Journal of the Technical University of Gabrovo

https://mc04.manuscriptcentral.com/jtug

EVALUATION OF THE GRADIENT BOOSTING ALGORITHM BASED ON TRAIN/TEST RATIOS IN SOLAR ENERGY POWER GENERATION FORECASTING

Dinçer Akal[1*], Taşkın Tez[2], İlhan Umut[3]

[1]*Department of Mechanical Engineering, Faculty of Engineering, Trakya University, 22030, Edirne, Turkey*
[2]*Edirne Vocational College of Technical Sciences, Trakya University, 22030, Edirne, Turkey*
[3]*Department of Electronics and Automation, Corlu Vocational School, Tekirdag Namik Kemal University, Tekirdağ, Turkiye*

ARTICLE INFO

Article history:
Received 23 September 2025
Revised 23 October 2025
Accepted 27 October 2025

Keywords:
solar energy, gradient boosting, machine learning, renewable energy prediction

http://doi.org/10.62853/HHZK1143

ABSTRACT

Accurate forecasting of solar energy generation is of critical importance for energy planning, resource management, and sustainability efforts. This study investigates the performance of the Gradient Boosting algorithm in predicting solar power output. The analysis utilizes the Solar Energy Power Generation Dataset obtained from the Kaggle platform. The dataset comprises hourly meteorological variables such as temperature, humidity, pressure, precipitation, various cloud cover types, shortwave radiation, wind speed and direction, solar angles, as well as the corresponding power generation values. During the preprocessing phase, the data were imported into the Orange open-source data analysis software, where variable names were standardized and transformed into a format suitable for modeling. Gradient Boosting was selected as the predictive algorithm, and its performance was evaluated under various train/test split ratios (50%, 60%, 66.6%, 70%, 75%, 80%, 90%, and 95%). Several essential performance metrics including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were employed to assess the model's performance. The highest R^2 value (0.790) and the lowest error rates were achieved with a 90% training ratio (RMSE=428.959, MAE=289.195). However, a slight performance decline observed at the 95% training ratio suggests a potential risk of overfitting. Overall, the findings demonstrate that Gradient Boosting is a reliable and effective method for forecasting solar energy generation, with optimal results obtained at the 90% training level. Future studies may achieve higher accuracy and generalization capacity through the integration of alternative boosting algorithms and hyperparameter optimization techniques.

1. INTRODUCTION

In today's world, societies are facing significant challenges due to climate change, which is largely driven by the increasing concentration of greenhouse gases in the atmosphere. In response, numerous initiatives have been launched with the aim of reducing anthropogenic CO_2 emissions. Simultaneously, the urgent need for an environmentally friendly transition has led to intensified research and development efforts focused on sustainable and eco-friendly technologies [1]. Investments in these areas continue to grow, driven by the goal of developing more cost-effective and higher-efficiency technologies and strategies. These efforts aim to facilitate the transition from conventional energy systems based on fossil fuels to structures reliant on renewable energy sources. In this context, machine learning and artificial intelligence-based approaches are increasingly being employed for solar energy forecasting, owing to their ability to model complex and non-linear relationships. The literature indicates that machine learning methods such as SVR, Random Forest, ANN, and boosting algorithms have demonstrated strong performance in energy forecasting tasks. For GHI and DNI prediction, ANN models utilizing weather forecast data as inputs have been widely applied. Input feature selection was commonly performed using Genetic Algorithms (GA) and the Gamma Test, resulting in significant performance gains over baseline models [2].

In another study, short-term PV power forecasting (1- and 2-hour ahead) was conducted using only intrinsic variables, employing ARIMA, kNN, ANN, and GA-optimized ANN (ANN/GA) models. ANN-based approaches generally outperformed others across multiple error metrics. However, ARIMA yielded lower Mean Bias Error (MBE) during certain intervals. The ANN/GA consistently outperformed the standard ANN, highlighting the benefits of simultaneous optimization of model parameters and input features. Notably, the ANN/GA model achieved a 32.2% reduction in RMSE compared to the persistence model in one-hour ahead forecasts [3]. In recent years, a wide array of methods and algorithms have been developed to improve the accuracy of energy production forecasting. The primary goal of these

* Corresponding author. E-mail: dincerakal@trakya.edu.tr

approaches is to ensure more efficient utilization of existing energy resources and to develop strategies that enhance the effectiveness of energy management processes. However, the inherently high variability of solar energy necessitates the use of more advanced forecasting models that go beyond traditional approaches. In this context, our study emphasizes the advantages provided by machine learning methods. Conversely, traditional statistical approaches such as ARIMA, linear regression, and heuristic methods often struggle to effectively model the sudden variations and complex nonlinear dynamics characteristic of solar energy output. These conventional methods generally assume stationarity and thus face difficulties adapting to rapidly evolving environmental factors or real-time inputs. Their predictive capabilities also tend to decline sharply when applied to high-dimensional and intricate datasets [4].

A recent development introduced a probabilistic ultra-short-term photovoltaic (PV) power forecasting framework that merges the Natural Gradient Boosting (NGBoost) algorithm with deep learning models. To extract abstract and meaningful patterns from time series data, the framework employs a neural network augmented with an attention mechanism, integrating Convolutional Neural Networks (CNN) and Bidirectional Long Short Term Memory (BILSTM) layers.The features extracted through this hybrid network serve as inputs to an optimized NGBoost model for final prediction generation.

In comparison with conventional quantile regression (QR) based deep learning models and traditional NGBoost techniques, the proposed hybrid approach demonstrates a markedly improved ability to capture PV power variability. The deterministic forecasting accuracy increased by 15% to 60%, depending on the scenario. For probabilistic forecasting, the model consistently exceeded the performance of baseline methods, offering greater precision and robustness. The Continuous Ranked Probability Score (CRPS) ranged between 0.0710 kW and 0.0898 kW, indicating an error reduction of 21–43% compared to QRbased models and 29–40% compared to standard NGBoost methods [5]. Among these algorithms, Gradient Boosting has gained significant attention due to its ability to construct a strong predictive model by iteratively correcting the errors of weak learners. Its high predictive accuracy, flexible structure, and adaptability to various types of data have made it one of the most prominent methods in the energy domain in recent years.

In this study, the performance of the Gradient Boosting algorithm in forecasting solar energy production is systematically examined under different train/test split ratios. The primary objective is to analyze the impact of varying training proportions on model accuracy and error metrics, and to identify the optimal data partitioning ratio. In this regard, the study aims to contribute to the existing literature both methodologically and practically.

2. MATERIALS AND METHODS

This study utilized the "Solar Energy Power Generation Dataset" [6] to explore how solar power output correlates with various meteorological variables. Sourced from Stucom via the Kaggle platform, the dataset comprises hourly measurements and includes a wide range of features, such as temperature, humidity, atmospheric pressure, total precipitation, snowfall, cloud cover across different atmospheric levels (low, mid, high), shortwave radiation, wind speed and direction at multiple altitudes, angular parameters (zenith and azimuth), and power generation values in kilowatts. During the modeling process, the power output served as the target (dependent) variable.

Initially, the dataset was imported into the open-source data analysis tool Orange. To ensure machine learning model compatibility, variable names were normalized by converting all characters to lowercase and substituting spaces and special characters with underscores (“_”). The Gradient Boosting algorithm was chosen as the primary predictive model due to its capability to iteratively improve performance by correcting errors from weak learners, commonly decision trees. Since the task was framed as a regression problem, Mean Squared Error (MSE) served as the loss function. Key hyperparameters such as learning rate, maximum depth of trees, and the number of resampling iterations were kept fixed to isolate the effects of different train/test split ratios.

As illustrated in Figure 1, the data processing workflow was established within the Orange platform. The dataset was loaded using the File widget and explored via the Data Table widget. The Gradient Boosting model was trained and evaluated using several Test and Score widgets, each configured with different training/test splits (50%, 60%, 66.6%, 70%, 75%, 80%, and 90%). Model evaluation employed three primary metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2), allowing for a comparative assessment of model performance.

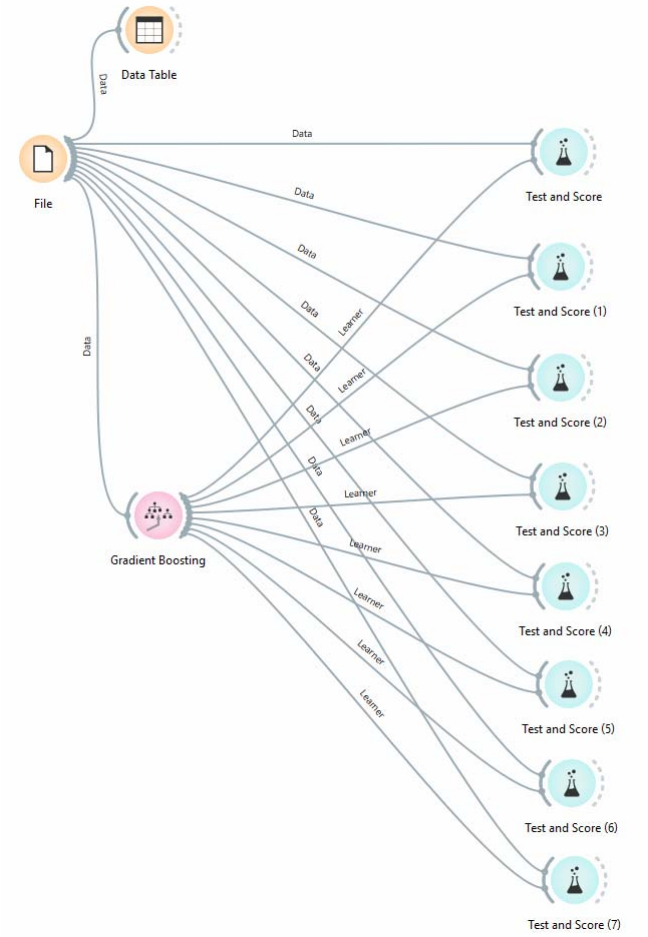

Fig. 1. Workflow for performance evaluation of the Gradient Boosting model under different train/test split ratios in the Orange open-source data analysis software

Gradient Boosting Algorithm

Gradient Boosting (GB) is a robust machine learning technique that consists of three key elements: a specified loss function, weak learners, and an additive modeling framework. The selection of the loss function varies according to the problem type; for regression tasks, squared error loss is typically employed, whereas classification problems often utilize the logarithmic loss function. In GB, weak learners are usually decision trees, each trained sequentially to correct the residual errors made by the preceding trees.

Because the model is constructed additively, new trees are appended one after another without modifying the previously built ones. This method uses gradient descent optimization to iteratively adjust the trees' parameters in order to minimize the loss function. As a result, Gradient Boosting effectively builds a strong ensemble by gradually reducing prediction errors with each added tree [7].

Performance Metrics

Three different evaluation metrics [8] were used to assess the performance of the machine learning regression models.

Coefficient of Determination (R^2):

The Coefficient of Determination, denoted as R^2, measures the proportion of variance in the dependent variable that is explained by the independent variables. It functions as an indicator of the regression model's goodness-of-fit, where values range from 0 to 1, with higher values signifying a better fit.

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y}_j)^2} \quad (1)$$

Root Mean Squared Error (RMSE):

Root Mean Squared Error (RMSE) is a widely used metric that measures the average magnitude of prediction errors by calculating the square root of the average squared differences between actual and predicted values. A smaller RMSE indicates better model accuracy and performance.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad (2)$$

Mean Absolute Error (MAE):

Mean Absolute Error (MAE) measures the average magnitude of errors between predicted and actual values, without considering their direction. It is calculated as the mean of the absolute differences between the forecasted and observed values. Lower MAE values indicate greater accuracy and better model performance.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \quad (3)$$

3. RESULTS

Table 1 summarizes the performance results of the Gradient Boosting algorithm across different train/test split ratios. The R^2 values ranged from 0.774 to 0.790, with the highest value achieved at a 90% training ratio. Similarly, RMSE and MAE values exhibited the same trend, with the lowest error metrics also observed at the 90% training proportion. These findings indicate that the model demonstrates stable performance and achieves optimal results particularly when trained with 90% of the data.

Table 1 *Performance comparison of the Gradient Boosting algorithm across different train/test split ratios*

% train/test	R^2	RMSE	MAE
%50	0.776	443.937	301.129
%60	0.774	444.231	299.870
%66	0.776	441.062	297.282
%70	0.779	439.022	296.235
%75	0.782	435.756	294.359
%80	0780	438.278	294.568
%90	0.79	428.959	289.195
%95	0.781	437.226	294.775

Figure 2 illustrates the variation of R^2 values according to different train/test split ratios. Upon examining the graph, a noticeable increase in the coefficient of determination is observed starting from the 75% training ratio, reaching its maximum at 90%. At a 95% training ratio, a slight decrease in R^2 is detected, which may indicate a tendency of the model towards overfitting.

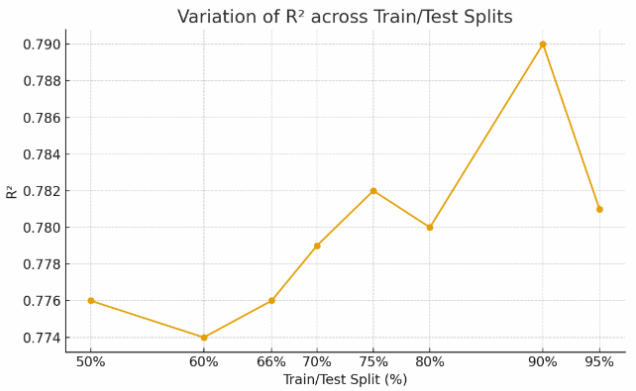

Fig. 2. Variation of the Coefficient of Determination (R^2) values across different train/test split ratios

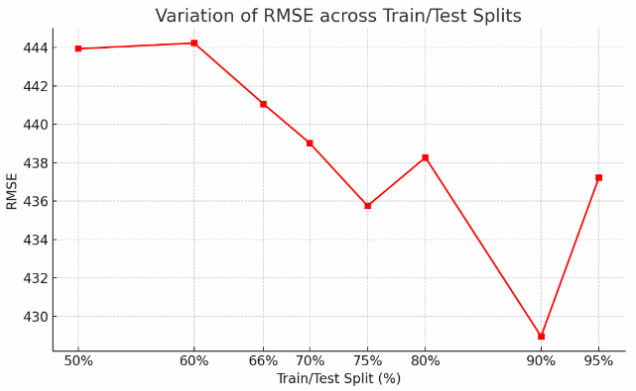

Fig. 3. Variation of Root Mean Squared Error (RMSE) values across different train/test split ratios

Figure 3 illustrates the variation of RMSE values across different train/test split ratios. It was observed that as the training ratio increased, the error values decreased, with the lowest RMSE recorded at the 90% training ratio. This indicates that using a larger amount of training data improves the model's error prediction performance. However, the increase in RMSE at the 95% training ratio suggests that the model's generalization capability may be limited when trained with excessively high proportions of the data.

Figure 4 illustrates the distribution of Mean Absolute Error (MAE) values across different train/test split ratios. According to the results, MAE values remain relatively close between 50% and 70% training ratios, followed by a gradual decreasing trend starting from 75%. The lowest MAE value was obtained at the 90% training ratio, with a slight increase observed at 95%. This pattern is consistent with the RMSE results, confirming that the model achieves its best performance at the 90% training ratio.

Fig. 4. Variation of Mean Absolute Error (MAE) values across different train/test split ratios

Overall, both the table and figures demonstrate that the Gradient Boosting algorithm exhibits a highly stable performance across different train/test split ratios. However, the most balanced and optimal results were observed at the 90% training ratio, leading to the conclusion that this ratio is the most suitable split for solar energy power generation forecasting.

4. CONCLUSION

In this study, the performance of the Gradient Boosting algorithm for solar energy power generation forecasting was evaluated under different train/test split ratios. The Kaggle dataset utilized provided a comprehensive sample illustrating the relationship between meteorological parameters and solar energy production. The analyses revealed that the algorithm demonstrated stable performance across all ratios, with the highest accuracy ($R^2 = 0.790$) and the lowest error values (RMSE=428.959, MAE=289.195) achieved specifically at the 90% training ratio. However, a slight decline in performance observed at the 95% training ratio indicated a potential risk of overfitting. These findings clearly establish Gradient Boosting as a reliable and effective method for solar energy production forecasting. Furthermore, the optimal performance at the 90% training ratio suggests that this split ratio is the most suitable choice for data partitioning. Future studies may consider employing other boosting-based algorithms (e.g., XGBoost, LightGBM, CatBoost) and hyperparameter optimization techniques to achieve higher accuracy and improved generalization capabilities.

REFERENCES

[1] Persson C., Bacher P., Shiga T., Madsen H., Multi-site solar power forecasting using gradient boosted regression trees, Solar Energy, 150 (2017) 423-436

[2] Marquez R., Coimbra C.F., Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database, Solar Energy, 85(5) (2011) 746-756

[3] Pedro H.T., Coimbra C.F. Assessment of forecasting techniques for solar power production with no exogenous inputs, Solar Energy, 86(7) (2012) 2017-2028

[4] Abdelsalam H., Souri A., İnanç N., A time series forecasting approach based on gradient boosting method for IoT-based solar energy production systems, Energy Storage and Saving (2025)

[5] Song Z., Xiao F., Chen Z., Madsen H., Probabilistic ultra-short-term solar photovoltaic power forecasting using natural gradient boosting with attention-enhanced neural networks, Energy and AI 20 (2025) 100496

[6] Stucom, Solar Energy Power Generation Dataset (2020) [Data set] Kaggle, https://www.kaggle.com/datasets/stucom/solar-energy-power-generation-dataset

[7] Ayyadevara V.K., Pro machine learning algorithms, Apress: Berkeley, CA, USA, (2018) 283-297

[8] Chicco D., Warrens M.J., Jurman G., The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, Peerj computer science, 7, e623 (2021)