Photovoltaic energy prediction using machine learning techniques

Gonzalo Surribas Sayago[1][0009−0004−6988−4825], Jose David Fernández-Rodríguez[1][0000−0003−3702−2230], and Enrique Dominguez[1][0000−0002−2232−4562]

Dept. of Computer Science, University of Malaga, 29071 - Malaga, Spain
{surribasg,josedavid,enriqued}@uma.es

Abstract. Solar energy is becoming one of the most promising power sources in residential, commercial, and industrial applications. Solar photovoltaic (PV) facilities use PV cells that convert solar irradiation into electric power. PV cells can be used in either standalone or grid-connected systems to supply power for home appliances, lighting, and commercial and industrial equipment. Managing uncertainty and fluctuations in energy production is a key challenge in integrating PV systems into power grids and using them as steady, standalone power sources. For this reason, it is very important to forecast solar energy power output. In this paper, we analyze and compare various methods to predict the production of photovoltaic energy for individual installations and network areas around the world, using statistical methods for time series and different machine learning techniques.

Keywords: forecasting · photovoltaic energy · machine learning

1 Introduction

In the last years, dramatic drops in the total cost of ownership for many types of renewable energy power generation have translated into significantly increased rates of installed power generation, both standalone and connected to the power grid. In this regard, solar energy has grown enormously, and it is considered to still have a considerable growth potential, as more and more solar power is installed to help meet energy demands at a worldwide scale [1]. However, solar energy comes with serious challenges: its maximum power output is very susceptible to the amount of solar radiation reaching the solar panels' availability. As both power grids and standalone facilities require electric power flows to be as steady as possible, accurate forecasts of available solar radiation become very important for managing solar facilities. In the case of commercial operators directly selling their output into the electricity market, accurate predictions are even more relevant, as their profit margins can be significantly affected by inaccuracies in the predictions [4]. Numerous approaches have been proposed in the literature to predict the availability of solar radiation [5]. Most are based on simple, empirical mathematical models that are easy to compute. These

Year_Month	Power
2016-08-01	2.556462e+06
2016-09-01	1.926004e+06
2016-10-01	1.658863e+06
2016-11-01	1.755318e+06
2016-12-01	1.592755e+06

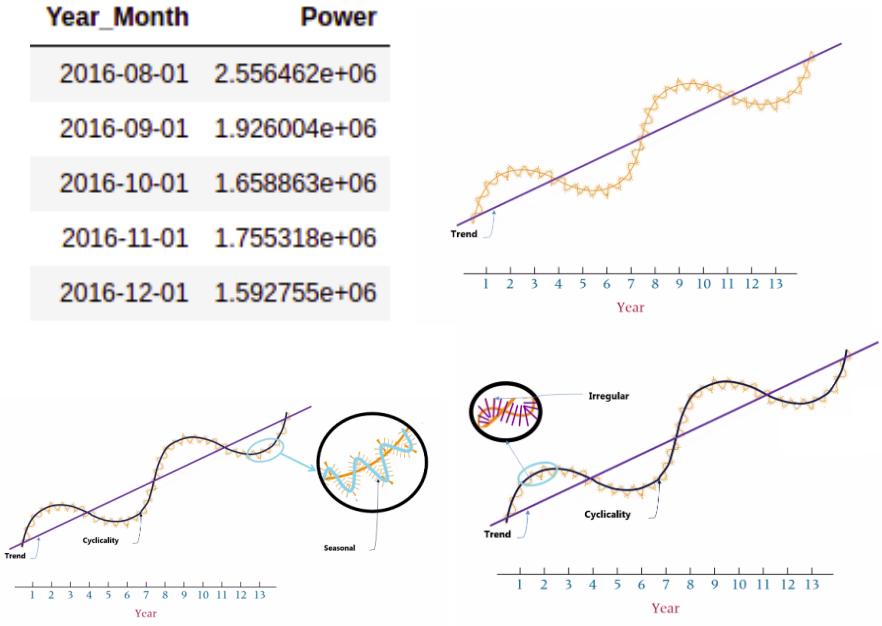

Fig. 1. Elements of a time series (Series, Trend, Seasonal, Irregular)

are widely regarded in the industry as valuable heuristics to predict average daily global solar radiation. Nevertheless, these simple models cannot accurately predict short-term solar radiation availability, as localized and rapid changes in weather conditions (such as cloud cover, intermittent rain, etc.) can significantly impact this availability. Furthermore, these models have been shown to be unable to reflect the complex and nonlinear relationships among dependent and independent variables in humid regions where solar radiation is strongly affected by heavy clouds throughout rainy days [2]. In this work, we propose several approaches based on machine learning to predict short-term solar radiation availability. The proposed techniques are analyzed, and their performance is compared using a common dataset.

2 Dataset and time series

The dataset used in this work has been provided by the SunLab platform [7], a collection of on-field PV laboratories installed throughout Portugal with the goal of characterizing the relative performance of various PV technologies. SunLab was set up by Energias De Portugal (EDP), a Portuguese power generation company, in order to support its business units in the acquisition of knowledge in the solar market field. The datasets provided by SunLab are organized by year: from 2014 to 2017. There are two datasets for each year: one with data

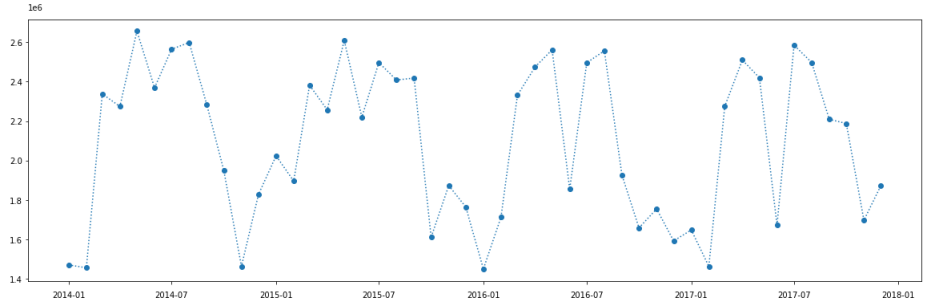

Fig. 2. Per-month SunLab power generation from 2014 to 2017.

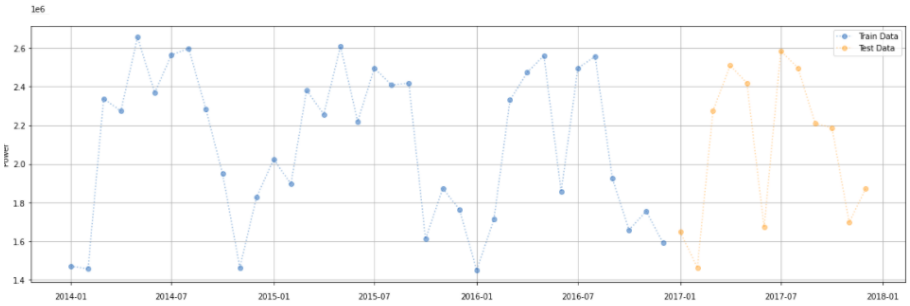

Fig. 3. The data from Figure 2 has been split into train and test datasets. In this case, the last year is set aside as testing data, and the rest as training data.

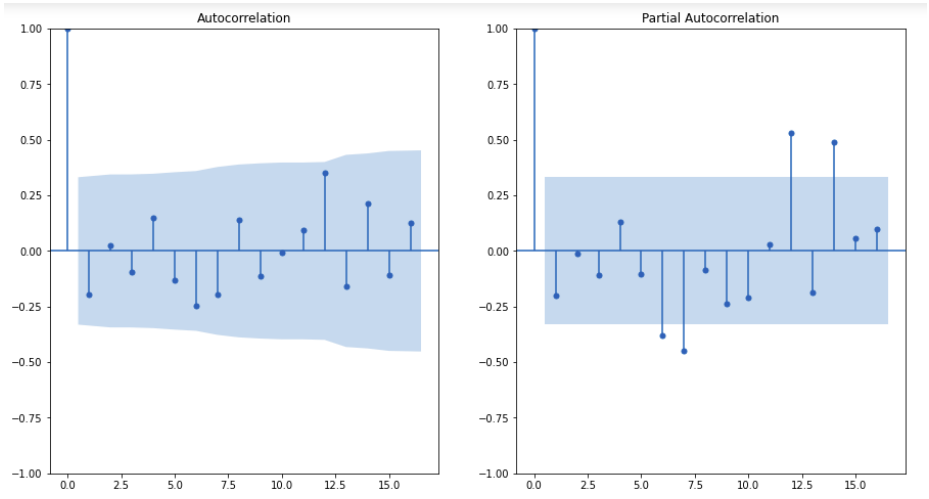

Fig. 4. Autocorrelation and partial autocorrelation functions for the training data in Figure 3. The highest non-zero value is at 12 in both functions.

from weather stations and the other with production and temperature data from the different PV modules. All these datasets are organized in time series, and all time series have a resolution of 1 minute.

Machine learning (ML) techniques can be applied to the time series to forecast solar radiation availability. Time series extracted from SunLab datasets are collections of observations of well-defined data elements obtained through measurements over time, such as the measurement of electricity production in a specific PV module. It is customary in data science environments to characterize time series from real-world data using the following concepts or elements (see Figure 1):

- Trend: long-term general direction of the time series.
- Cyclicality: repeating patterns of high and low values (cycles), typically over periods of multiple years.
- Seasonality: like cyclicality, but referring to shorter cycles, usually repeating with a frequency of one year.
- Irregularity: rapid changes ("bleeps") in the data, occurring in very short time frames.

Autoregressive models are very useful tools for analyzing time series data and performing forecasts. In order to fit an autoregressive model to a specific example, such as the total power generated at SunLab facilities per month from 2014 to 2017 (Figure 2), the data must be split into training and testing sets. In this case, the data corresponding to the last year (2017) is used as the testing set, with the rest used as the training set. Then, inspecting the autocorrelation and partial autocorrelation functions (Figure 4), we can identify the most significant autocorrelation value in order to apply a forecast for the test data (Figure 5).

3 Proposed Models

In this section, we present four machine-learning models and benchmark them by modeling per-day, fine-grained solar radiation availability, and computing a solar radiation forecast for a specific day used as testing data (May 26, 2015), while using the rest of the dataset as training data. The four models are multivariate linear regression (included as a baseline), decision trees, and two ensemble models based on decision trees: random forest regression and XGBoost.

3.1 Multivariate Linear Regression

Multivariate linear regression (MLR) is the simplest regression method considered in this work, used as a baseline. MLR can be regarded as a tool for building linear statistical models that characterize relationships among multiple dependent variables and multiple independent variables, and can be regarded as a collection of multiple linear regressions, each one for a different dependent variable, all of them sharing the same independent variables. Multi-linear regression

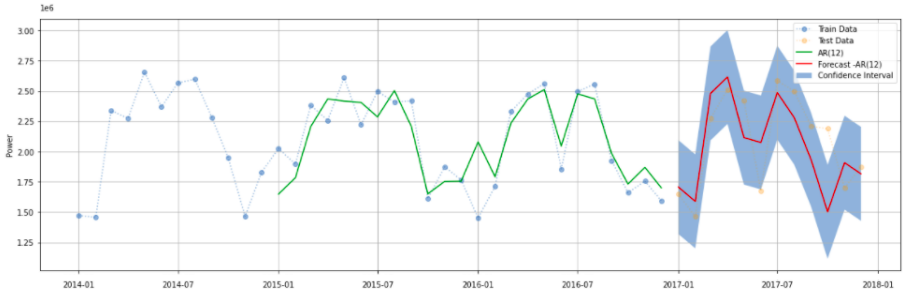

Fig. 5. Autoregressive Model with forecast based on largest autocorrelations (see Figure 4).

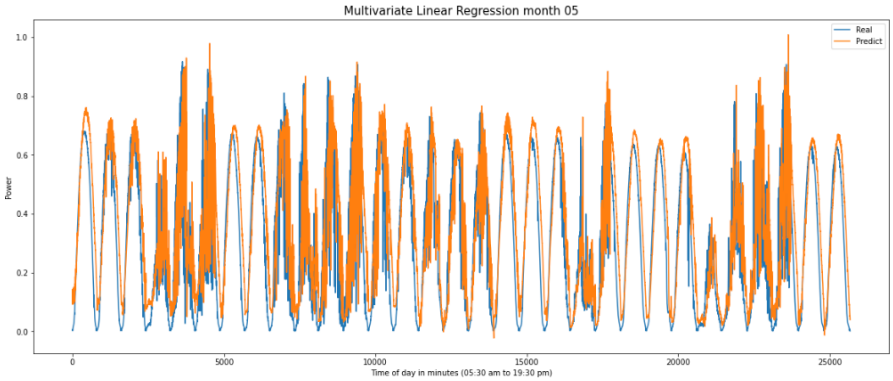

Fig. 6. Multilinear regression forecast for may 2015. The X-axis is in minutes; the Y-axis is the insolation coefficient). The X-axis covers the whole month, but only 14 daylight hours for each day (night periods are omitted from the series).

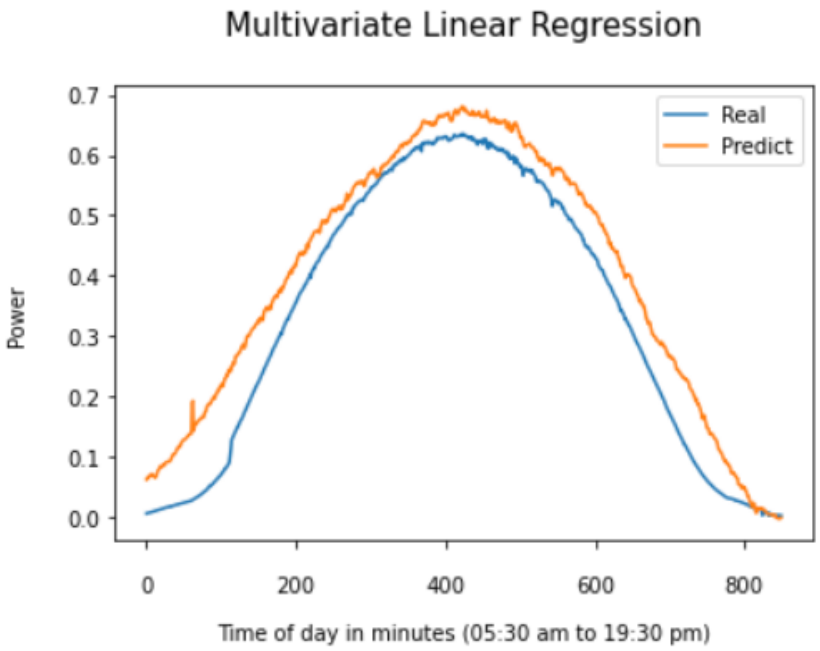

Fig. 7. Multilinear regression forecast for May 26, 2015. The X-axis is in minutes; the Y-axis is the insolation coefficient). The X-axis covers the 14 daylight hours (840 minutes) for that day.

can be written as $\hat{y} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_n * X_n$, where $\hat{y}$ is the dependent variable (predicted value), β_0 is the estimated intercept, and β_n is the n-estimated slope coefficient.

Related algorithms have been used for solar radiation forecast. Wang et al. [8] proposed a daily power output forecasting for PV facilities based on the Partial Functional Linear Regression Model (PFLRM) method. The PFLRM was integrated by using both functional and multiple linear regression models.

After fitting an MLR model to the training data, we can use it to forecast solar radiation availability in the testing data. In Figure 7, we can see the relatively large discrepancies between actual data (blue line) and the forecast (orange line) for the testing data (i.e., the daylight hours for May 26, 2015).

3.2 Decision Trees

Decision Trees are an important type of machine learning algorithm for predictive modeling, where a hierarchy of very simple regression models is built from the training data, so that samples are broken down into progressively smaller

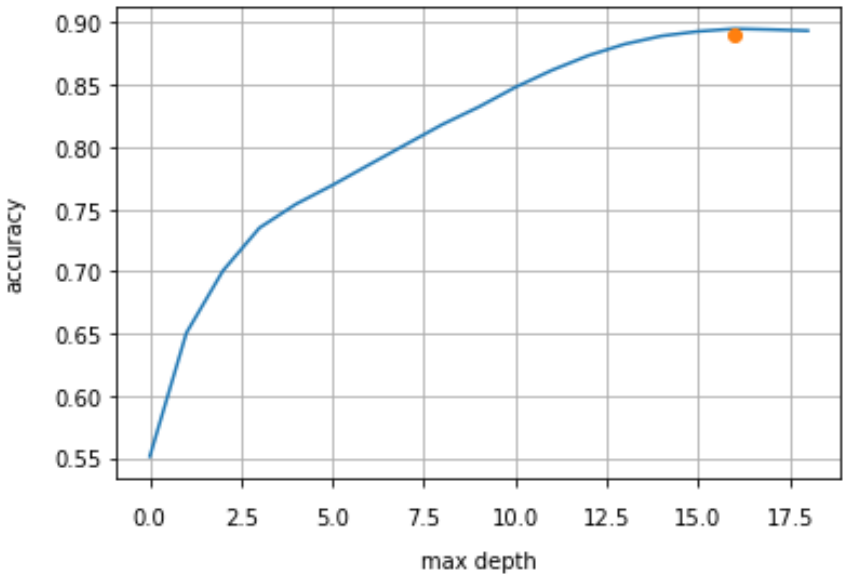

Fig. 8. Accuracy by decision tree depth for our SunLab training data.

subsets, according to a rule to minimize the prediction error, gradually building up an associated decision tree. In general, as the tree gets deeper, the accuracy increases.

This type of algorithm has already been used by other researchers to predict energy production and its relationship with climatic factors. In [3], authors aimed to predict the output production of solar power plants in kWh and how climatic factors influence that production. The maximum achieved accuracy was around 81% with a maximum tree depth value of 8. However, when using a decision tree to model the SunLab dataset, we found an increased accuracy of 94%, with an optimal depth of 16 (see Figure 8). With this decision tree model, the predictions for the testing set (May 26, 2015) are significantly better than with the baseline method (see Figure 9).

3.3 Random Forest Regression

Random Forest Regression can be regarded as a generalization of decision trees: The algorithm works by building multiple decision trees at training time (in parallel, with no interaction between the trees) and using the average of the outputs of the trees as its prediction. When applying this method to our SunLab dataset, a maximum accuracy of 0.95 was found, averaging results from 300 trees

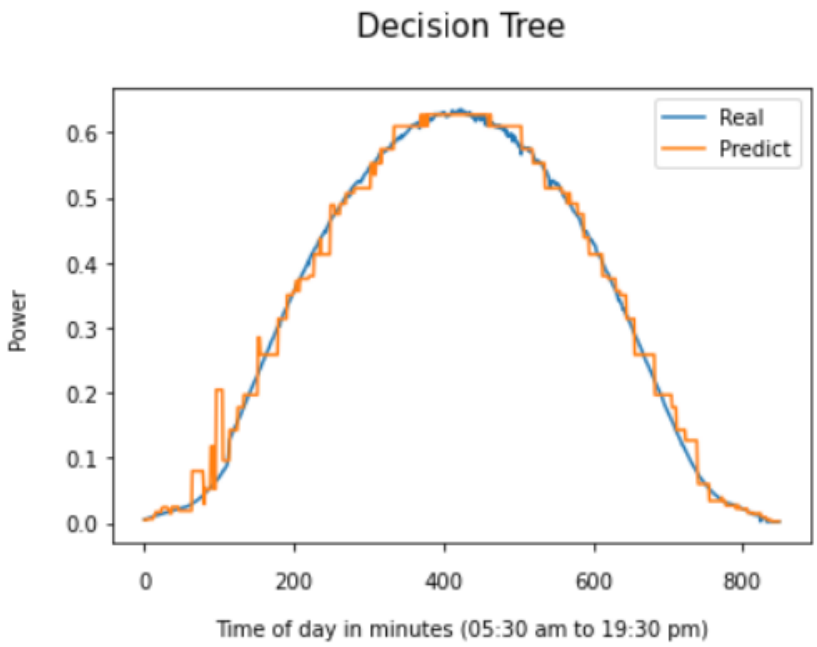

Fig. 9. Decision tree forecast for May 26, 2015. The X-axis is in minutes; the Y-axis is the insolation coefficient). The X-axis covers the 14 daylight hours (840 minutes) for that day.

at a tree depth of 16. The forecast of this aggregate model for May 26, 2015 is shown in Figure 10.

3.4 XGboost (Extreme Gradient Boosting)

Gradient boosting is a family of machine learning algorithms to build an ensemble of models that significantly outperform any models in the ensemble. Typically, the base models are decision trees, so gradient boosting can be regarded as a generalization of these. Broadly speaking, gradient boosting algorithms build decision trees one after another. Crucially, each decision tree is not independent of the rest but is built and fit to correct the prediction errors from previous trees, such that each new tree refines the predictions from previous trees. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "gradient boosting," as the loss gradient is minimized as the model is fit, much like a neural network.

XGBoost is a specific implementation of gradient boosting, deploying a wide array of optimizations for speed and performance. XGBoost is regarded as a

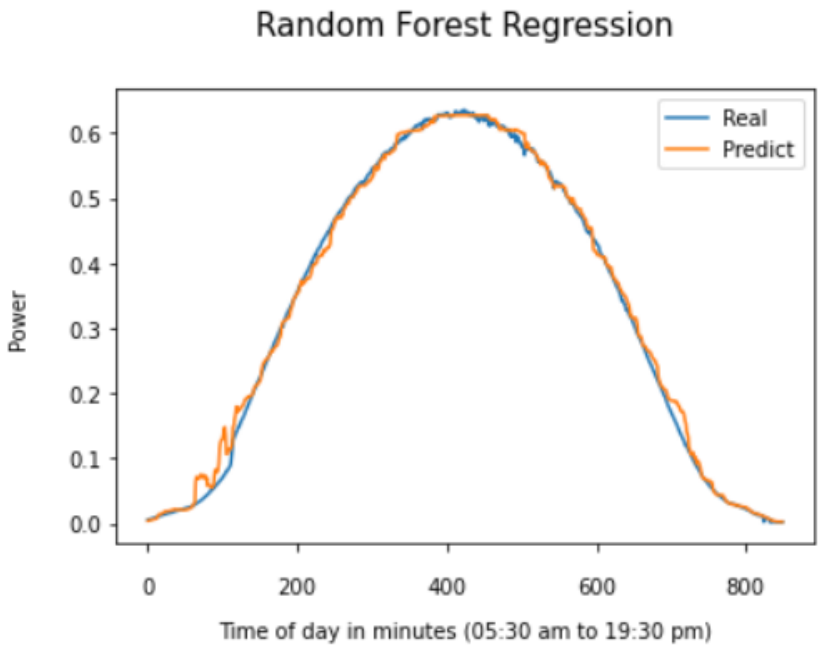

Fig. 10. Random Forest Regression forecast for May 26, 2015. The X-axis is in minutes; the Y-axis is the insolation coefficient). The X-axis covers the 14 daylight hours (840 minutes) for that day.

very competitive implementation of gradient boosting, being used by the winners of many machine learning contests. Obiora et al. [6] obtained very good results predicting solar radiation with this algorithm. When applying XGBoost to our training data, the maximum accuracy is 0.96, using 360 estimators with a maximum tree depth of 8 (see Figure 11).

4 Experimental Results

Results from the methods discussed in the previous section are gathered together in Table 1. This Table shows the best accuracy achieved with each method. Linear regression models perform significantly worse than the rest; this is to be expected, since solar radiation does not linearly depend on the variables in the SunLab dataset. Accordingly, vanilla decision trees perform significantly better, and the ensemble models provide additional increases in accuracy. While the increases might not seem substantial (Random Forest Regression increases the accuracy over decision trees in 0.01, and XGBoost in 0.02), they are pretty

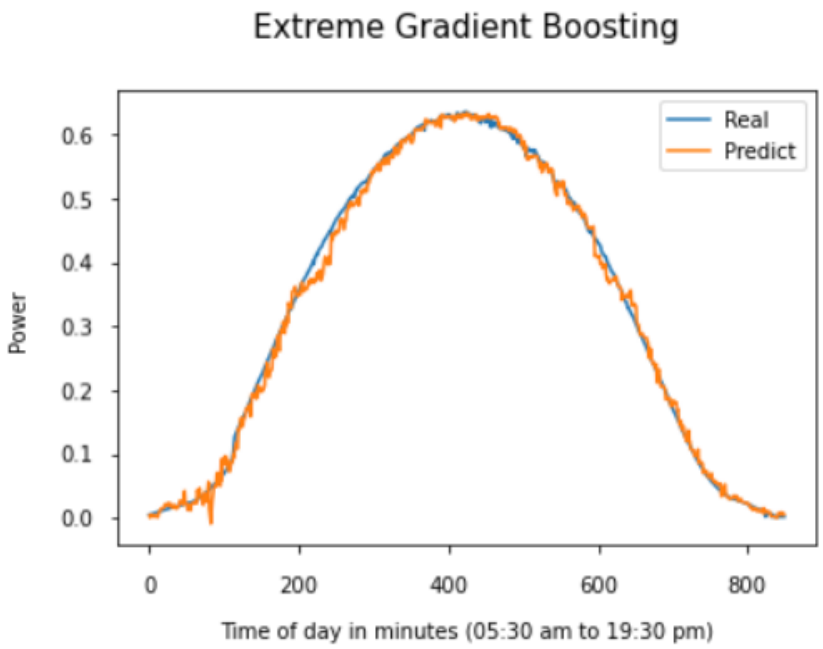

Fig. 11. Extreme gradient boosting forecast for May 26, 2015. The X-axis is in minutes; the Y-axis is the insolation coefficient). The X-axis covers the 14 daylight hours (840 minutes) for that day.

significative, as can be seen by comparing the discrepancies between actual and predicted solar radiation in Figures 9 (vanilla decision tree), 10 (Random Forest Regression) and 11 (XGBoost).

5 Conclusions

A lot of research has been directed at optimizing power generation in PV facilities at multiple levels. In particular, the sizing of a PV installation (number of PV modules, storage and inverter capacity, etc.) is a crucial part of the PV system's design, as a correctly sized PV facility with proper energy storage scheduling is a more stable source of electric power, and thus can be more effectively used both as a standalone power source and as a power plant for a public power grid. In this context, accurate forecasting of weather conditions that may affect solar radiation availability can become a very effective tool, not only for more effective optimization of PV facility size but also for managing the balance between power

Table 1. Results of the proposed techniques.

Method	R^2 score
Linear Regression	0.73
Decision Tree	0.94
Random Forest Regression	0.95
XGBoost	0.96

generation and load demand, as balance problems can destabilize the power grid and cause significant economic losses.

This work shows that ensemble algorithms based on decision trees can achieve excellent results in forecasting solar radiation availability. The best accuracy has been achieved using XGBoost with a maximum tree depth of 8 and an ensemble size of 360. While achieving slightly less accuracy, the Random Forest Regression algorithm can achieve results almost as good, with the benefit of being significantly simpler and less computationally intensive to train. In future work, we expect to achieve even better accuracy by using deep learning models built from the ground up to model and effectively generalize time patterns, such as LTSM networks.

References

1. Choudhary, P., Srivastava, R.K.: Sustainability perspectives-a review for solar photovoltaic trends and growth opportunities. Journal of Cleaner Production **227**, 589–612 (2019)
2. Fan, J., Wang, X., Wu, L., Zhang, F., Bai, H., Lu, X., Xiang, Y.: New combined models for estimating daily global solar radiation based on sunshine duration in humid regions: a case study in south china. Energy Conversion and Management **156**, 618–625 (2018)
3. Gupta, A., Bansal, A., Roy, K., et al.: Solar energy prediction using decision tree regressor. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). pp. 489–495. IEEE (2021)
4. Gürel, A.E., Ağbulut, Ü., Biçen, Y.: Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation. Journal of Cleaner Production **277**, 122353 (2020)
5. Liu, Y., Zhou, Y., Chen, Y., Wang, D., Wang, Y., Zhu, Y.: Comparison of support vector machine and copula-based nonlinear quantile regression for estimating the daily diffuse solar radiation: A case study in china. Renewable Energy **146**, 1101–1112 (2020)

6. Obiora, C.N., Ali, A., Hasan, A.N.: Implementing extreme gradient boosting (xgboost) algorithm in predicting solar irradiance. In: 2021 IEEE PES/IAS PowerAfrica. pp. 1–5. IEEE (2021)
7. Sunlab: Edp open data. https://opendata.edp.com/open-data/en/data.html, accessed: March 2023
8. Wang, G., Su, Y., Shu, L.: One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models. Renewable Energy **96**, 469–478 (2016)