Meteorological Station at Medellín, Colombia. Taken from [38].

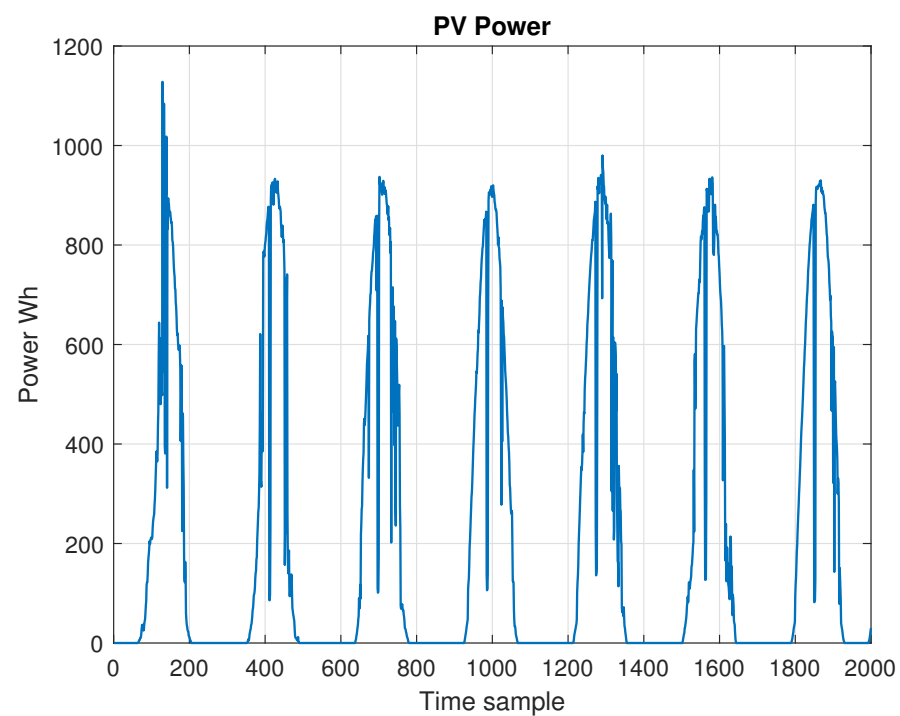


Figure 7. PV power output for the period between January 8 and January 15 2020.

The meteorological station database registered the minute-to-minute information of relative humidity, solar radiation, temperature and wind speed for the same observation period. The complete system database involves 50,198 samples in 5 min intervals with information of one output variable (PV power) and four input variables (relative humidity, solar radiation, temperature and wind speed). A correlation analysis using the Pearson coefficient (r) [39] was performed to corroborate the influence of input variables on the output, and the results are shown in Table 1.

Table 1. Values of Pearson coefficient between output variables and each on the input variables.

	Relative Humidity	Solar Radiation	Temperature	Wind Speed
PV Power	−0.0692	0.9122	0.0329	−0.0303

Pearson’s correlation coefficient is a measure of linear dependence between two quantitative random variables [39]. The r values show a clear dependence between PV power and solar irradiance, making the latter the most significant input variable affecting the system’s power generation. A r value close to zero is observed between PV power output and the other inputs. Although this indicates a weak linear relationship, it does not necessarily imply that the variables are independent: nonlinear relationships may still exist between the variables. Thus, relative humidity, temperature and wind speed remain as input variables for the proposed models despite the small correlation values.

The literature recommends a stage of data normalization, transforming the input and output variables into a fixed range, generally between 0 and 1. This step intends to handle the different measurement scales better and to improve the computational methods’ performance [15]. Thus, the database was normalized using Equation (9), where x_i denotes a variable, $x_{i,min}$ and $x_{i,max}$ are the minimum and maximum values of x_i and $x_{i,norm}$ is the normalized value of x_i . The index i denotes the number of variables in the system.

$$x_{i,norm} = \frac{x_i - x_{i,min}}{x_{i,max} - x_{i,min}}. \quad (9)$$

After normalization, the data were ready to be applied to the selected Machine Learning algorithms, as described in the next section.

5. Results and Discussion

Based on the analysis carried out in the previous sections, we compared four selected Machine Learning techniques to predict electricity generation from the PV power generation system of Institucion Universitaria Pascual Bravo. The techniques are K-Nearest Neighbors (KNN), Linear Regression (LR), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The developed models use input variables such as solar irradiance, temperature, relative humidity, wind speed and the historical data of PV power to estimate future PV generation. Solar irradiance in Medellín is available from 06:00 to 18:00 during almost the entire year.

Data were divided into training and evaluation subsets. Training data provided the first training of the algorithms, with a series of inputs (predictors) and known results (output), and the model used these data to relate the predictors with the results. The following are some considerations about the process of building the models:

- About 70% of data were used for the training stage, roughly corresponding to the data from January to April 2020 (35,138 samples). That left the data from May and June (the remaining 30%) to validate each method. The maximum PV power output recorded in the database was 937.67 W.
- ANN: The selected neural network is the feed-forward back-propagation type. After many tests with different network architectures, we found a configuration with four neurons in the input layer, fifteen neurons in the hidden layer and one output that offered the best results. The training involved the Levenberg–Marquardt Algorithm.
- KNN: The KNN algorithm was used as a regressor in this case, with $k = 5$ neighbors and Euclidean distance as the metric.
- LM: A linear model was developed from the training data, using Matlab function “fitlm” with five K-Fold for Cross-Validation.
- SVM: We used the Support Vector Machines as a regressor, employing Matlab function “fitrsvm” with a linear Kernel function and five K-Fold for Cross-Validation.

If the obtained relationship is not accurate enough, the model can be retrained until the process reaches adequate results. After that, the algorithms are validated with the testing portion of the data. In this step, the algorithm only receives the input data to test if the model can correctly make predictions. Once the predictions are made, the predicted data and actual observed data are compared to evaluate the better model.

Figure 8 displays the comparison graphs between the actual electrical energy generation and the prediction made by the Machine Learning techniques. These figures correspond to about four days from the months chosen to carry out the testing. As can be seen, the actual generation graph (blue color) and the prediction graph (orange color) almost overlap for all days analyzed with the different methods, which shows that the prediction made captures the trends in the actual data. Although all estimations look roughly the same, the SVM method is the only one whose estimation surpasses the peak values of actual data for each observed day.

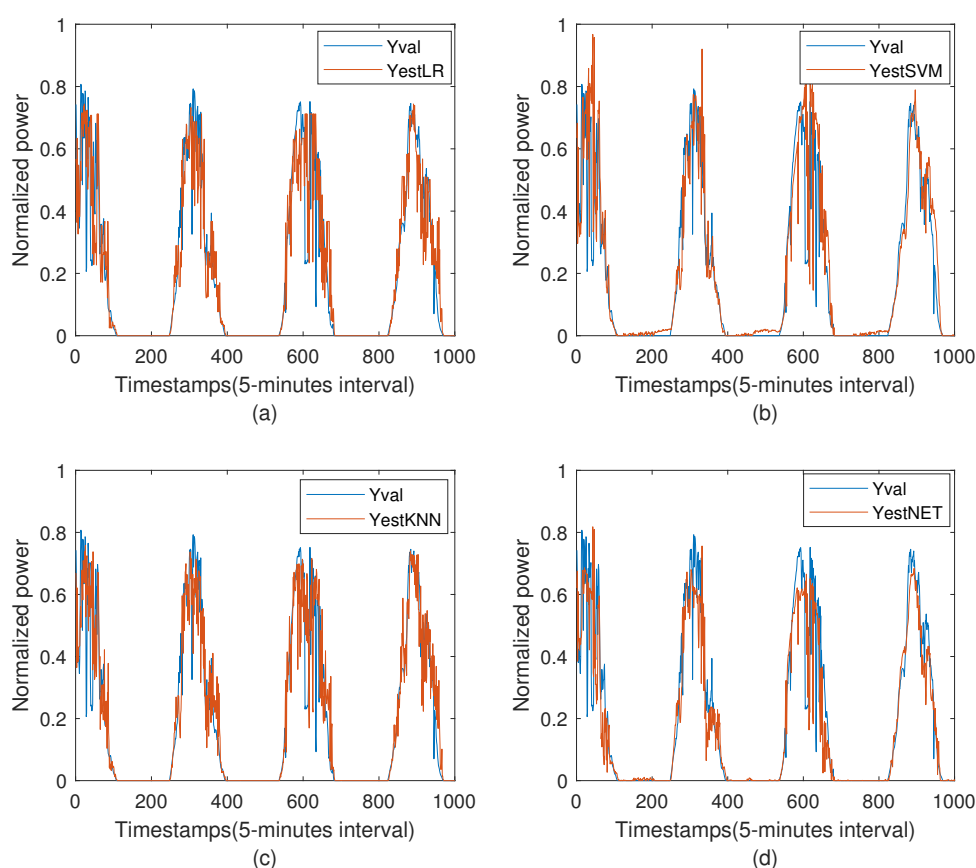


Figure 8. A 1000 sample graph of actual PV power data (blue) versus the estimation with each one of the techniques (orange): Linear Regression (a); Support Vector Machines (b); K-Nearest Neighbors (c); and Artificial Neural Networks (d).

Figure 9 shows one-day comparison graphs between the actual electrical energy generation and the prediction made by the Machine Learning techniques. The actual PV profile depicts a day with a high variation in the power production, probably due to a cloudy day. This causes a high variability of the power generated by the photovoltaic installation, which could add uncertainty in the prediction. From the curves, it is striking how the methods on the left side (LR and KNN) show a more variable (or distorted) prediction than those on the right side (SVM and ANN), whose estimations are much smoother.

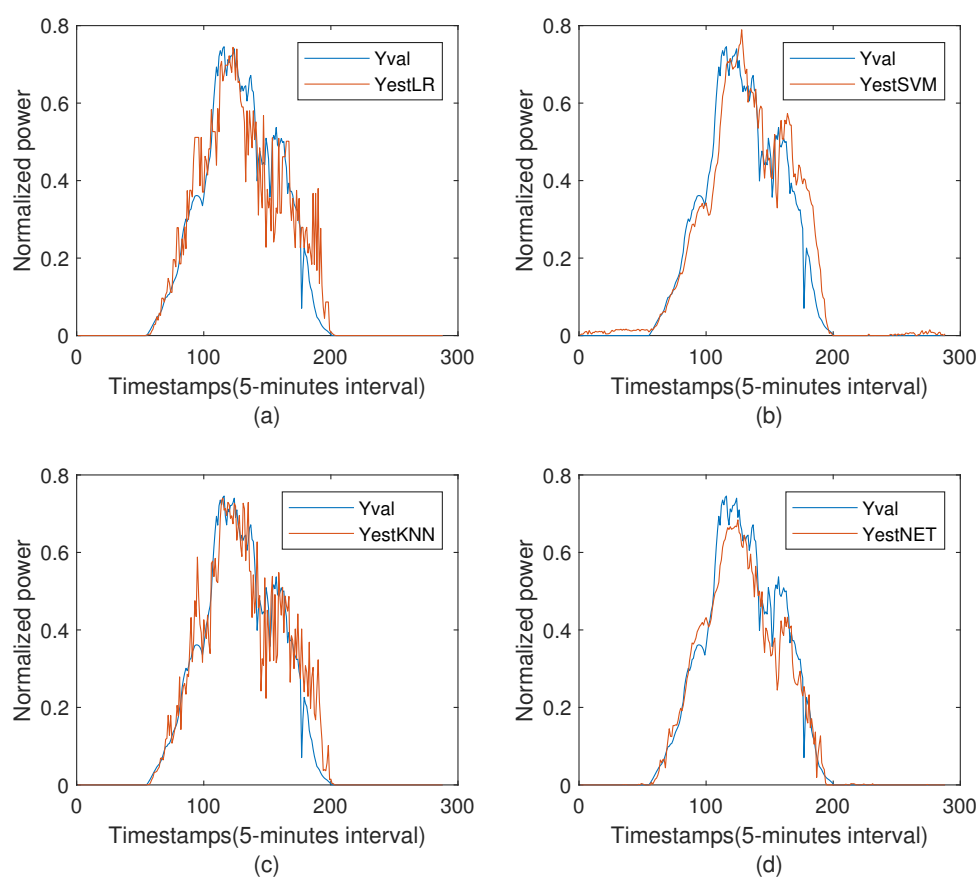


Figure 9. A one-day graph of actual PV power data (blue) versus the estimation with each one of the techniques (orange): Linear Regression (a); Support Vector Machines (b); K-Nearest Neighbors (c); and Artificial Neural Networks (d).

From these images, it is hard to judge which method performed better. Thus, the prediction error must be determined to assess the quality of the predictions made. Table 2 presents the values of RMSE and MAE for the compared methods.

Table 2. Values of error metrics for the validation of the compared methods. Units in Watts (W).

	KNN	LR	SVM	ANN
RMSE	92.857	94.583	93.644	86.466
MAE	8.8279	8.9632	9.6209	8.409

As observed in the table above, the values of RMSE and MAE for each of the proposed Machine Learning techniques are small (the maximum PV power production was 937.67 W), which shows that the prediction made for the test dataset is adequate. Both metrics show that the minimum estimation error is reached with the ANN model. On the other hand, the LR method produces the highest value of RMSE, while, according to MAE, the worse estimation performance belongs to the SVM model. The errors obtained for all the methods are very similar in magnitude. This could be attributed to the lack of drastic variations in the meteorological conditions of Medellín.

According to the figures and tables presented in this section, the four evaluated models produced adequate estimations of photovoltaic energy generation. In a previous study for Medellín [21], the electrical characteristics of a single 55 W-mono-crystalline silicon PV panel were modeled using ANN, reporting an average error of 1.6 W with a 50% confidence in the results. These values come from a different metric and experimental setups not entirely comparable with our study. Despite the lack of similar studies using Machine

Learning for PV power estimation in Medellín, the results can be compared with previously reported works found in the literature. In [9], a season-customized ANN is proposed to forecast the PV power of a system in Italy, with an average MAE of 17 W. The work of Das et al. [19] reports average MAE values of 33.63 W for SVR and 50.69 W for ANN estimation of PV power output in Malaysia. Values reported in Table 2 are at worst on the same level of the reported values, if we consider the different geographical conditions.

The modeling process involved data measured over six months, due to the restrictions in the availability of PV power measurements. This database size limits the prediction horizon of the models. Despite this, the validation results of the proposed Machine Learning based models for the PV plant in Institución Universitaria Pascual Bravo are also congruent with values previously reported in the literature. Although from different locations, those works also demonstrate the application of Machine Learning techniques to estimate PV power production and their importance in the increased usage of these resources around the world.

6. Conclusions

This work presents a performance comparison of supervised learning algorithms for PV solar power prediction in Medellín, Colombia. Four different forecasting models were generated using the techniques of K-Nearest Neighbors (KNN), Linear Regression (LR), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The performance of these methods was evaluated using data from a PV energy system located on the university campus of the I. U. Pascual Bravo in Medellín, Colombia.

The different methods used to predict electrical energy generation in photovoltaic solar parks are described by reviewing the international bibliography related to the subject. The analysis carried out shows that most of the studies recommend using Machine Learning methods to carry out this type of study due to their effectiveness in estimation and prediction tasks. Non-conventional power generation systems present a wide variation due to meteorological conditions, making supervised learning algorithms a valuable tool to predict the power generated from meteorological variables. As there is a need to obtain PV power production models to explore the potential of penetration of renewable systems in a city such as Medellín, Colombia, this study demonstrated Machine Learning techniques as practical and effective solutions to forecast PV power generation in Medellín.

The actual electricity generation values in the photovoltaic installation were compared with the predictions made by different methods (ANN, KNN, LR and SVM) for two months, reaching similar values in the comparison of error metrics. The comparison with reported studies showed a performance at worst on the same level if the different geographical conditions are considered. However, Artificial Neural Networks showed error values of 86.466 in RMSE and 8.409 in MAE, outperforming the techniques of K-Nearest Neighbors, Linear Regression and Support Vector Machines for PV power forecasting in Medellín.

Author Contributions: Conceptualization, L.G. and J.P.; methodology, L.G.; software, L.G.; validation, L.G., J.P. and E.D.-G.; formal analysis, L.G. and J.P.; investigation, L.G. and J.P.; resources, L.G., J.P. and E.D.-G.; data curation, L.G.; writing—original draft preparation, L.G.; writing—review and editing, J.P. and E.D.-G.; visualization, E.D.-G.; supervision, J.P. and E.D.-G.; project administration, J.P.; and funding acquisition, E.D.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in by Institución Universitaria Pascual Bravo by project number IN201904 and project number IN202106. The APC was funded by Institución Universitaria Pascual Bravo through the program “Convocatoria permanente para Article Processing Charges (APC) y/o corrección de estilo en idioma inglés”.

Acknowledgments: The authors thank Paola Ortiz for technical support.

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

Abbreviations

The following abbreviations are used in this manuscript:

PV	Photovoltaic
ANN	Artificial Neural Network
KNN	K-Nearest Neighbors
LR	Linear Regression
SVM	Support Vector Machine
RMSE	Root Mean Square Error
MAE	Mean Absolute Error

References

1. Patiño, J.; López, J.D.; Espinosa, J. Analysis of Control Sensitivity Functions for Power System Frequency Regulation. In *Applied Computer Sciences in Engineering*; Figueroa-García, J.C., López-Santana, E.R., Rodríguez-Molano, J.I., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 915, pp. 606–617. [\[CrossRef\]](#)
2. Patiño, J.; López, J.D.; Espinosa, J. Sensitivity Analysis of Frequency Regulation Parameters in Power Systems with Wind Generation. In *Advanced Control and Optimization Paradigms for Wind Energy Systems*; Precup, R.E., Kamal, T., Zulqadar Hassan, S., Eds.; Springer: Singapore, 2019; pp. 67–87. [\[CrossRef\]](#)
3. Abas, N.; Kalair, A.; Khan, N. Review of fossil fuels and future energy technologies. *Futures* **2015**, *69*, 31–49. [\[CrossRef\]](#)
4. Mohr, S.H.; Wang, J.; Ellem, G.; Ward, J.; Giurco, D. Projection of world fossil fuels by country. *Fuel* **2015**, *141*, 120–135. [\[CrossRef\]](#)
5. Dogan, E.; Seker, F. The influence of real output, renewable and non-renewable energy, trade and financial development on carbon emissions in the top renewable energy countries. *Renew. Sustain. Energy Rev.* **2016**, *60*, 1074–1085. [\[CrossRef\]](#)
6. Montoya Giraldo, O.D.; Grajales, A.; Grisales, L.F.; Castro, C.A. Ubicación y operación eficiente de almacenadores de energía en micro-redes en presencia de generación distribuida. *Rev. CINTEX* **2017**, *22*, 97–117. [\[CrossRef\]](#)
7. Lausset, C.; Lund, K.; Brattebø, H. LCA and scenario analysis of a Norwegian net-zero GHG emission neighbourhood: The importance of mobility and surplus energy from PV technologies. *Build. Environ.* **2021**, *189*, 107528. [\[CrossRef\]](#)
8. Niccolai, A.; Dolara, A.; Ogliaresi, E. Hybrid PV Power Forecasting Methods: A Comparison of Different Approaches. *Energies* **2021**, *14*, 451. [\[CrossRef\]](#)
9. Radicioni, M.; Lucaferri, V.; De Lia, F.; Laudani, A.; Lo Presti, R.; Lozito, G.M.; Riganti Fulginei, F.; Schioppa, R.; Tucci, M. Power Forecasting of a Photovoltaic Plant Located in ENEA Casaccia Research Center. *Energies* **2021**, *14*, 707. [\[CrossRef\]](#)
10. Tovar, M.; Robles, M.; Rashid, F. PV Power Prediction, Using CNN-LSTM Hybrid Neural Network Model. Case of Study: Temixco-Morelos, México. *Energies* **2020**, *13*, 6512. [\[CrossRef\]](#)
11. Badwawi, R.A.; Abusara, M.; Mallick, T. A Review of Hybrid Solar PV and Wind Energy System. *Smart Sci.* **2015**, *3*, 127–138. [\[CrossRef\]](#)
12. Das, U.K.; Tey, K.S.; Seyedmahmoudian, M.; Mekhilef, S.; Idris, M.Y.I.; Deventer, W.V.; Horan, B.; Stojcevski, A. Forecasting of photovoltaic power generation and model optimization: A review. *Renew. Sustain. Energy Rev.* **2018**, *81*, 912–928. [\[CrossRef\]](#)
13. Ahmed, R.; Sreeram, V.; Mishra, Y.; Arif, M. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renew. Sustain. Energy Rev.* **2020**, *124*, 109792. [\[CrossRef\]](#)
14. Sun, H.; Zhao, N.; Zeng, X.; Yan, D. Study of solar radiation prediction and modeling of relationships between solar radiation and meteorological variables. *Energy Convers. Manag.* **2015**, *105*, 880–890. [\[CrossRef\]](#)
15. Bou-Rabee, M.; Sulaiman, S.A.; Saleh, M.S.; Marafi, S. Using artificial neural networks to estimate solar radiation in Kuwait. *Renew. Sustain. Energy Rev.* **2017**, *72*, 434–438. [\[CrossRef\]](#)
16. Urrego-Ortiz, J.; Martínez, J.A.; Arias, P.A.; Jaramillo-Duque, A. Assessment and Day-Ahead Forecasting of Hourly Solar Radiation in Medellín, Colombia. *Energies* **2019**, *12*, 4402. [\[CrossRef\]](#)
17. Ruiz, S.; Patiño, J.; Marquez-Ruiz, A.; Espinosa, J.; Duque, E.; Ortiz, P. Optimal Design of a Diesel-PV-Wind-Battery-Hydro Pumped POWER system with the Integration of ELECTRIC vehicles in a Colombian Community. *Energies* **2019**, *12*, 4542. [\[CrossRef\]](#)
18. Ferrero Bermejo, J.; Gómez Fernández, J.F.; Olivencia Polo, F.; Crespo Márquez, A. A Review of the Use of Artificial Neural Network Models for Energy and Reliability Prediction. A Study of the Solar PV, Hydraulic and Wind Energy Sources. *Appl. Sci.* **2019**, *9*, 1844. [\[CrossRef\]](#)
19. Das, U.; Tey, K.; Seyedmahmoudian, M.; Idna Idris, M.; Mekhilef, S.; Horan, B.; Stojcevski, A. SVR-Based Model to Forecast PV Power Generation under Different Weather Conditions. *Energies* **2017**, *10*, 876. [\[CrossRef\]](#)
20. Chen, C.R.; Kartini, U. k-Nearest Neighbor Neural Network Models for Very Short-Term Global Solar Irradiance Forecasting Based on Meteorological Data. *Energies* **2017**, *10*, 186. [\[CrossRef\]](#)
21. Velilla, E.; Valencia, J.; Jaramillo, F. Performance evaluation of two solar photovoltaic technologies under atmospheric exposure using artificial neural network models. *Sol. Energy* **2014**, *107*, 260–271. [\[CrossRef\]](#)
22. Vrettos, E.; Kara, E.C.; Stewart, E.M.; Roberts, C. Estimating PV power from aggregate power measurements within the distribution grid. *J. Renew. Sustain. Energy* **2019**, *11*, 023707. [\[CrossRef\]](#)

23. Saleh, A.E.; Moustafa, M.S.; Abo-Al-Ez, K.M.; Abdullah, A.A. A hybrid neuro-fuzzy power prediction system for wind energy generation. *Int. J. Electr. Power Energy Syst.* **2016**, *74*, 384–395. [\[CrossRef\]](#)
24. Mirzapour, F.; Lakzaei, M.; Varamini, G.; Teimourian, M.; Ghadimi, N. A new prediction model of battery and wind-solar output in hybrid power system. *J. Ambient. Intell. Humaniz. Comput.* **2019**, *10*, 77–87. [\[CrossRef\]](#)
25. Kumar, K.P.; Saravanan, B. Recent techniques to model uncertainties in power generation from renewable energy sources and loads in microgrids—A review. *Renew. Sustain. Energy Rev.* **2017**, *71*, 348–358. [\[CrossRef\]](#)
26. Abdel-Nasser, M.; Mahmoud, K. Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Comput. Appl.* **2019**, *31*, 2727–2740. [\[CrossRef\]](#)
27. Wang, K.; Qi, X.; Liu, H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. *Appl. Energy* **2019**, *251*, 113315. [\[CrossRef\]](#)
28. Sun, Y.; Wang, F.; Zhen, Z.; Mi, Z.; Liu, C.; Wang, B.; Lu, J. Research on short-term module temperature prediction model based on BP neural network for photovoltaic power forecasting. In Proceedings of the IEEE Power & Energy Society General Meeting, Denver, CO, USA, 26–30 July 2015; pp. 1–5. [\[CrossRef\]](#)
29. Halabi, L.M.; Mekhilef, S.; Hossain, M. Performance evaluation of hybrid adaptive neuro-fuzzy inference system models for predicting monthly global solar radiation. *Appl. Energy* **2018**, *213*, 247–261. [\[CrossRef\]](#)
30. Yaniktepe, B.; Genc, Y.A. Establishing new model for predicting the global solar radiation on horizontal surface. *Int. J. Hydrogen Energy* **2015**, *40*, 15278–15283. [\[CrossRef\]](#)
31. Ahmad, M.W.; Mourshed, M.; Rezugui, Y. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy* **2018**, *164*, 465–474. [\[CrossRef\]](#)
32. Wang, J.; Li, P.; Ran, R.; Che, Y.; Zhou, Y. A Short-Term Photovoltaic Power Prediction Model Based on the Gradient Boost Decision Tree. *Appl. Sci.* **2018**, *8*, 689. [\[CrossRef\]](#)
33. Powers, J.G.; Klemp, J.B.; Skamarock, W.C.; Davis, C.A.; Dudhia, J.; Gill, D.O.; Coen, J.L.; Gochis, D.J.; Ahmadov, R.; Peckham, S.E.; et al. The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 1717–1737. [\[CrossRef\]](#)
34. Deossa, P.; Patino, J.; Espinosa, J.; Valencia, F. A comparison of Extended Kalman Filter and Levenberg-Marquardt methods for neural network training. In Proceedings of the Robotics Symposium, 2011 IEEE IX Latin American and IEEE Colombian Conference on Automatic Control and Industry Applications (LARC), Bogota, Colombia, 1–4 October 2011; pp. 1–5. [\[CrossRef\]](#)
35. Chu, Y.; Urquhart, B.; Gohari, S.M.; Pedro, H.T.; Kleissl, J.; Coimbra, C.F. Short-term reforecasting of power output from a 48 MWe solar PV plant. *Sol. Energy* **2015**, *112*, 68–77. [\[CrossRef\]](#)
36. Yousif, J.H.; Kazem, H.A.; Alattar, N.N.; Elhassan, I.I. A comparison study based on artificial neural network for assessing PV/T solar energy production. *Case Stud. Therm. Eng.* **2019**, *13*, 100407. [\[CrossRef\]](#)
37. SIATA. SIATA Website Download Portal. 2021. Available online: <https://bit.ly/2U1L7Tx> (accessed on 16 October 2020).
38. Escobar, J.D. Telemedellin Website. 2019. Available online: <https://telemedellin.tv/wp-content/uploads/2019/03/torre-siata.jpg> (accessed on 24 April 2021).
39. Hajjaj, C.; Alami Merrouni, A.; Bouaichi, A.; Benhmida, M.; Sahnoun, S.; Ghennioui, A.; Zitouni, H. Evaluation, comparison and experimental validation of different PV power prediction models under semi-arid climate. *Energy Convers. Manag.* **2018**, *173*, 476–488. [\[CrossRef\]](#)