Photovoltaic Energy Prediction: Evaluating Feature Selection Methods for Enhanced Forecasting

Marcin Zalasiński
Częstochowa University of Technology/Department of Artificial Intelligence
Częstochowa, Poland *marcin.zalasinski@pcz.pl*

Tomasz Szczepanik
Częstochowa University of Technology/Department of Artificial Intelligence
Częstochowa, Poland *tomasz.szczepanik_22@pcz.pl*

Magdalena M. Scherer
Częstochowa University of Technology/Faculty of Management
Częstochowa, Poland *magdalena.scherer@pcz.pl*

Jolanta Zalasińska
CEO - ERP Serwis sp. z o.o. sp. k.
Częstochowa, Poland *jolanta.zalasinska@erpserwis.net*

Abstract

Renewable energy, particularly photovoltaic (PV) systems, plays a crucial role in sustainable energy development. Its production is largely dependent on external factors, especially weather conditions, making the forecasting of generated energy a significant research challenge. Selecting appropriate features that influence electricity production can enhance forecasting accuracy. This paper evaluates various feature selection methods relevant to energy output, aiming to identify the most effective selection strategy and determine the most influential variables. Three groups of methods were analyzed: correlation-based statistical methods, ensemble-based importance metrics, and univariate significance tests. The results highlight the importance of choosing suitable feature selection algorithms to improve the accuracy of PV energy production forecasting.

Keywords: Predictive models, Energy forecasting, Feature selection, Renewable energy, Photovoltaic systems

1. Introduction

Accurate energy production forecasting is crucial for optimizing the performance of photovoltaic (PV) systems. This task is challenging due to the variability and complexity of weather data that influence energy output. Proper feature selection plays a vital role in identifying the most relevant variables affecting the process under study [2]. This study focuses on evaluating three groups of feature selection methods used in PV energy production forecasting: correlation-based statistical methods, ensemble-based importance metrics, and univariate significance tests. The goal is to identify the most effective combination of feature selection strategies and predictive algorithms for photovoltaic datasets, ultimately improving the accuracy and robustness of energy forecasting models.

2. Brief Literature Review of Feature Selection Methods in PV Applications

In the field of PV energy forecasting, feature selection plays a key role in improving the accuracy, efficiency, and interpretability of predictive models. Due to the high dimensionality and variability of meteorological and operational data, various approaches have been developed and applied in the literature to identify the most relevant input variables.

Some studies adopt correlation-based methods to select features that exhibit strong linear relationships with PV energy output. These approaches often rely on metrics such as the Pearson correlation coefficient to identify solar-related variables - particularly global horizontal irradiance and clearness index - as the most significant predictors of system performance (see e.g. [1]). Other research utilizes importance measures derived from ensemble models, especially Random Forests, which are well suited to handling non-linear interactions and high-dimensional datasets. These methods evaluate the contribution of each feature to reducing model variance, effectively identifying variables such as temperature, wind speed, and solar irradiance as crucial for short-term and day-ahead forecasting (see e.g. [3]). Additionally, univariate statistical tests, including ANOVA F-tests and chi-squared tests, are used in several works as a computationally efficient means of ranking feature relevance. These tests assess the statistical significance of each variable with respect to the target output and are often applied in early stages of model development (see e.g. [5]). However, such methods typically ignore multicollinearity and potential feature interactions.

Given the diversity of approaches, feature selection techniques in PV forecasting can generally be categorized into three major groups frequently discussed in the literature: correlation-based statistical methods, ensemble-based importance metrics, and univariate significance testing. Each of these has specific strengths and limitations depending on the characteristics of the data and the type of predictive model employed.

3. Methodology

This study analyzes the effectiveness of different feature selection approaches, grouped into three categories, aimed at optimizing PV energy forecasting. The objective is to evaluate and compare various feature selection techniques to identify the group of methods that most effectively reduces redundancy and noise in the dataset while preserving variables that significantly enhance predictive performance. The focus is on meteorological variables most relevant to PV systems.

The dataset used in this study includes three years (2022-2024) of PV energy production records from the University of Mein [4], combined with comprehensive meteorological data obtained from satellite observations and publicly available online platforms [6, 7]. A total of 18 different features were considered. This diverse set of environmental descriptors enabled a broad characterization of solar conditions relevant to energy production.

This study evaluates feature selection methods categorized into three groups: correlation-based statistical methods, ensemble-based importance metrics, and univariate significance testing. As a representative of the first group, the Pearson correlation-based feature selection method was used. It employs the Pearson correlation coefficient, calculated as:

$$\rho(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2]}}, \tag{1}$$

where X and Y represent the feature and the target variable, X_i and Y_i are individual observations, and $\bar{X}$ and $\bar{Y}$ denote their means. Features with a high absolute correlation value (closer to ±1) are considered strongly predictive of the target.

The second group of methods was represented by Random Forest-based feature selection. It uses the feature importance measure I_j for feature j, calculated as:

$$I_j = \frac{1}{T} \sum_{t=1}^{T} \Delta_{t,j}, \tag{2}$$

where T is the total number of trees in the forest, and $\Delta_{t,j}$ is the contribution of feature j in tree t.

The third group employed ANOVA F-test feature selection. It uses the F-statistic, which measures the ratio of variability between groups to variability within groups, expressed as:

$$F = \frac{\text{Between-group variability}}{\text{Within-group variability}} = \frac{\sum n_i(\bar{X}_i - \bar{X})^2}{\sum(X_{ij} - \bar{X}_i)^2}, \tag{3}$$

where n_i is the number of observations in group i, X_{ij} denotes the j-th observation in group i, $\bar{X}_i$ is the mean of group i, and $\bar{X}$ is the overall mean.

To evaluate the impact of each feature selection method on model performance, three regression algorithms were employed: K-Nearest Neighbors (KNN), Decision Trees (DT), and XGBoost (XGB). These models were chosen for their capacity to capture both linear and non-linear relationships, as well as for their established utility in renewable energy forecasting tasks.

Model performance was evaluated using three standard metrics: (1) RMSE, to capture the magnitude of prediction errors with emphasis on large deviations, (2) MAPE, to express accuracy as a percentage of actual values, (3) R^2, to indicate the proportion of variance explained by the model. Together, these metrics provide a comprehensive assessment of forecasting accuracy and generalization.

4. Simulation Results

As part of this study, simulations were carried out in a custom testing environment implemented in Python, in accordance with the assumptions outlined in Section 3.

Table 1 summarizes the features selected by each of the three feature selection methods from the 18 considered. Pearson correlation-based selection identified features with strong linear relationships to energy output - such as K_{T}, I_{UV}, and R_{DIFF} - effectively reducing dimensionality and noise. Random Forest-based selection captured non-linear dependencies, selecting variables like T_{max} and R_{TOA}, but sometimes introduced seemingly less relevant features. ANOVA F-test selection emphasized individually significant features like H_{sun}, I_{UV}, and K_{T}, but could not account for multicollinearity or feature interactions. All methods were evaluated using three regression models: XGBoost, Decision Tree, and K-Nearest Neighbors.

In Table 2, the performance of three regression models for the considered feature selection techniques is presented. The results revealed significant variation in model performance depending on the selected method. The best results were achieved by combining Pearson correlation-based feature selection (a correlation-based statistical method) with the K-Nearest Neighbors model, reaching an RMSE of 424.33, a MAPE of 6.55%, and an R^2 of 0.907. This feature selection method also achieved the best average results across all regression models (RMSE: 502.46, MAPE: 7.33%, R^2: 0.862). In contrast, the average prediction performance of all models without feature selection was the worst (RMSE: 608.43, MAPE: 8.94%, R^2: 0.801), confirming the validity of the adopted assumptions.

The obtained results confirm that the appropriate selection of weather-related features used for predicting the electricity production level of photovoltaic panels has a significant impact on forecast accuracy. The high effectiveness of the Pearson correlation-based method highlights the value of linear dependency-based filtering in identifying the most relevant input variables. Among the analyzed feature selection methods, the weakest performance was observed for the Random Forest-based approach when combined with the Decision Tree regression model. Although both this method and the ANOVA F-test selected the same number of features, the performance of the Decision Tree model was lower when using features selected by Random Forest. This suggests that, beyond the number of selected features, the type and structure of the variables - as well as their interaction with the applied regression model - play a crucial role in prediction quality.

Table 1. Description of features selected from a set of 18, based on all considered feature selection techniques.

Symbol	Feature Name	Unit	Selected by technique using		
			Correlation-Based Methods	**Random Forest-Based Importance Measures**	**Univariate Statistical Tests**
T_{max}	The highest temperature	°C	no	yes	yes
H_{sun}	The total number of sunlight hours	h	no	yes	yes
I_{UV}	Strength of sunburn-producing ultraviolet radiation	Unitless	yes	yes	yes
H_{rel}	Relative humidity (the amount of water vapor present in air)	%	no	yes	no
R_{TOA}	The total solar irradiance incident on a horizontal plane at the top of the atmosphere	W/m^2	no	yes	yes
R_{DNI}	Direct solar irradiance on a horizontal plane aligned perpendicularly to the sun	W/m^2	no	yes	yes
R_{DIFF}	The diffuse solar irradiance incident on a horizontal plane at the surface of the earth	W/m^2	yes	no	no
K_{T}	A fraction representing clearness of the atmosphere	Unitless	yes	no	yes

Table 2. Model Performance for Different Feature Selection Methods. Average results for each method are bold. Lower values of RMSE and MAPE indicate better performance, whereas higher values of R^2 are preferable.

Feature Selection Methods	Predictive Model	RMSE	MAPE	R^2
None (all features are used)	XGBoost	598.25	9.45%	0.811
	Decision Tree	669.42	8.47%	0.758
	K-Nearest Neighbors	557.63	8.91%	0.835
	Average values	**608.43**	**8.94%**	**0.801**
Pearson correlation-based feature selection (correlation-based statistical method)	XGBoost	458.16	6.72%	0.889
	Decision Tree	624.89	8.71%	0.791
	K-Nearest Neighbors	424.33	6.55%	0.907
	Average values	**502.46**	**7.33%**	**0.862**
Random Forest-based feature selection (ensemble-based importance metric)	XGBoost	496.45	7.68%	0.868
	Decision Tree	729.37	8.21%	0.715
	K-Nearest Neighbors	517.80	9.91%	0.857
	Average values	**581.21**	**8.60%**	**0.813**
ANOVA F-test feature selection (univariate significance testing)	XGBoost	488.92	7.43%	0.871
	Decision Tree	558.20	7.79%	0.832
	K-Nearest Neighbors	529.41	8.48%	0.851
	Average values	**525.51**	**7.90%**	**0.851**

5. Conclusions

This study addressed a key issue in PV energy forecasting: the selection of the most relevant weather-related features affecting production levels. Three feature selection methods were tested, each representing a different category of techniques commonly used in photovoltaic energy prediction: Pearson correlation-based feature selection (a correlation-based statistical method), Random Forest-based feature selection (an ensemble-based importance metric), and ANOVA F-test feature selection (univariate significance testing). The conducted simulations demonstrated that the choice of feature selection method significantly impacts prediction accuracy. Correlation-based feature selection methods proved to be the most effective, offering an optimal balance between predictive accuracy, simplicity, and robustness. They focused on the most relevant linear features, such as the clearness index, the strength of sunburn-producing ultraviolet radiation, and the diffuse solar irradiance incident on the Earth's surface. These variables helped reduce noise and overfitting while keeping the models efficient. As part of future research, a hybrid feature selection approach is planned, combining the strengths of all the methods considered. We also intend to explore additional machine learning-based feature selection methods, including wrapper-based and model-agnostic approaches. Furthermore, the development of an interpretable regression model based on a flexible neuro-fuzzy system is envisioned.

Acknowledgments

The authors would like to acknowledge the use of the artificial intelligence tool ChatGPT for support in language editing and enhancing the clarity of the manuscript. All scientific content and intellectual contributions are entirely the work of the authors.

References

[1] Amer, H.N., Dahlan, N.Y., Azmi, A.M., Latip, M.F.A., Onn, M.S., Tumian, A.: Solar power prediction based on artificial neural network guided by feature selection for large-scale solar photovoltaic plant. Energy Reports 9(Supplement 12), pp. 262–266 (November 2023), https://doi.org/10.1016/j.egyr.2023.09.141, under a Creative Commons license

[2] Castangia, M., Aliberti, A., Bottaccioli, L., Macii, E., Patti, E.: A compound of feature selection techniques to improve solar radiation forecasting. Expert Systems with Applications 178, pp. 114979 (September 2021), https://doi.org/10.1016/j.eswa.2021.114979

[3] Massaoudi, M., Chihi, I., Sidhom, L., Trabelsi, M., Refaat, S.S., Oueslati, F.S.: Enhanced random forest model for robust short-term photovoltaic power forecasting using weather measurements. Energies 14(13), pp. 3992 (2021), https://www.mdpi.com/1996-1073/14/13/3992, this article belongs to the Topic Innovative Techniques for Smart Grids

[4] OpenEI - energy information on hundreds of topics crowdsourced from industry and government agencies. https://openei.org/wiki/PVDAQ/Sites/Univ._of_Maine_-_Presque_Isle (2024), accessed May 1, 2025

[5] Rahman, N.H.A., Hussin, M.Z., Sulaiman, S.I., Hairuddin, M.A., Saat, E.H.M.: Univariate and multivariate short-term solar power forecasting of 25mwac pasir gudang utility-scale photovoltaic system using lstm approach. Energy Reports 9(Supplement 11), pp. 387–393 (October 2023), https://doi.org/10.1016/j.egyr.2023.09.018, presented at the 7th International Conference on Renewable Energy and Conservation, ICREC 2022, Paris, France. Under a Creative Commons license

[6] The POWER project: Solar and meteorological data sets from NASA. https://power.larc.nasa.gov/ (2024), accessed May 1, 2025

[7] World Weather Online - current, forecast and historical weather. https://www.worldweatheronline.com/ (2024), accessed May 1, 2025