

Research Article

Comparison of The Performance of Machine Learning Methods for Solar Energy Power Prediction

Ozlem BEZEK GÜRE^{1*} 

^{1*} Batman University, Department of Medical Documentation, Secretariat Program Health Services Vocational School, Batman, Turkey.
(e-mail: ozlem.bezekgure@batman.edu.tr)

ARTICLE INFO

Received: May, 30 2025

Revised: Sep., 01, 2025

Accepted: Oct., 22, 2025

Keywords:

Machine Learning
Predictive Models
Number of Variables
Regression
Solar Energy
Power Generation

Corresponding author:

Özlem Bezek Güre

ISSN: 2536-5010 | e-ISSN: 2536-5134

DOI: <https://doi.org/10.36222/ejt.1709782>

ABSTRACT

In parallel with population growth, the demand for electrical energy continues to rise. As traditional energy sources have proved insufficient to meet this demand, attention has shifted to environmentally friendly renewable energy sources, among which solar energy stands out. Electricity generation from solar energy is achieved via photovoltaic (PV) systems. Accurately predicting generated power using machine learning (ML) methods with low error supports effective and efficient electricity planning. This study aims to predict solar energy power using ML methods—Gradient Boosting Regressor (GBR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR), and AdaBoost Regressor (AR). For this purpose, the “Solar energy power generation” dataset from Kaggle was utilized. The dataset includes meteorological variables such as temperature, humidity, and radiation, consisting of 21 variables and 4,212 measurements. The Relief feature selection method was used to identify the most informative independent variables for predicting solar energy power. Seventy percent of the data was used for training and the remainder for testing. Model performance was examined with respect to the number of variables. To compare methods, we employed MSE, RMSE, MAE, R^2 , and training time, and applied cross-validation to enhance model performance. The results indicate that AR achieved superior predictive performance, yielding lower errors (MSE = 0.111, RMSE = 0.333, MAE = 0.145) than the other methods. In contrast, RFR was found to be better in terms of both speed (training time = 2.896) and explained variance ($R^2=0.926$). Additionally, except for DTR, all evaluated techniques exhibited improved performance as the number of variables increased.

1. INTRODUCTION

Renewable energy sources are widely recognized as clean alternatives to conventional energy. They have increasingly supplanted traditional sources owing to their role in reducing greenhouse gas emissions and mitigating global warming. When optimally utilized, renewable resources also minimize environmental impacts such as waste generation [1]. Among these resources, solar energy holds a prominent place because of its potential to reduce carbon footprints and alleviate rising electricity costs [2-3]. Energy production from solar energy is carried out through photovoltaic systems.

The decreasing costs of PV technologies and the relative ease of installation have accelerated their adoption in recent years. PV systems typically comprise solar panels, inverters, charge controllers, solar cables, connection elements (e.g., MC4 connectors), and structural mounting components. Broadly, PV installations are categorized as on-grid and off-grid systems. On-grid (grid-tied) systems convert solar energy into grid-quality electricity in real time via an on-grid inverter. As they do not include storage, they are often described as battery-less systems and thus cannot be used effectively when

solar irradiance is insufficient. These systems are commonly deployed for self-consumption and for selling surplus electricity to the distribution utility. In addition to purely on-grid systems, grid-assisted and hybrid configurations are also used. A representative schematic is provided in Figure 1.

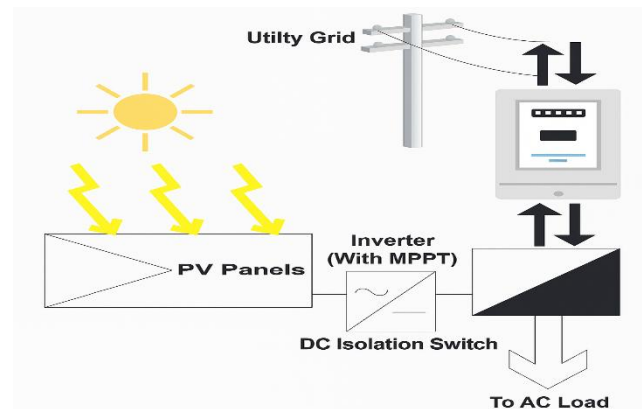


Figure 1. On-grid photovoltaic system design [4]

Off-grid systems, by contrast, are battery-based and are typically installed in locations without access to the power grid or where extending transmission lines is impractical. In such systems, energy harvested from solar panels is stored in batteries and subsequently used to meet demand during hours when panels cannot generate electricity (e.g., at night). A representative schematic is provided in Figure 2.

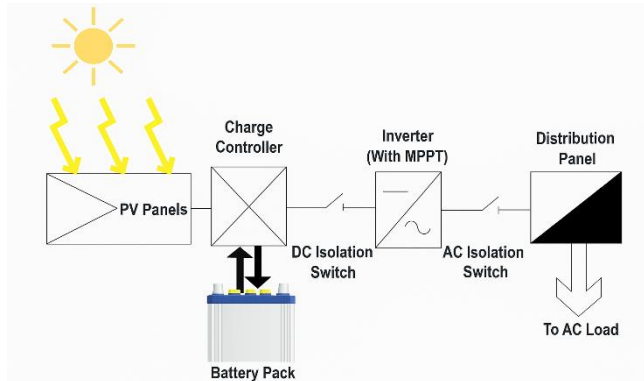


Figure 2. Off-grid photovoltaic system design [4].

The power delivered by PV systems is influenced by project design, climatic conditions, and panel characteristics. Accurate power prediction is therefore essential for cost-effective energy planning [4]. In power calculations, panel capacity is generally sized above the demand load to ensure efficient operation. Although initial capital costs can be high, payback periods may be relatively short. PV generation is particularly sensitive to solar radiation intensity, temperature, humidity, and wind speed. Photovoltaic energy production is affected by factors such as solar radiation intensity, temperature, humidity, and wind speed [5]. Because these variables fluctuate over time, uncertainty in energy production is unavoidable. Consequently, there is a pressing need to design systems capable of operating under varying conditions and to predict the amount of electricity they will produce [6-7]. Reliable predictions, in turn, support effective strategies for electricity production, distribution, and storage across the grid [8-9]. The literature accordingly documents a growing number of studies that employ machine learning (ML) to predict solar power.

Machine learning comprises a family of methods that learn from historical data to make predictions, uncovering complex relationships among variables that may be difficult to detect with traditional techniques [10]. The success of ML depends on both data characteristics and algorithmic performance [11]. As a field, ML integrates concepts and tools from artificial intelligence, information technology, and statistics [12].

This study aims to predict solar energy power using ML methods—specifically Gradient Boosting Regressor (GBR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR), and AdaBoost Regressor (AR). To our knowledge, few studies have jointly applied this particular set of methods to the problem of solar power prediction, and even fewer have examined how prediction performance changes with different numbers of predictor variables.

The contributions of this study to the literature include:

- It employs ML methods to predict solar energy production from weather-related variables.
- It identifies the most informative predictors of solar power using the Relief feature selection method.
- It examines model performance as a function of the number of variables included.
- It evaluates models using MSE, RMSE, MAE, R^2 , and training time to determine the best-performing approach.
- It offers a comparative assessment across regression-based ML methods.
- By developing a model that minimizes prediction error, it highlights the practical impact of ML for solar power forecasting.

The remainder of the paper is organized as follows. Section 2 reviews prior work on solar power prediction and the methods evaluated in this study. Section 3 describes the dataset and methodological procedures. Section 4 reports the analysis results, and Section 5 discusses the findings and offers recommendations.

2. RELATED WORKS

The literature reports extensive use of machine learning (ML) techniques to forecast photovoltaic (PV) power output. Rajulu et al. [13] aimed to predict power using Linear Regression (LR), Random Forest Regressor (RFR), Support Vector Regression (SVR), k-Nearest Neighbor Regression (kNNR), Extreme Gradient Boosting (XGBoost), Decision Tree Regressor (DTR), Lasso Regression (LassoR), and hybrid ML approaches. Sandeep et al. [14] employed XGBoost, DTR, RFR, SVR, LR, k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and hybrid ensemble methods for the same purpose. Similarly, Abdullah et al. [10] utilized LR, SVR, kNNR, DTR, RFR, Multilayer Perceptron (MLP), and Gradient Boosting Regressor (GBR).

Erten and Aydilek [15] investigated solar power prediction with LR, Ridge Regression (RidgeR), LassoR, and Elastic Net Regression (ElasticR). In a related study, Babu et al. [16] applied Multiple Linear Regression (MLR), RidgeR, LassoR, DTR, SVR, RFR, AdaBoost, Bagging, and GBR. Büyükbıçakcı [17] explored genetic algorithms (GA), Support Vector Machines (SVM), and GA-SVM. Nur et al. [4] employed Artificial Neural Networks (ANNs). Mohanty et al. [18] used Bayesian Linear Regression and Light Gradient Boosting Machine (LightGBM) to model PV power. Pachauri et al. [19] examined solar energy production using GA, while Chang and Jia [20] proposed a hybrid system. Moreover, Mohanty, Pati, and Mahapatra [21] investigated deep learning architectures—Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM)—for solar power prediction.

These prior studies demonstrate a wide spectrum of ML and deep learning approaches for PV power forecasting. However, relatively few have jointly evaluated the specific combination of GBR, DTR, RFR, and AdaBoost Regressor (AR), and even fewer have systematically analyzed how performance varies with the number of predictor variables—an empirical gap the present study addresses.

3. MATERIAL AND METHOD

This study employed the “Solar Energy Power Generation” dataset from Kaggle to predict solar energy power [22]. The dataset comprises 21 weather-related variables and 4,212 observations. Descriptive statistics for all variables used in the analysis are presented in Table 1. In subsequent analyses, the Relief feature selection method was used to rank predictor importance (see Results, Table 2).

TABLE 1
DESCRIPTIVE STATISTICS OF THE VARIABLES

Variables	Unit	Min	Max	Mean	Std. Deviation
Temperature 2m above ground	°C	-5.35	34.90	15.07	8.85
Relative humidity 2m above	%	7	100	51.36	23.53
Mean sea level pressure	%	998	1047	1019.34	7.02
Total precipitation	%	0	3	0.03	0.17
Snowfall amount	%	0	2	0.00	0.04
Total cloud cover	%	0	100	34.06	42.84
High cloud cover	%	0	100	14.46	30.71
Medium cloud cover	%	0	100	20.02	36.39
Low cloud cover	%	0	100	21.37	38.01
Shortwave radiation backwards	$\frac{W}{m^2}$	0	952	387.76	278.46
Wind speed at 10 m above ground	$\frac{m}{s}$	0.00	61.18	16.23	9.88
Wind direction at 10 m above ground	$\frac{m}{s}$	0.54	360.00	195.08	106.63
Wind speed at 80 m above ground	$\frac{m}{s}$	0.00	66.88	18.98	12.00
Wind direction at 80 m above ground	$\frac{m}{s}$	1.12	360.00	191.17	108.76
Wind speed	$\frac{m}{s}$	0.00	61.11	16.36	9.89
Wind direction	$\frac{m}{s}$	1.12	360.00	192.45	106.52
Wind gust at 10 m above ground	$\frac{m}{s}$	0.72	84.96	20.58	12.65
Angle_of_incidence	Degree	3.75	121.63	50.84	26.64
Zenith	Degree	17.72	128.41	59.98	19.86
Azimuth	Degree	54.37	289.04	169.17	64.57
Generated_power_	KW	0.0006	3056.79	1134.35	937.96

3.1. Gradient Boosting Regressor

The Gradient Boosting Regressor is an ensemble technique introduced by Friedman that combines multiple weak learners—typically regression trees—via stage-wise additive modeling and gradient-based optimization [23]. GBR iteratively fits trees to the residuals of the current model to minimize a specified loss (here, mean squared error) [24-25]. Model accuracy is improved by appropriately selecting the number of iterations (trees) and related hyperparameters [26]. Concretely, a first tree is fit to the data; residuals (actual minus predicted) are computed and used to fit the next tree, and so on, thereby emphasizing harder-to-predict observations and reducing overall error through gradient optimization [23, 27].

3.2. Decision Tree Regressor

Decision trees provide supervised models for both classification and prediction and are used to derive transparent decision structures [28-30]. For regression, the tree is built by

iteratively applying **binary splits** that minimize impurity—commonly the sum of squared deviations from the mean—until stopping criteria are met [31-33]. Internal (decision) nodes specify feature thresholds; leaf nodes store numerical targets. The root node corresponds to the split offering the largest error reduction [32]. Tree depth and minimum node size are typically constrained to avoid overfitting.

3.3. Random Forest Regressor

Random Forest (RF) is an ensemble of decision trees constructed on bootstrapped samples and randomized feature subsets [34-36]. It integrates ideas from bootstrap aggregation and randomized splits to reduce correlation among trees, thereby lowering variance and improving generalization [37-40]. In regression, the Random Forest Regressor (RFR) builds many trees in parallel and averages their predictions. As an extension of CART-style trees, RFR often achieves higher accuracy than a single tree by averaging many noisy but unbiased learners [26, 38-39].

3.4. Adaboost Regressor

Boosting algorithms aim to improve weak learners by iteratively reweighting training instances so that subsequent learners focus more on previously misfit cases [41]. Among boosting methods, AdaBoost (Adaptive Boosting)—developed by Freund and Schapire (1997)—is widely used and theoretically well-founded [42-43]. AdaBoost can address both classification (e.g., AdaBoost.M1/M2) and regression tasks (AdaBoost.R, hereafter AR) [44]. AR typically employs shallow trees as base learners; at each iteration, instance weights are updated according to residual errors, and the final prediction is a weighted aggregation of base learners [47-50]. AR is fast and simple to implement and can reduce both bias and variance, though it may be sensitive to outliers and noisy data; if base learners are weak in performance, the aggregate model may also underperform [38, 44].

Performance Metrics

The following metrics were used to evaluate model performance: MSE, RMSE, MAE, and R^2 .

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

(1)

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

(2)

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

(3)

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

(4)

[51,52].

4. RESULTS

Before conducting the main analyses, multicollinearity diagnostics were performed. Multicollinearity is typically indicated when VIF > 10 and tolerance < 0.1 [53]. In the present data, VIFs ranged from 1.752 to 4.642 and tolerance values from 0.215 to 0.571, which indicates no multicollinearity concerns. After confirming that the dataset contained no missing values, a z-transformation was applied. The dependent variable was defined as solar energy power.

Next, the Relief feature selection method was used to identify the most informative predictors (Table 2). Relief

employs a nearest-neighbor search to capture feature interactions in high-dimensional data and was preferred for its ability to detect complex patterns, handle mixed data types, and offer computational efficiency [54]. To examine how predictor set size affects performance, models were trained with the top 5 variables, top 10 variables, and all variables. To support generalizability, 10-fold cross-validation was employed, and models were evaluated using the software's default hyperparameters.

TABLE 2

VARIABLES CONSIDERED IMPORTANT ACCORDING TO RELIEF FEATURE SELECTION METHODS

	Variables	Relief (Score)
1	Angle of incidence	0.118
2	Azimuth	0.113
3	Shortwave radiation backwards	0.097
4	Relative humidity 2m above grid	0.091
5	Zenith	0.086
6	Temperature 2m above grid	0.079
7	Mean sea level pressure	0.074
8	High cloud cover high	0.071
9	Wind speed 80m above grid	0.055
10	Wind direction 80 m above grid	0.055
11	Wind gust 10m above grid	0.053
12	Wind speed 900 mb	0.051
13	Wind speed 10 m above grid	0.050
14	Wind direction 900 mb	0.044
15	Wind direction 10m above grid	0.044
16	Total cloud cover low	0.031
17	Low cloud cover low	0.026
18	Medium cloud cover mid	0.021
19	Total precipitation	0.006
20	Snowfall amount	0.000

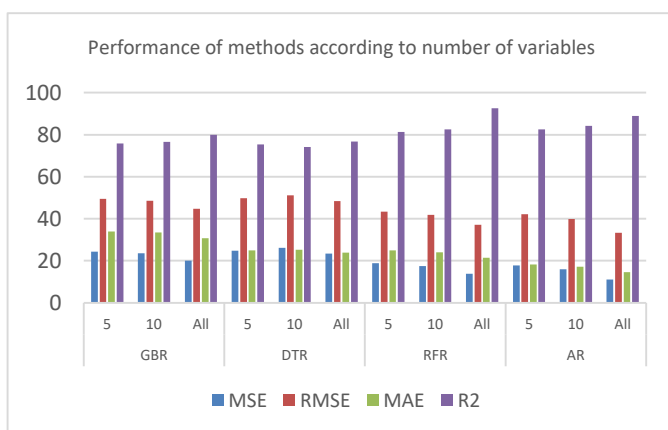


Figure 3. Performance of the methods with respect to the number of variables

As seen in Table 2, according to the Relief feature selection method, the most important factor affecting solar energy power is the angle of incidence. The next most important variable is azimuth. A similar result was also

obtained in the study conducted by Babu. The least important variables are snowfall amount and total precipitation.

Before proceeding with the analyses, 70% of the data was designated as training data, and 30% as test data. On the other hand, in order to determine the relationship between the dependent variable—solar energy power—and the other variables used in the study, Spearman's correlation coefficient was employed. According to the Spearman correlation analysis, the variables with the highest correlation with solar energy power are, respectively, zenith, angle of incidence, and shortwave radiation. The correlation coefficient between solar energy power and zenith was -0.68 , between solar energy power and angle of incidence -0.67 , and between solar energy power and shortwave radiation 0.55 . Lower correlations were found between solar energy power and the other variables. In addition, a positive relationship was identified with shortwave radiation and temperature, while negative relationships were found with the other variables.

TABLE 3.

PERFORMANCE METRICS FOR THE METHODS

Method	Variables number	MSE	RMSE	MAE	R ²	Training time
GBR	5	0.244	0.494	0.340	0.758	8.276
	10	0.236	0.486	0.334	0.766	18.760
	All	0.201	0.448	0.307	0.800	26.028
DTR	5	0.248	0.498	0.249	0.754	10.928
	10	0.262	0.511	0.252	0.741	19.527
	All	0.234	0.484	0.238	0.768	38.264
RFR	5	0.188	0.434	0.249	0.813	2.896
	10	0.175	0.418	0.240	0.826	5.625
	All	0.138	0.371	0.214	0.926	8.913
AR	5	0.177	0.421	0.183	0.825	11.084
	10	0.159	0.399	0.172	0.842	20.524
	All	0.111	0.333	0.145	0.890	33.346

Table 3 illustrates that, when all variables were considered, the AR method achieved superior predictive performance—evidenced by lower errors—compared to the other approaches. For the AR method, the MSE, RMSE, and MAE values were determined to be 0.111, 0.333, and 0.145, respectively. Conversely, with respect to the R² metric, the RFR method was observed to outperform the alternatives. The R² value indicates the proportion of variance in the dependent variable explained by the independent variables. According to the RFR method, solar power generation is explained by the independent variables at a rate of 92.6%. Furthermore, in terms of computational speed, the RFR method demonstrated superior efficiency when five variables were employed. The analysis also revealed that, with the exception of the DT method, all evaluated techniques exhibited enhanced performance as the number of variables increased. Figure 3 illustrates the performance of the methods with respect to the number of variables.

The literature shows that there are studies supporting our findings [55–61]. In contrast to our results, some studies have also demonstrated that the use of feature selection methods has a positive effect on the performance of the methods [63–64]. Furthermore, the literature indicates that there are studies

suggesting that the number of variables does not have a significant impact on the performance of the methods [41, 64–68].

In the present study, it was observed that the GBR, RFR, and AR methods, belonging to the family of ensemble learning approaches, demonstrated superior predictive performance compared to the DTR method. As is well known, ensemble methods are algorithms that combine multiple decision trees to generate stronger regressors. It has been noted that these methods provide more robust predictions compared to individual trees [69–72].

5. CONCLUSION AND SUGGESTIONS

The present study utilized machine learning methods, specifically GBR, DTR, RFR, and AR, to predict solar energy power. Although a small body of work has combined the methods adopted in this study, the literature lacks systematic evaluations of their comparative performance as a function of variable-set size. Table 3 compares the results of this study with those of previous studies conducted to predict solar energy power.

Table 4 illustrates that the present study's results exhibit superior performance—evidenced by lower error rates—compared to those of other studies,

Recently, due to the increasing world population and the parallel development of technology, the global demand for energy is rising [73]. Among the efforts to meet this demand, there is growing interest in environmentally friendly renewable energy sources, Solar energy, being a reliable, cheap, and clean energy source, is the most preferred among these sources [74]. Accurately predicting the power generated by PV systems through machine learning methods makes it possible to optimally plan energy production, which is a key indicator of economic growth. In this way, strategies can be developed based on the prediction results.

Therefore, using machine learning methods to make more robust predictions of solar energy power is considered important. Based on prediction results, it is crucial for countries to develop strategies related to energy management. In this study, we evaluated the performance of the ensemble methods GBR, RFR, and AR, alongside the single-model baseline DT, for forecasting solar power. The findings indicate that the ensemble approaches outperform DTR. We also examined performance as a function of the number of predictors. In machine-learning models, the predictor count directly affects model complexity, generalization capacity, and computational efficiency: too few variables simplify the model and may limit its ability to learn complex relationships, leading to underfitting; too many variables can induce overfitting to the training data and degrade test performance. The appropriate number of variables depends not only on data volume but also on model class and application context. For these reasons, careful variable selection is essential for improving model performance [75].

Future research could further analyze how performance varies across different predictor set sizes. Beyond the methods investigated here, alternative algorithms such as LSTM, GRU, Transformer architectures, Bagging, MARS, and artificial neural networks could be considered. Researchers might also study the effects of varying sample sizes, cross-validation schemes, and hyperparameter settings. In the present work, feature selection was performed with the Relief algorithm;

subsequent studies could compare alternative feature-selection and dimensionality-reduction techniques. Finally, to enhance external validity, models should be tested on datasets drawn from diverse geographic regions and climatic conditions.

TABLE 4.

COMPARISON OF THE CURRENT STUDY WITH STUDIES IN THE LITERATURE

Study	Method	MSE	RMSE	MAE	R ²	Training time
The Current Study (All variables)	GBR	0.201	0.448	0.307	0.800	26.028
	DTR	0.234	0.484	0.238	0.768	38.264
	RFR	0.138	0.371	0.214	0.926	8.913
	AR	0.111	0.333	0.145	0.890	33.346
Nur et al. [4]	ANN	0.59	0.77	0.42	0.84	-
Rajulu et al. [13]	Linear R	0.28	0.53	0.35	0.71	-
	RFR	0.23	0.48	0.28	0.77	-
	SVM	0.28	0.53	0.35	0.72	-
	DTR	0.31	0.56	0.28	0.69	-
	KNN	0.24	0.49	0.30	0.76	-
	XGBoost	0.18	0.42	0.23	0.82	-
	LassoR	0.28	0.530	0.35	0.72	-
	Hybrid ML	0.17	0.410	0.23	0.83	-
Sandeep et al. [14]	XGBoost	0.27	0.51	0.33	0.74	-
	DTR	0.42	0.65	0.37	0.60	-
	RFR	0.22	0.47	0.30	0.77	-
	ANN	0.24	0.49	0.32	0.76	-
	SVR	0.23	0.48	0.29	0.76	-
	LR	0.30	0.55	0.42	0.69	-
	KNN	0.23	0.48	0.31	0.75	-
	Hybrid ensemble	0.21	0.21	0.29	0.79	-
Babu et al. [16]	MLR	352465.70	559.60	412.27	0.60	-
	RR	252465.60	502.46	386.69	0.73	-
	LassoR	252465.6	502.46	386.69	0.73	-
	GBR	158559.33	399.44	253.62	0.83	-
	DTR	222190.13	471.37	286.03	0.75	-
	SVR	557635.43	746.75	616.352	0.39	-
	RFR	164088.80	405.07	258.06	0.82	-
	AR	284581.40	533.46	427.64	0.69	-
	BR	178778.30	422.82	265.82	0.81	-
Büyükbıça kci et al. [17]	GA	-	0.07	0.13	-	-
	SVM	-	0.04	0.09	-	-
	GA-SVM	-	0.02	0.03	-	-

REFERENCES

- [1] Panwar, N, L., Kaushik, S, C., & Kothari, S, (2011), Role of renewable energy sources in environmental protection: A review, *Renewable and sustainable energy reviews*, 15(3), 1513-1524,

- [2] Tasnin, W., Saikia, L. C., & Raju, M. (2018), Deregulated AGC of multi-area system incorporating dish-Stirling solar thermal and geothermal power plants using fractional order cascade controller, *International Journal of Electrical Power & Energy Systems*, 101, 60-74.
- [3] Borunda, M., Ramírez, A., Garduno, R., Ruíz, G., Hernandez, S., & Jaramillo, O. A. (2022), Photovoltaic power generation forecasting for regional assessment using machine learning, *Energies*, 15(23), 8895.
- [4] Nur, A., Güre, B., Rüstemli, S. & Bezek Güre, Ö. (2024) Solar Power Estimation by Using Artificial Neural Networks, he International Conference on Energy and Environmental Technologies in Engineering and Architecture (ICETEA 2024,
- [5] Deng, X., Da, F., Shao, H., & Wang, X. (2023), A Survey of the Researches on Grid-Connected Solar Power Generation Systems and Power Forecasting Methods Based on Ground-Based Cloud Atlas, *Energy Engineering*, 120(2), 385-408.
- [6] Shahid, F., Zameer, A., Afzal, M., & Hassan, M. (2020), Short term solar energy prediction by machine learning algorithms, *arXiv preprint arXiv:2012.00688*,
- [7] Nourani, V., Elkiran, G., Abdullahi, J., & Tahsin, A. (2019), Multi-region modeling of daily global solar radiation with artificial intelligence ensemble, *Natural Resources Research*, 28, 1217-1238.
- [8] Ates, K. T. (2022), Çok Katmanlı Yapay Sinir Ağı Modeli ve Kültürel Algoritma Modeli Kullanılarak Geliştirilen Melez Yöntem ile Kısa Vadeli Fotovoltaik Enerji Santrali Çıkış Gücü Tahmini, *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 5(1), 342-354.
- [9] Yang, H.-T.; Huang, C.-M.; Huang, Y.-C.; Pai, Y.-S. A weather-based hybrid method for 1-day ahead hourly forecasting of solar power output, *IEEE Trans, Sustain, Energy* **2014**, 5, 917–926.
- [10] Abdullah, B. U. D., Khanday, S. A., Islam, N. U., Lata, S., Fatima, H., & Nengroo, S. H. (2024), Comparative Analysis Using Multiple Regression Models for Forecasting Photovoltaic Power Generation, *Energies*, 17(7), 1564.
- [11] Sarker, I. H. (2021), Machine learning: Algorithms, real-world applications and research directions, *SN computer science*, 2(3), 160.
- [12] Nasteski, V. (2017), An overview of the supervised machine learning methods, *Horizons*, b, 4(51-62), 56.
- [13] Rajulu, A. G., Mohanty, P., Pati, U. C., & Mahapatra, K. (2024, November), SolarPred: Hybrid Machine Learning Based Time Series Prediction Model for Solar Energy Applications, In 2024 IEEE 4th International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC) (pp. 1-6), IEEE.
- [14] Sandeep, C. S., Mohanty, P., & Pati, U. C. (2023, August), Machine learning based framework for prediction of photovoltaic output power, In 2023 IEEE 3rd International Conference on Sustainable Energy and Future Electric Transportation (SEFET) (pp. 1-6), IEEE.
- [15] Erten, M. Y., & Aydılek, H. (2022), Solar power prediction using regression models, *International Journal of Engineering Research and Development*, 14(3), 333-342.
- [16] Babu, A. R. V., Kumar, N. B., Narasipuram, R. P., Periyannan, S., Hosseinpour, A., & Flah, A. (2025). Solar Energy Forecasting using Machine Learning Techniques for Enhanced Grid Stability. IEEE Access.
- [17] Büyükbıçakci, E., Atli, C., Dumanli, M., & Bulat, S. (2024, December), Enhanced Photovoltaic Power Forecasting in Renewable Energy Systems Using Genetic Algorithm and SVM Approaches, In 2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNC) (pp. 1-6), IEEE.
- [18] Mohanty, P., Subhadarshini, K., Nayak, R., Pati, U. C., & Mahapatra, K. (2025), Exploring data-driven multivariate statistical models for the prediction of solar energy, In *Computer Vision and Machine Intelligence for Renewable Energy Systems* (pp. 85-101), Elsevier.
- [19] Pachauri, R. K., Yadav, J., Torto, S. O. G., Minai, A. F., Pandey, V., Shashikant, & Sharma, P. (2024), Solar Photovoltaic System-Based Power Generation: Imperative Role of Artificial Intelligence and Machine Learning, *Clean and Renewable Energy Production*, 267-286.
- [20] Chang, J., & Jia, S. Y. (2009, July), Modeling and application of wind-solar energy hybrid power generation system based on multi-agent technology, In *2009 International Conference on Machine Learning and Cybernetics* (Vol. 3, pp. 1754-1758), IEEE.
- [21] Mohanty, P., Pati, U. C., & Mahapatra, K. (2023, October), Deep Learning Based Framework for Forecasting Solar Panel Output Power, In *IFIP International Internet of Things Conference* (pp. 229-239), Cham: Springer Nature Switzerland.
- [22] <https://www.kaggle.com/datasets/stucom/solar-energy-power-generation-dataset>
- [23] Huang, Y., Liu, Y., Li, C., & Wang, C. (2019), GBRTVis: online analysis of gradient boosting regression tree, *Journal of Visualization*, 22, 125-140.
- [24] Gumaï, A., Al-Rakhani, M., Al Rahhal, M. M., Albogamy, F. R., Al Maghayreh, E., & AlSalman, H. (2021), Prediction of COVID-19 confirmed cases using gradient boosting regression method, *Comput Mater Continua*, 66, 315-329.
- [25] Touzani, S., Granderson, J., & Fernandes, S. (2018), Gradient boosting machine for modeling the energy consumption of commercial buildings, *Energy and Buildings*, 158, 1533-1543.
- [26] Zhang, Y., & Haghani, A. (2015), A gradient boosting method to improve travel time prediction, *Transportation Research Part C: Emerging Technologies*, 58, 308-324.
- [27] Islam, S., & Amin, S. H. (2020), Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques, *Journal of Big Data*, 7(1), 65.
- [28] Pathak, S., Mishra, I., & Swetapadma, A. (2018, November), An assessment of decision tree based classification and regression algorithms, In *2018 3rd International Conference on Inventive Computation Technologies (ICICT)* (pp. 92-95), IEEE.
- [29] Chang, C. L., & Chen, C. H. (2009), Applying decision tree and neural network to increase quality of dermatologic diagnosis, *Expert Systems with Applications*, 36(2), 4035-4041.
- [30] Chen, M. Y. (2011), Predicting corporate financial distress based on integration of decision tree classification and logistic regression, *Expert systems with applications*, 38(9), 11261-11272.
- [31] Pekel, E. (2020), Estimation of soil moisture using decision tree regression, *Theoretical and Applied Climatology*, 139(3-4), 1111-1119.
- [32] Patil, S., Patil, A., & Phalle, V. M. (2018, December), Life prediction of bearing by using adaboost regressor, In *Proceedings of TRIBOINDIA-2018 An International Conference on Tribology*,
- [33] Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005), Decision tree regression for soft classification of remote sensing data, *Remote Sensing of Environment*, 97(3), 322-336.
- [34] Wang, L., Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016), Estimation of biomass in wheat using random forest regression algorithm and remote sensing data, *The Crop Journal*, 4, 212-219.
- [35] Biau, G., Cadre, B., & Rouvière, L. (2019), Accelerated gradient boosting, *Machine learning*, 108, 971-992.
- [36] Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020), Analysis of Factors Effecting PISA 2015 Mathematics Literacy via Educational Data Mining, *Education & Science/Eğitim ve Bilim*, 45(202),
- [37] Güre, Ö. B., Şevgin, H., & Kayri, M. (2023), Reviewing the Factors Affecting PISA Reading Skills by Using Random Forest and MARS Methods, *International Journal of Contemporary Educational Research*, 10(1), 181-196.
- [38] Shanmugasundar, G., Vanitha, M., Çep, R., Kumar, V., Kalita, K., & Ramachandran, M. (2021), A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining, *Processes*, 9(11), 2015.
- [39] Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C. W., Van den Bossche, P., ... & Omar, N. (2018), Random forest regression for online capacity estimation of lithium-ion batteries, *Applied energy*, 232, 197-210.
- [40] Geneur, R., Poggi, J. M., Tuleao Malot, C. ve Villa-Vialaneix, N. (2017), Random forest for big data, *Big Data Research*
- [41] Bezek Güre, Ö. (2023), Investigation of ensemble methods in terms of statistics: TIMMS 2019 example, *Neural Computing and Applications*, 35(32), 23507-23520
- [42] Ferreira, A. J., & Figueiredo, M. A. (2012), Boosting algorithms: A review of methods, theory.
- [43] Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013), Advance and prospects of AdaBoost algorithm, *Acta Automatica Sinica*, 39(6), 745-758.
- [44] Shrestha, D. L., & Solomatine, D. P. (2006), Experiments with AdaBoost, RT, an improved boosting scheme for regression, *Neural computation*, 18(7), 1678-1710.
- [45] Gupta, K. K., Kalita, K., Ghadai, R. K., Ramachandran, M., & Gao, X. Z. (2021), Machine learning-based predictive modelling of biodiesel production—A comparative perspective, *Energies*, 14(4), 1122.
- [46] Kummer, N., & Najjaran, H. (2014), Adaboost, MRT: Boosting regression for multivariate estimation, *Artif, Intell, Res.*, 3(4), 64-76.

- [47] Wang, F., Li, Z., He, F., Wang, R., Yu, W., & Nie, F. (2019), Feature learning viewpoint of AdaBoost and a new algorithm, *IEEE Access*, 7, 149890-149899,
- [48] Rokach L (2010) Ensemble-based classifiers, *Artif Intell Rev* 33(1):1-39
- [49] Sagi O, Rokach L (2018) Ensemble learning: a survey, *Wiley Interdiscip Rev Data Min Knowl Discov* 8(4):e1249
- [50] B. Kégl. Introduction to AdaBoost. Accessed: Nov. 15, 2009. [Online]. Available: <https://users.lal.in2p3.fr/kegl/teaching/stages/notes/tutorial>
- [51] Kayri, M., (2015), An intelligent approach to educational data: Performance comparison of the multilayer perceptron and the radial basis function artificial neural networks, *Educational Sciences: Theory & Practice*, 15(5),
- [52] Kayri, M., Kayri, I., & Gençoğlu, M, T, (2017), The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data, In 2017 14th International Conference on Engineering of Modern Electric Systems (EMES), 1-4,
- [53] Keller, P, S., El-Sheikh, M., Granger, D, A., & Buckhalt, J, A, (2012), Interactions between salivary cortisol and alphaamylase as predictors of children's cognitive functioning and academic performance, *Physiology & Behavior*, 105, 987-995,
- [54] Le, T. T., Urbanowicz, R. J., Moore, J. H., & McKinney, B. A. (2019). Statistical inference relief (STIR) feature selection. *Bioinformatics*, 35(8), 1358-1365.
- [55] Sordo M, Zeng Q (2005) On sample size and classification accuracy: a performance comparison. In: International symposium on biological and medical data analysis. Springer, Berlin, pp 193-201
- [56] Temel GO, Erdoğan S, Ankaralı H (2012) Usage of resampling methods for evaluating the performance of classification model. *J Inf Technol* 5(3):1-8
- [57] Dolgun MO (2014) Comparison of the performances of data mining classification methods based on prevalence of the dependent variable, sample size and the correlation of the independent variables. PhD thesis, Hacettepe University
- [58] Kanik EA, Temel GO, Erdoğan S, Kaya I'E (2013) Affected states soft independent modeling by class analogy from the relation between independent variables, number of independent variables and sample size. *Balkan Med J* 2013(1):28-32
- [59] Yabancı A (2017) Comparison of tree-based methods used in survival data. PhD thesis, Uludağ University
- [60] Nuray SE, Genc dal HB, Arama ZA (2021) Zeminlerin kıvam ve kompaksiyon özelliklerinin tahmininde rastgele orman regresyonu yönteminin uygulanabilirliği. *Mühendislik Bilimleri ve Tasarım Dergisi* 9(1):265-281
- [61] Kumar S, Chong I (2018) Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states. *Int J Environ Res Public Health* 15(12):2907
- [62] Latha CBC, Jeeva SC (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inf Med Unlock* 16:100203
- [63] Sun Y, Li Z, Li X, Zhang J (2021) Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction. *Appl Artif Intell* 35(4):290-303
- [64] Ceyhan G (2020) Comparison of performance of data mining methods used for classification in terms of data characteristics. PhD thesis, Gazi University
- [65] Kwon O, Sim JM (2013) Effects of data set features on the performances of classification algorithms. *Expert Syst Appl* 40(5):1847-1857
- [66] Dag H, Sayın KE, Yenidogan I, Albayrak S, Acar C (2012) Comparison of feature selection algorithms for medical data. In: 2012 International symposium on innovations in intelligent systems and applications. IEEE, pp 1-5
- [67] Huang S, Fang N (2013) Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. *Comput Educ* 61:133-145
- [68] Kasap Y, Dog an N, Koc ak C (2021) Determining variables that predict reading comprehension success by data mining in PISA 2018. *Manisa Celal Bayar Univ J Soc Sci* 19(4):241-258
- [69] Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging boosting and variants. *Mach Learn* 36(1):105-139
- [70] Rokach L (2010) Ensemble-based classifiers. *Artif Intell Rev* 33(1):1-39
- [71] Dimitriadou E, Weingessel A, Hornik K (2003) A cluster ensembles framework. Design and application of hybrid intelligent systems. IOS Press, Amsterdam
- [72] Sidhu, R. K., Kumar, R., & Rana, P. S. (2020). Machine learning based crop water demand forecasting using minimum climatological data. *Multimedia Tools and Applications*, 79(19), 13109-13124.
- [73] Geçmez, A., & Gençer, C. (2020), Güneş Enerji Santrali Üretim Verilerinin Meteorolojik Verilere Bağlı olarak Yapay Zeka Yöntemleri ile Tahmini, Doç, Dr, Aydın Ruşen Doç, Dr, Sadık Alper Yıldız, 39,
- [74] Uğuz, S., Oral, O., & Çağlayan, N, (2019), PV Güç Santrallerinden Elde Edilecek Enerjinin Makine Öğrenmesi Metotları Kullanılarak Tahmin Edilmesi, *International Journal of Engineering Research and Development*, 11(3), 769-779,
- [75] Guvon, I., & Elisseeff, A. (2003). *An introduction to variable and feature selection*. *Journal of Machine Learning Research*, 3, 1157-1182.

BIOGRAPHIES

Özlem Bezek Güre graduated in Statistics from Anadolu University. She received her MSc in Biostatistics from Fırat University, and her PhD from Van Yüzüncü Yıl University in Statistics. She is currently an instructor at Batman University, working on machine learning, data mining, and statistic. She has studies published in national and international scientific platforms.