A Novel Hybrid Model Proposal derived from Prevalent Methods for Power Generation Prediction of Solar Power Plants

Necati Aksoy[1] and İstemihan Genç[1]

[1] Department of Electrical Engineering, Istanbul Technical University, Istanbul, Türkiye
aksoyn18@itu.edu.tr, gencis@itu.edu.tr

Abstract

Renewable energy sources play a pivotal role in contemporary distributed energy generation, owing to their significance in reducing energy costs and mitigating carbon emissions. Ensuring predictability in energy demand and production is crucial for effective future planning, wherein intuitive predictions for renewable energy sources are indispensable. In this study, we propose a novel method for predicting power generation in solar power plants. We develop a hybrid prediction model by combining prevalent machine learning models trained with meteorological data, yielding superior results compared to individual model outcomes. Through analysis, we evaluate the performance of the models trained with real meteorological and production data, while emphasizing the advantages of the proposed hybrid approach. The proposed method offers valuable insights into enhancing the predictability of solar power plant generation, thereby contributing to the advancement of renewable energy utilization.

1. Introduction

Renewable energy sources, with their feature of reducing dependence on fossil fuels, appear as a powerful solution to the energy crises that will confront humanity in the future. Reducing reliance on fossil fuels will not only provide flexibility in the use of energy, but also promise fairness to energy sharing around the world. On the other hand, power plants using the sun, which is an almost inexhaustible source of energy, are the leading factor of renewable energy sources. While the technologies for converting solar energy to electrical energy are in continuous development, the issue of efficient use of conventional solar panels is also in academic rising trends.In this regard, increasing the efficiency of the energy generation is directly related to the level of use of solar energy, specifically the radiation. While sensor-based solar tracking systems emerge as a solution to increase the use of radiation on solar panels, estimating the energy that the panels will generate with certain radiation level is another approach. Prediction can be made intuitively or rule-based, or it can be based on artificial intelligence (AI). Just in this regard, machine learning, one of the sub-titles of AI, includes various well-working prediction algorithms. Prediction can be made intuitively or rule-based, or it can be based on artificial intelligence. Just in this regard, machine learning, one of the sub-titles of AI, includes various well-working prediction algorithms. Integration of these machine learning (ML) techniques into solar power forecast is one of the topics studied [1]. In addition, the advantages of deep learning, which is a special sub-topic of machine learning, for solar power predictions are among the subjects studied [2]. Direct satellite photos can be used for the dataset for forecasting [3], and meteorological data can be preferred for training forecast models together with satellite images [4]. In addition to estimating the generation of a solar power plant, predictive models have also been developed in the light of data collected from a single panel [5]. Moreover, forecasting the energy generation of only a solar panel or solar power plant, studies focusing on the power generation prediction of a solar power plant in a microgrid and the effects of these results on the microgrid have also been carried out [6].

It is not enough to create prediction models for power generation alone, the performance outputs of these models should be examined and improvements should be made accordingly. Although the created model provides an estimate, this estimate may not be as accurate as desired. Or, the prediction model may fail with sharp changes in power generation. Furthermore, [7] examines the performance of three innovative forecasting algorithms based on gradient boosting machine, trained with weather data, in solar power forecasting. It has been claimed in some studies that bootstrapping can improve models [8]. The combined use of more than one forecasting model, which was put forward as another approach, emerges as an alternative to the mathematical improvement of a single model. This kind of solution proposal, which is generally called a hybrid model, focuses on the use of predictive models that use the same algorithm but are created with different parameters [9], [10]. Although this approach allows for certain improvements, the combined use of forecast models created using different algorithms has yielded more satisfactory results [11], [12].

Considering all these improvements in solar power generation prediction, this study offers an innovative proposal for the combined use of prediction models created using different machine learning algorithms. Models are combined with a voting approach in creating hybrid models and the results are compared with models that would be used alone.

2. Predictive Techniques

This section briefly describes the four prevalent machine learning-based forecasting methods that make up the hybrid forecasting model proposed in this paper.

2.1. Polynomial Regression

Polynomial regression, which is a special type of linear regression, or rather multi linear regression, is preferred when the effect of the data on the result is not linear.First, let's look at the equation of the linear regression model:

$$\mathbf{Y} \approx \beta_0 + \beta_1 \mathbf{X} + \epsilon \tag{1}$$

where; **Y** is label vector and **X** is input vector This model, which is used to predict outputs with an input data, has been enriched to establish a linear connection between multiple inputs and outputs. The equation called multi linear regression:

$$\mathbf{Y} \approx \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \cdots + \beta_n \mathbf{X}_n + \epsilon \tag{2}$$

This regression equation, which produces a suitable estimation result for many problems, may be insufficient in cases where the relationship between input data and output is not linear. In this case, the polynomial regression equation comes into play:

$$\mathbf{Y} \approx \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2^2 + \beta_3 \mathbf{X}_3^3 + \cdots + \beta_n \mathbf{X}_n^n + \epsilon \tag{3}$$

This type of regression can give appropriate results in creating an appropriate estimation model for the problem. The prediction model specific to this study using meteorological data was created using polynomial regression and 4 degrees were used.

2.2. Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression problems. However, it is mostly used in classification problems.The regression type of this algorithm, which was preferred in this study to create a prediction model, is called support vector regressor (SVR). SVR is a supervised learning algorithm that is used to predict discrete values. SVR uses the same principle as the SVMs. The fundamental idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points. Further, the main purpose in SVR is to find a kernel that will contain as many points as possible with support vectors that will fit on the label points as much as possible. The aim here is to get as many data points as possible

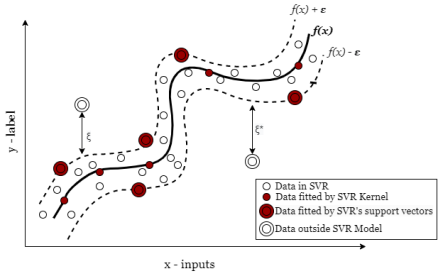

Fig. 1. Support Vector Regression

between the support vectors with the help of the kernel function to be selected, and to minimize the error between the data outside the support vectors and also, in the support vectors 1.

2.3. Decision Tree

Decision Tree (DT) is a machine learning algorithm based on the logic of dividing a dataset of variables into certain regions. Although it is generally used for classification problems, it can also be preferred as a regression tree for regression problems and good results can be obtained. A tree structure is created from the sections created from the dataset.It is used that regression trees as the prediction algorithm and the goal is to generate non-overlapping regions using the learning set variables. $X_1, X_2, X_3, \ldots X_p \longrightarrow R_1, R_2, R_3, \ldots R_j$ where: **Xs** represent features of the dataset and **Rs** are regions that the DT algorithms creates. Therefore, the estimation is the average of the output values (y_j) of the training data in the R_j region, for example, the j region. So how do we decide on these regions ($R_1, R_2, R_3, \ldots R_j$) ? At this point, our model uses the Residual Squares Error function as in linear connection. Which is:

$$\sum_{j=1}^{J} \sum_{j \in 1} \left(y_i - \hat{y}_{R_j}\right)^2 \tag{4}$$

$\hat{y}_{R_j}$: average of outputs in R_j region
In decision trees,regions are separated according to the recursive binary splitting method. Here is the formula:

$$R_1(j,s) = \left\{X \| X_j < s\right\} \quad R_2(j,s) = \left\{X \| X_j \geq s\right\} \tag{5}$$

Here we are trying to find the j and s values that minimize the following expression.

$$\sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 \tag{6}$$

After dividing into regions, if there are very few points left in any region, it will be stopped.

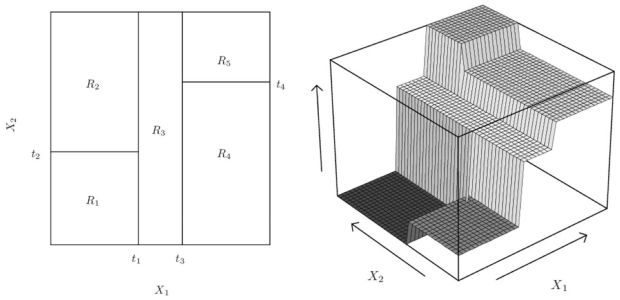

Fig. 2. Identifying regions in DT

If the values shown by the upward arrow in the right figure in Fig. 2 are considered as the average values of the separated regions, they show our estimation values. If we divide our data into too many regions in our learning set, our model will over-learn (be overfitted) because it will memorize each data and its output value. In order to prevent over-learning, we do tree pruning, which is basically based on adjusting the size of our tree.

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} \left(y_i - \hat{y}_{R_j}\right)^2 + \alpha |T| \tag{7}$$

where:

- $T \subset T_0$: Subtree
- $|T|$: Number of terminal nodes in T tree
- R_m :the region corresponding to the terminal node m
- α: hyperparemeter

In order to optimize the output of the formula, it should be either reduced the number of nodes in the tree or our squared error function. The value that determines the size of the tree is done by k-fold cross validation, in this way the most appropriate value is determined and the best prediction result is obtained.

2.4. Random Forrest

The random forests algorithm is basically based on the decision trees method and is an Ensemble Learning (EL) method. Bagging which is the branch of EL is when constructing multiple decision trees in the same learning set and reflecting the average of their outputs as the result. The Random Forest method aims to reduce the similarity between the randomly selected sub-datasets from the dataset and the trees built by establishing smaller trees. The average of the results from more than one tree with reduced correlation returns to us as output data. First of all, this algorithm can be used in both classification and regression problems and gives good results. The second advantage is that overfitting or over-learning is a critical problem that adversely affects the results, but for the Random Forest algorithm, if there are enough trees in the forest, the probability of the overfitting problem will decrease. In the use of this algorithm, the number of random trees can be entered as a hyperparameter. Appropriate selection of this parameter changes the quality of learning. This forest structure needs to be designed specifically for the problem. The number of trees in the forest to be established and the level of randomness in the random mixing of the data are very critical for the correct solution of the problem.

3. Implementation and Proposed Method

3.1. Dataset

The power generated by a solar power plant depends on two main variables. These are the material properties of the panels used in this power plant and changes in weather conditions. The production characteristic of the panels affects the power generation stably, not dynamically. However, instantaneously changing weather data affects the amount of energy generated dynamically, a positive or a negative way. The meteorological data collected in less than five minutes for the past five years and the total energy produced by the solar power plants operating in the region where these data were obtained are used in the training of the models. Before starting the models' training, data preprocessing was performed and missing data were filled with appropriate data. In addition, the time intervals were rearranged as 15 minutes, 30 minutes and hourly. The dataset tuned at these three frequencies is saved separately. Rather than using all weather data types, four data types that are thought to affect generation directly are used. These are: ambient temperature, global radiation, diffuse radiation and ultraviolet data. Fig 3 shows how the ambient

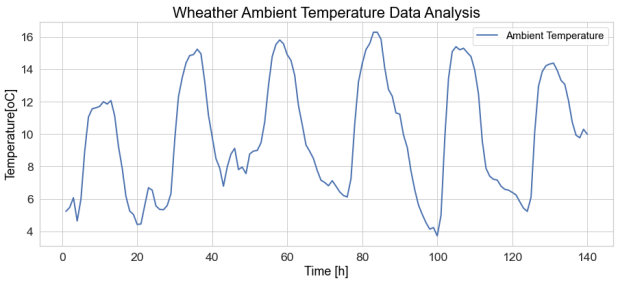

Fig. 3. Ambient Temperature for randomly selected 140 hours

temperature changes for a randomly selected 140 hours. The ambient temperature changes the temperature of the panels, which is to change the amount of energy produced. In addition, it can be said that the data type that most affects the energy produced is radiation data. If these data are to be examined. Three different radiation types are visualized together in Fig. 4. for 140 randomly selected hours. Global radiation is the most dominant type of radiation. Diffuse radiation comes second and ultraviolet comes third. It can be thought that the global radiation data will be the factor that affects the production the most.

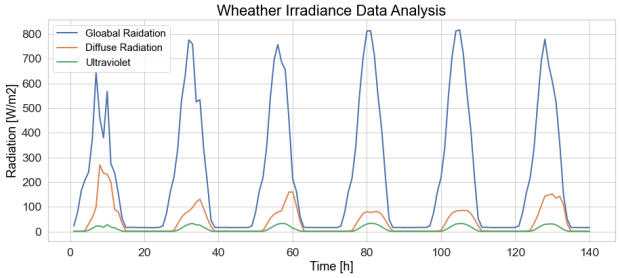

Fig. 4. Radiation Data for randomly selected 140 hours

It can be deduced from the similarity between Fig. 4 and power generation from panels that there is a considerable relationship between the variation of energy produced from solar power plants and the radiation data. In addition to global radiation, it can be deduced that diffuse radiation and ultraviolet radiation level also have an effect on power generation. In the light of all these inferences, ML models are trained using these data and the label.

3.2. Evaluation Methods for Model Performances

Loss (error)functions, which are frequently used in many different branches of AI, are basically used to evaluate the trained model. While it is used to readjust the weights and biases of the neural networks built with back-propagation in deep learning, it is used to see the difference between the outputs of the model and the actual values in ML. If the mean square error (MSE), one of the most used loss functions, is to be examined:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

Where: y_i represents actual data and $\hat{y}_i$ represents predicted values for index i. MSE is a very powerful metric for evaluating models. Another similar error function is root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

While MSE shows the average squared difference between the predicted values made by the model and the actual values, RMSE indicates the average deviation between the predicted points scored and the actual points. So, these two metrics give us useful information about different model outputs.In addition to these two loss functions, mean absolute error (MAE) is also used in the study.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

These three useful error outputs are used both in the improvement of the created models and in the co-use (unification) strategy.

3.3. Proposed Method

Creating a prediction model, training it and using it once satisfactory results are obtained is a subject that has been studied many times. In addition, studies that train a single model with different kernel functions and hyperparameters a few times and then use it are another approach that produces positive results. Creating a hybrid forecasting model, which is another important approach, is one of the subjects studied. Similarly, in this study, the issue of creating more than one ML model and using them together is emphasized. Models are created using the algorithms described in the second section and the dataset described in the title *A.Dataset* of this section. All models are tested with the test dataset and predictive values are obtained. These models are combined with an innovative voting method. This is the proposed unification method:

$$\hat{y} = \frac{R_1^2.y_1 + R_2^2.y_2 + R_3^2.y_3 + R_4^2.y_4}{4} + \left(y^{\alpha} - y^{\beta}\right)\left(c_0\left(R^{2\alpha} - R^{2\beta}\right)\right) \quad (11)$$

Where:

- $\hat{y}$ is final prediction value
- y_1, y_2, y_3, y_4 are predicted values of four ML models respectively.
- $R_1^2, R_2^2, R_3^2, R_4^2$ R square scores of four ML models respectively.
- y^{α} is predicted value of the model that has highest accuracy score
- y^{β} is predicted value of the model that has lowest accuracy score
- c_0 consolidation and improvement constant which is selected 1.5 for four ML models
- $R^{2\alpha}$–$R^{2\beta}$ difference between best and worst accuracy score

Thanks to this proposed hybrid method approach, the model with the most accuracy score contributes the most to the final prediction values, while the model with the worst score contributes the least. It also contributes to the final value in difference between best and worst ML models. In Fig. 5, unification strategy is demonstrated.

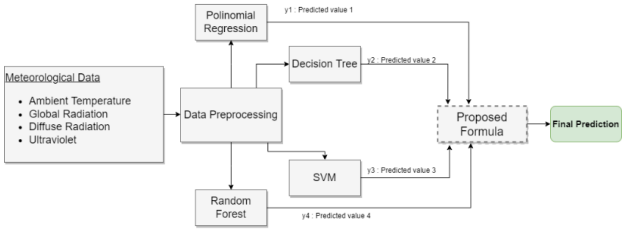

Fig. 5. Overall demonstration of prediction model that is proposed.

4. Performance Analysis

Predictive models using four different ML algorithms are trained using four different meteorological features and power generation amounts over the past five years. This section focuses on the performance outputs of these created models. In addition to the loss(error) functions described in the third section, accuracy is used as the performance metrics of the models. R squared value is used while calculating the accuracy. R squared, a performance output score, is calculated as:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (12)$$

The R^2 is calculated by dividing the sum of squares of residuals from the regression model (given by SS_{RES}) by the total sum of squares of errors from the average model (given by SS_{TOT}) and then subtracting it from 1. The result is between zero and one, and the closer it is to one, the better the model yielded. In addition, as a result of multiplying the R^2 value by 100, the accuracy percentage value is obtained. For the performance analysis of the trained models,accuracy as percentage, R^2, MSE, RMSE and MAE values of all models are acquired. In addition, the prediction results obtained with the proposed unification method and the actual results are evaluated with these metrics. Table 1 presents all these models and metrics together.

Table 1. Performance Output Metrics

	Accuracy(%)	**MSE**	**RMSE**	**MAE**	R^2
PolyReg	89.83%	10387.84	101.92	64.53	0.8983
SVM	81.92%	14893.92	133.04	72.45	0.8192
DT	87.47%	12992.07	113.98	42.45	0.8747
RF	93.01%	6933.98	83.27	35.82	0.9301
Hybrid	94.22%	5899.44	76.81	32.21	0.9422

In Tab. 1, the first column contains the model names. These are: PolyReg as polynomial regression, SVM is support vector regressor, DT as Decision Tree model, RF represents Random Forrest model and Hybrid name has been chosen as the proposed unification model. As it can be seen from the table, RF algorithm gave the best result, if we set the proposed joining method aside. The polynomial regression model gave the second and worst accuracy result as 81.12% with the SVM based prediction algorithm.

If the error results are to be evaluated, the RF algorithm gives the minimum deviation between the actual values and the predicted values, while this deviation value is maximum in the SVM algorithm. Moreover, in terms of MAE, the DT algorithm gave better MAE than the polynomial regressor, even though it had less accuracy score, which shows that the difference between the actual values and the predicted values is less. Tab.1 also indicates that the hybrid forecasting model built with the proposed unification approach gives the best results in terms of both accuracy and error outputs. This hybrid model showed that with the lowest RMSE value it had the least deviation in its predictions and with the lowest MAE value it had, the difference between the actual and its predictions is the minimum than others.

Fig. 6 presents a visual comparison of the hybrid prediction model created with the proposed method with other ML prediction models. Here, the predicted results of the proposed hybrid

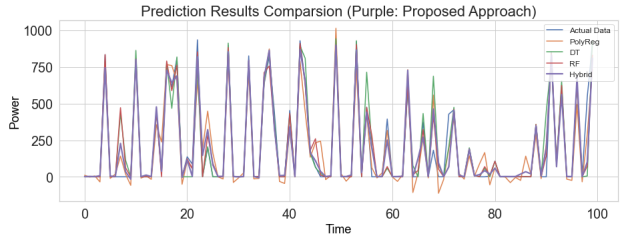

Fig. 6. Comparison of the proposed method with other ML models

model are shown in purple, while the actual values are shown in blue. Moreover, the prediction results of the polynomial regression in orange, the DT model in green, and the RF model in red are plotted in the Fig. 6. When the graph is examined, it can be seen that the polynomial regression model gives an overestimation result at the sharp ends. Furthermore, it can be observed that the DT model gives results much higher than the true value in some cases. Although the prediction model created with RF gave relatively satisfactory prediction results, the hybrid prediction model created with the proposed method produced almost the same results with the real values. In order to better examine

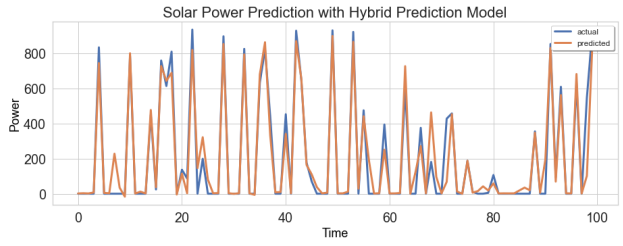

Fig. 7. Solar Power Prediction using Unified Hybrid Model

the performance of the proposed combined model, the actual data and the estimation results of this model are visualized together in Fig. 7. Here, estimation results are shown in orange versus actual values represented in blue. Except for a few sharp transition points, these two outcomes overlapped. In addition, this model preferred to give lower estimates at these sharp points rather than overestimate the actual power generation value. This feature can avoid making useless decisions in the management of the solar power plant.

5. Conclusion

The power generation amount of the solar power plant primarily depends on the production method of the solar panels and the purity of the material used. Although this changes the power generation efficiency, it does not dynamically change the power generation. A secondary but more important phenomenon is the fact that meteorological data constantly changes the instantaneous power generation. On the other hand, estimating how much power a solar power plant will produce in minutes and hours will significantly affect the decisions to be taken for this power plant. While contributing to the control of battery banks, if any, it will also enable more accurate decisions such as decommissioning or taking power plant sections. Considering these and similar advantages, this study focuses on solar power generation forecasting. In this context, four different machine learning-based prediction models are created and trained using selected weather data. Moreover, these forecasting models are combined using an innovative unification strategy and a hybrid prediction model is created. This model, which not only gives better accuracy than the ML prediction models used alone, but also offers better error outputs.Thanks to this method, it has been seen that the difference between the actual and predicted values is reduced, and the deviation in the estimations is lowered and improved.

6. References

[1] E. D. Obando, S. X. Carvajal, and J. Pineda Agudelo, "Solar radiation prediction using machine learning techniques: A review," *IEEE Latin America Transactions*, vol. 17, no. 04, pp. 684–697, 2019.

[2] M. Elsaraiti and A. Merabet, "Solar power forecasting using deep learning techniques," *IEEE Access*, vol. 10, pp. 31 692–31 698, 2022.

[3] H. S. Jang, K. Y. Bae, H.-S. Park, and D. K. Sung, "Solar power prediction based on satellite images and support vector machine," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1255–1263, 2016.

[4] J. Zheng, H. Zhang, Y. Dai, B. Wang, T. Zheng, Q. Liao, Y. Liang, F. Zhang, and X. Song, "Time series prediction for output of multi-region solar power plants," *Applied Energy*, vol. 257, p. 114001, 2020.

[5] F. Serttas, F. O. Hocaoglu, and E. Akarslan, "Short term solar power generation forecasting: A novel approach," in *2018 International Conference on Photovoltaic Science and Technologies (PVCon)*, 2018, pp. 1–4.

[6] R. H. M. Zargar and M. H. Yaghmaee Moghaddam, "Development of a markov-chain-based solar generation model for smart microgrid energy management system," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 2, pp. 736–745, 2020.

[7] N. Aksoy and I. Genc, "Predictive models development using gradient boosting based methods for solar power plants," *Journal of Computational Science*, vol. 67, p. 101958, 2023. [Online]. Available: https://doi.org/10.1016/j.jocs.2023.101958

[8] K. Li, R. Wang, H. Lei, T. Zhang, Y. Liu, and X. Zheng, "Interval prediction of solar power using an improved bootstrap method," *Solar Energy*, vol. 159, pp. 97–112, 2018.

[9] A. Asrari, T. X. Wu, and B. Ramos, "A hybrid algorithm for short-term solar power prediction—sunshine state case study," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 582–591, 2017.

[10] B. Zazoum, "Solar photovoltaic power prediction using different machine learning methods," *Energy Reports*, vol. 8, pp. 19–25, 2022, 2021 The 8th International Conference on Power and Energy Systems Engineering.

[11] M. Nejati and N. Amjady, "A new solar power prediction method based on feature clustering and hybrid-classification-regression forecasting," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 2, pp. 1188–1198, 2022.

[12] S. Wen, C. Zhang, H. Lan, Y. Xu, Y. Tang, and Y. Huang, "A hybrid ensemble model for interval prediction of solar power output in ship onboard power systems," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 14–24, 2021.