Research Article

Machine Learning Approach on Time Series for PV-Solar Energy

S. Sivakumar, B. Neeraja, M. Jamuna Rani, Harishchander Anandaram, S. Ramya, Girish Padhan, and Saravanakumar Gurusamy

1. Introduction

The generation of power on a global scale is one of the primary challenges. Every nation in the world has access to nontraditional or renewable forms of energy in order to mitigate the issue. They can lower the cost of producing power by making use of renewable energy sources [1, 2]. In recent years, a number of nations have placed a significant emphasis on solar energy systems as sources of renewable energy [3]. The photovoltaic (PV) panel is the primary component in the solar energy system responsible for generating electricity [4–6]. Solar energy will lead to a large reduction in the amount of energy saved. Solar panels produce different amounts of energy depending on the circumstances when it is gloomy, wet, or bright outside [7, 8].

There are a wide variety of methods available for estimating the amount of energy produced by PV panels. On the other hand, the industry should establish the standard according to the capacity of the PV panels [9–11]. In order to enhance and more accurately estimate the energy produced by PV panels over the long term, a numerical weather prediction system is employed as an input. Machine learning will be used in this study, and it will use a variety of time series [12]. The accuracy of the energy projection in relation to the allotted time is essential for this task. The research findings indicate that both conventional methods and machine learning are applied in predicting energy use [13]. In order to make predictions about the performance of solar PV panels, the supportive vector approach is used.

The purpose of this study is to investigate several ways of forecasting in order to estimate the amount of energy...
produced by solar panels. The diverse climates may be understood via the use of machine learning and time-series approaches. This study aims to analyze the time series utilized to improve prediction accuracy by using machine learning methods [14]. This research has some limitations:

1. In the prediction, the developed model compares the different tools based on the performance in previous research forecasting [15, 16].

2. This research sets threshold values for the machine learning and time-series technique. Most of the existing techniques also have threshold values. So, the main aim is to introduce the relative performance of the various methods [16, 17].

The current work developed two methods. One is the time series model in the machine learning technique for short time series, and another is machine learning to optimize or predict the accuracy of the ANN model [18]. The entire work will give a comparative analysis of the machine learning techniques.

2. Time Series-Static Process

In time series, we should model the stationary process. Formally, the series of time \([t] \geq 0\), we can say that static process [19].

2.1. Autoregression Process. It depends on the historical values.

2.2. Moving Average Process. It is another regression process. Here, we do not consider historical values. Consider the prediction parameters, which depend on external shocks [20].

3. Artificial Neural Network (ANN) Functions

ANN can handle both regression and classification. ANN will work as a nerve system, but transferable mathematical functions establish that system [21, 22]. For regression, the process can choose only one, but the classification method can handle several classes. Each class can take several layers; each layer may combine the number of input and output variables or parameters and linear transformation. Some layers may be hidden layers. There is no prescribed structure to design the ANN method. But ANN can develop according to all parameters, as shown in Figure 1 [18, 23, 24].

The following can be seen from Figure 1:

1. We consider the original data as an input variable or parameter

2. We transform the input variable or parameter into the linear, which will activate the network

3. Finally, it gives the output in a prescribed format it may image structure; it is called final formation

3.1. Fitting Neural Networks

3.1.1. Backpropagation and Gradient. Batch learning is the discussion about the batch, which is added collection-wise at all the learning stages [25]. Online learning will discuss the learning of every individual data set. Conclusion: online learning will give more accuracy by comparing every data set and is less comprehensive and expensive [24].

3.1.2. Batch Learning and Online Learning. Batch learning is the discussion about the batch, which is added collection-wise at all the learning stages. Online learning will discuss the learning of every individual data set. Conclusion: online learning will give more accuracy by comparing every data set and is less comprehensive and expensive [26, 27].

4. Data Processing and Methodology

4.1. Data Collection. We collect data from small-scale solar PV-panels, which have 20–200 KW capacity. The energy output was sampled every 30 minutes from 4 different cities [28] (Table 1).

The weather data are prescribed in Table 2. This data are forecasted every 6 hours, i.e., 0–6 hrs, 6–12 hrs, 12–18 hrs, and 18–24 hrs. By using the method, we organized limited datasets. The data were stored with city names, which make the data prescribed with coordinating PV solar cells [29, 30].

4.2. Validating the Data. As handling the two different output data types is challenging, we change the input datasets for validation between sunny and cloudy days. Otherwise, we will get some errors, i.e., called local errors. We combine the input dataset with local errors to produce accurate forecasting to get the global errors [10, 14].

4.3. Data Processing. It depends on the collection of data and validation of data. We collected data every 6 hours. We also stored the data and converted it into prescribed data sets. Each data set should not have any missing parts. When total data is stored correctly, then the data should be kept as numerical, whether the production data or variables [17].

It is essential to eliminate the nighttime energy output. So, overall performance is also reduced due to the performance of PV solar cells, and night time data are considered the mean of the daytime and nighttime data. However, in most cases, it varies; for example, nighttime temperatures vary in summer and winter. More than 12 w/m² sky radiation was considered at the time [1, 20].

4.3.1. Visual Inspection. For all the data inspected, we should find out whether there are any same variables/data or not so that we can interpolate the following data concerning previous data.

4.3.2. Normalization Technique. When all the basic parameters are the same with respective times, solar PV-energy
data is also the same so that we can normalize the condition to reduce the nonstationary and we can reduce the data sets.

4.3.3. Assumptions

(1) If we want longer production, we must assume that solar PV energy is stored for one day, i.e., today’s energy is considered as yesterday only
(2) Numerical weather production data forecast according to city names may not be exact values
(3) Solar zenith angle can be estimated by the function of SAZ() in machine learning

4.4. Cross-Validation Technique. Cross-validation may refer to picking one data set from existing data sets randomly. It will test that data set from existing data sets. It will try that data set; the remaining data sets are learning data sets. A common choice is between 5–10 data sets. It will calculate the average performance of the data sets. It will take one for hyperparameters to determine the best performance matrices from cross-validation. But our total system depends on the time. So, there are several correlations to make the regulation cross-validation.

5. Forecasting Models

In this section, the step-by-step performance of different forecasting models is discussed.

5.1. Lasso Regression. In this technique, overfit the value in between the variables to zero. An algorithm is defined as feature selection performance in the algorithm selection process before modeling. This entire process is done in R package.

5.2. K-NN Model. The K-NN technique was performed with the help of R package; in the K-NN model and GBRT model, five different variables are used to reduce the time. The variables are chosen after executing the model in separate runs.

5.3. GBRT (Gradient Boosting Regression Trees). This package was performed with the R package.

5.4. ANN Model. In the ANN model, 2- and 3-layer networks are used with nine nodes per layer, and six active layers are fitted. The test was performed in several layers, and weight decay allows the ANN to reduce feature space to adopt the number of samples.

5.5. Persistence Model. It is used when more than one model is executed in the algorithm. The basic concept in this model is next coming value is equal to the current value.

5.6. Climatology Model. The climatology model is an average of 100 most recent data sets in this research. The last 100 data sets in 30-minute intervals will be used 6 hours ahead.

5.7. Performance Calculation System

5.7.1. RMSE. It is a common performance matrix:

$$RMSE = \sqrt{\frac{1}{S} \sum_{i=1}^{n} (V - \bar{V})^2}.$$  (1)
5.7.2. Normalized RMSE (nRMSE). It will compare the different five cities with high capacities to be calculated:
\[
n\text{RMSE} = \frac{\sqrt{\sum_{i=1}^{V} (V_i - \bar{V})^2}}{S \times \max(v)},
\]
where \( \max(v) \) represents the highest value of output.

5.7.3. Skill Score. It will compare one to other.
\[
SS = \frac{X_{\text{model}} - X_{\text{reference}}}{X_{\text{perfect model}} - X_{\text{reference}}} \times \frac{\text{RMSE}_{\text{model}}}{\text{RMSE}_{\text{reference}}},
\]
(3)

6. Results and Discussion

A higher shrinkage will be covered every 30 min in 6 hrs. The GBRT analyzes the data every 30 min for 6 hrs. RSME process is improved due to fitting all shrinkages. Note: most results are similar to their respective time and are eliminated.

From Figures 2–4, the primary problem with time series is that many parameters are not stationary, like time and weather reports. So, these parameters are considered in the time series model. From the analysis of the stored data, we can conclude that, in the winter season and on some rainy days, the energy output is meagre; at a specific time, like a night of the winter or at the time of rain, the energy output is almost equal to zero. So, when considering all these situations, it is tough to convert to time series. For this reason, we need to optimize the data sets to reduce the time series. Consider a more mathematical view and model the data for validation.

From Figures 5–13 and the observation of data sets, sometimes two different cities have the same result. It indicates that several data points are affected by ANN and GBRT models. Observing numerical weather production data and inputs takes a longer time horizon, represented in Figures 5–13.

ANN and GBRT are showing low performance. However, considering the KNN and Lasso models are more flexible in all models, it is also integrated into a more complex data structure.

When numerical weather production data contrast, nREME and gradient boosting regression tree get the better result. When comparing nRSME with others, we realize that a high reduction in all the results varies from 20% to 13% in a small gap for ANN. Using this, we can use numerical weather prediction data input to explain.

The developed data sets may not be sufficient for the better performance of the ANN and GBRT model when compared to Lasso and KNN. Lasso is relatively simple; it can analyze linear regression and does not analyze the nonlinear structure in the data sets. When ANN is compared to the other models, we considered weights randomly at the starting stage, and very few trials were conducted to get a well-optimized output. We need to increase the number of paths. However, we can finalize that ANN is the best in machine learning models by our tests.

It may be low when considering the average morning and night output. At the peak time, the output gets very high compared to all the times. In our research, we removed the peak time data from the stored data, which boosted the result. The major problem hears it is error fixing, so I implemented error fixing in this prediction.
Figure 5: City 1 RMSE with all models.

Figure 6: City 1 machine learning algorithm with all models.

Figure 7: City 2 RMSE with all models.

Figure 8: City 2 machine learning algorithm with all models.

Figure 9: City 3 RSME with all models.

Figure 10: City 3 machine learning algorithm with all models.
This methodology has improved the result based on considering a year. These data are divided into three seasons, i.e., summer, winter, and rainy; individual data are improved in the result. This method boosts the result by adopting the data in peak hours, both morning and night. Different models were performed based on climate change. Operational settings are changed from time to time.

7. Conclusion

This research considers the time to forecast solar energy in four cities using machine learning. Machine learning compares the number of models. Time series is complicated because all parameters are changed w.r.t. time. For this purpose, machine learning helps us forecast solar energy output using different models. In those models, we find ANN and GBRT performance is the best compared to remaining all models.

Data Availability

The data are available upon request to the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References