A Machine Learning System for Energy Forecasting with Feature Importance Analysis

Marcin Zalasiński[1][0000−0002−0009−6124], Tomasz Szczepanik[1][0009−0008−2170−2497], and Piotr Dobosz[2][0009−0003−5711−9141]

[1] Department of Artificial Intelligence, Częstochowa University of Technology, Al. Armii Krajowej 36, 42-200 Częstochowa, Poland
{marcin.zalasinski,tomasz.szczepanik}@pcz.pl
[2] Radom Academy of Economics, 26-610 Radom, Poland
pdobosz@ahns.pl

Abstract. This article presents a novel approach to forecasting energy production using machine learning techniques with feature selection. The study utilizes machine learning algorithms optimized using historical weather and real-time energy data. Feature selection is performed through a mathematical framework based on correlation coefficients and mutual information, ensuring the use of only the most significant predictors. SHAP (SHapley Additive exPlanations) is employed to analyze the impact of selected features on prediction outcomes, enhancing model transparency. Experimental results show high forecasting accuracy and robustness, highlighting the role of meteorological variables in energy production. This work advances predictive modeling for renewable energy systems and provides a comprehensive framework for identifying key factors influencing PV energy production, supporting improved forecasting and decision-making in energy management.

Keywords: Photovoltaic Energy Forecasting · Machine Learning Models · Feature Selection · Feature Importance.

1 Introduction

Accurate forecasting of photovoltaic (PV) energy production is essential for efficient energy management, operational planning, and grid stability. With the increasing adoption of renewable energy systems, reliable forecasting methods are crucial to address the dynamic nature of weather conditions and the nonlinear relationships between environmental factors and energy output. Machine learning techniques provide effective solutions by leveraging large datasets to model these complexities, enabling better integration of renewable energy into power grids and optimizing energy storage and distribution.

Recent studies emphasize the growing role of artificial intelligence in photovoltaic forecasting, highlighting the importance of model interpretability [15]. Explainable AI enhances both predictive performance and transparency, which is essential in energy management where decisions must be traceable, robust, and responsive to changing environmental conditions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

This paper presents a machine learning system for PV energy forecasting that leverages feature selection techniques to enhance predictive accuracy and model efficiency. The study utilizes a comprehensive dataset comprising historical PV energy production data, enriched with high-resolution meteorological parameters such as solar radiation, temperature, wind speed, and humidity. These datasets were preprocessed and synchronized to ensure consistency, facilitating the identification of the most impactful predictors.

The feature selection process focuses on identifying variables that maximize predictive performance while reducing redundancy and computational complexity. The proposed method considers both statistical relationships, such as correlation coefficients, and the contribution of individual features to model outputs. Unlike traditional approaches, this methodology accounts for feature interactions, ensuring a more informative and compact feature set that improves forecasting performance.

By performing feature importance analysis using SHapley Additive exPlanations (see e.g. [12]) after the feature selection process, this study provides valuable insights into the contribution of selected predictors to model performance. The results highlight key environmental factors affecting PV energy production, supporting better decision-making in energy management. Additionally, the findings demonstrate the effectiveness of the proposed system in enhancing forecasting accuracy while maintaining model efficiency, contributing to the broader adoption of renewable energy technologies in diverse operational scenarios.

The paper is organized into 5 sections. Section 2 discusses related works. Section 3 describes the methodology of the proposed algorithm. Section 4 presents the simulation results. Finally, conclusions are drawn in Section 5.

2 Related works

Recent years have seen a surge in machine learning applications for predicting solar energy production. Various approaches have been explored, with studies comparing different supervised learning algorithms, such as linear regression, Support Vector Machines (SVM), and Random Forest, for photovoltaic output forecasting [9]. Tree-based methods, particularly Random Forest, have demonstrated superior accuracy compared to conventional techniques. Additionally, feature selection has been identified as a crucial factor, as excessive input variables can lead to overfitting and reduced generalization [16]. Techniques like Pearson correlation [8] have been employed to determine the most influential features, highlighting the importance of solar radiation, temperature, and wind speed in forecast accuracy.

Hybrid models have also gained attention in solar energy forecasting [17]. One approach integrates neural networks with multiscale correlation-based methods to enhance short-term prediction stability, particularly under varying weather conditions. Another study combined time-series algorithms with Random Forest to leverage additional environmental data, such as air quality and weather met-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

rics, improving short-term forecast accuracy and capturing seasonal variations [7]. These findings underscore the potential of incorporating diverse meteorological and environmental factors to refine predictive performance.

Despite recent advancements, challenges remain in optimizing the selection and weighting of input variables. Many models aim to improve performance but lack a systematic approach to feature importance analysis [13]. Recent works have introduced hybrid models and advanced regression techniques to enhance forecasting robustness [11], yet the identification of the most relevant predictors remains underutilized. Decision tree-based models and hybrid frameworks continue to play a dominant role, with increasing emphasis on integrating relevant external variables, such as air quality indices and seasonal trends, to further enhance solar energy forecasting reliability.

3 Methodology

The proposed methodology (see Fig. 1) for predicting photovoltaic energy production follows a structured approach consisting of data preprocessing, feature selection, model development, and model evaluation, with an additional focus on feature importance analysis using SHapley Additive exPlanations (SHAP). Feature importance analysis using SHAP is performed after the feature selection process to gain deeper insights into the contribution of selected predictors to model performance.

The following subsections of the chapter provide detailed explanations of each step of the algorithm.

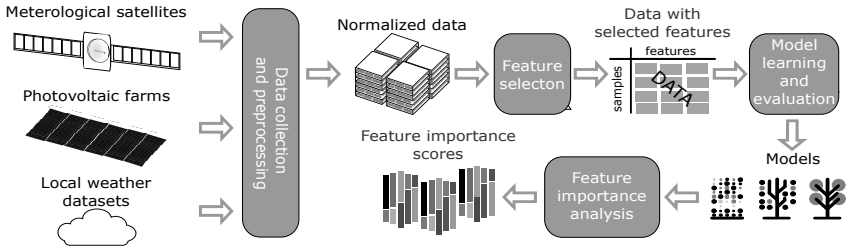

Fig. 1. General scheme of the proposed method.

3.1 Preprocessing

The preprocessing phase ensured data reliability for predictive modeling by cleaning meteorological and PV data from sources like NASA POWER [10] and World Weather Online [19]. Steps included duplicate removal, outlier detection (Z-score), missing value imputation, one-hot encoding for categorical variables, and normalization of continuous features. This produced a clean, balanced dataset, improving model stability and accuracy.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

3.2 Feature Selection

The feature selection method used in this study is purposefully designed to optimize variable choice. This enhances both prediction accuracy and model explainability while reducing computational complexity. Feature selection algorithm consists of the following steps: 1) Pearson correlation filtering to remove redundant features, 2) correlation-based clustering with representative selection, 3) tree-based Information Gain filtering, 4) XGBoost-based ranking using five models for robustness, 5) Recursive Feature Elimination to obtain the final subset.

The first stage involves analyzing correlations between variables and eliminating those that are highly correlated. This process is formalized using the Pearson correlation coefficient matrix R, defined as:

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}, \tag{1}$$

where r_{ij} is the correlation coefficient between feature i and feature j, n is the number of observations, and $\bar{x}_i$ is the mean of the i-th feature. This step is crucial as it ensures that highly correlated features ($r_{ij} > \theta$, where θ is a defined threshold, e.g., 0.8) are removed, minimizing redundancy and the risk of multicollinearity. Highly correlated features can distort the model's ability to accurately attribute importance to individual variables, as they effectively carry overlapping information. This redundancy can lead to inflated coefficients in linear models or diminished interpretability in machine learning models, making it difficult to discern the true contribution of each feature to the predictions.

After the initial filtering based on correlation, a more advanced clustering method is applied to group similar features. The goal is to form clusters G such that each group contains features with similar correlation structures. A representative feature x_G is chosen from each cluster using the formula:

$$x_G = \arg\min_{x_i \in G} \sum_{j \in G} r_{ij}, \tag{2}$$

where r_{ij} is defined as the correlation coefficient between feature i and j within the same cluster G. This step retains only the most informative features, ensuring that the retained subset captures all necessary information while minimizing dimensionality. Reducing the number of features is critical for improving model efficiency and reducing computational complexity, particularly in high-dimensional datasets. Minimizing features also helps to mitigate the risk of overfitting, where the model becomes too tailored to the training data and performs poorly on unseen data. By focusing on the most relevant features, the methodology enhances both the accuracy and interpretability of the predictive model.

The retained features are then evaluated using tree-based models, which are particularly effective for feature importance ranking. Tree-based algorithms calculate feature importance based on how much a feature reduces an impurity measure, such as entropy or Gini impurity, at each split. In the proposed method,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

Information Gain $IG(x_i)$ is employed as the criterion for evaluating feature importance. It measures how much a variable x_i contributes to reducing impurity when making decisions in the tree and is defined as follows:

$$IG(x_i) = \sum_{t \in T} p(t) \cdot \left[H(t) - \sum_{k \in \text{children}(t)} p(k \mid t) H(k) \right], \tag{3}$$

where $H(t)$ represents the impurity measure at node t, which can be entropy or Gini impurity depending on the algorithm used, $p(t)$ is the probability of reaching node t, and $p(k \mid t)$ is the probability of reaching child node k given node t.

The calculation of Information Gain involves assessing how much splitting the data based on a specific feature reduces the impurity at a node in a decision tree. Impurity measures how "mixed" the data is, with lower impurity indicating greater homogeneity concerning the target variable, such as energy production levels or classifications.

At the start, the data at a parent node has some degree of impurity, meaning the target values are a mix of different outcomes. When the data is split into child nodes using a particular feature, the impurity of each child node is calculated. The impurity of the overall split is determined as the weighted average of the impurity of the child nodes, where larger child nodes have a proportionally greater influence. Information Gain is then computed as the difference between the impurity of the parent node and the weighted impurity of the child nodes. A higher Information Gain indicates that the feature has effectively separated the data, making it a more valuable attribute for constructing the decision tree.

Removing highly correlated features helps ensure that the model isn't influenced by redundant information, which can inflate the importance of certain features or make the model less interpretable. This process also reduces complexity, improves computational efficiency, and enhances the reliability of the predictions.

To further enhance this process, XGBoost (Extreme Gradient Boosting) (see e.g. [2]) was employed. XGBoost is known for its efficiency and scalability, making it particularly suitable for large datasets with complex interactions. It builds an ensemble of weak prediction models, such as decision trees, to achieve strong predictive performance. In this study, XGBoost ranks features by calculating their F-scores, which measure how frequently and effectively a feature reduces prediction error across all trees in the ensemble. By using XGBoost, the model ensures that only the most impactful features are retained, improving both prediction accuracy and model explainability.

XGBoost's ability to handle non-linear relationships and its built-in regularization mechanism make it particularly useful for avoiding overfitting, ensuring that the model remains generalizable to new data. The combination of tree-based algorithms and XGBoost creates a powerful framework for feature selection, optimizing both predictive performance and computational efficiency.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

To further refine the selected subset, a Recursive Feature Elimination (RFE) (see e.g. [14]) method is applied. During each iteration, the least important feature $x_{\min}$ is removed based on its contribution to impurity reduction:

$$x_{\min} = \arg\min_{x_i \in S} \mathrm{IG}(x_i), \tag{4}$$

where S is the set of currently selected features. The iteration continues until only the most critical predictors remain. This process ensures that the final subset is highly optimized for prediction while minimizing the risk of overfitting.

In the process of feature selection, approximations play a crucial role in maintaining computational feasibility. The use of Pearson correlation analysis to eliminate highly correlated features is based on the assumption that linear relationships sufficiently capture redundancies between variables, even though some non-linear dependencies might remain. Similarly, the clustering of features to form groups with similar correlation structures and the subsequent selection of a single representative feature from each cluster is an approximation. This step simplifies the dataset while preserving its most critical information. These approximations, while not exhaustive, ensure that the feature selection process remains efficient and practical for large datasets, enabling subsequent model development to focus on the most impactful predictors.

3.3 Model Development and Evaluation Metrics

After selecting the optimal features, machine learning models - K-Nearest Neighbors (KNN) (see e.g. [18]), Decision Trees (see e.g. [4]), and Random Forest (see e.g. [7]) - were developed to predict energy production, leveraging their distinct capabilities. KNN is effective at capturing localized patterns and non-linear relationships by relying on proximity-based predictions, though it may face scalability challenges with large datasets. Decision Trees provide an interpretable, rule-based structure that highlights key decision paths and effectively handles both numerical and categorical data, requiring careful tuning to avoid overfitting. Random Forest enhances prediction robustness by combining multiple decision trees trained on random subsets of data, offering improved accuracy and resistance to noise while ranking feature importance. The evaluation of these models provided valuable insights into their strengths and applicability to energy production forecasting.

To assess the effectiveness of each model, standard error metrics are used. These include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2) (see e.g. [3]).

3.4 Feature Importance Analysis

Feature importance analysis focuses on understanding the impact of individual features on the predictive accuracy of a model. This is particularly crucial in applications such as energy forecasting, where the reliability of predictions directly

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

influences decision-making. By quantifying the contribution of each feature, feature importance analysis enhances transparency and supports the interpretability of machine learning models.

To achieve this, SHapley Additive exPlanations (SHAP) are employed as a robust method for assessing feature influence. SHAP values are grounded in cooperative game theory, providing a mathematically rigorous approach to distributing the contribution of each feature in a model's prediction. The SHapley value for a feature i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{5}$$

where S is any subset of features not containing i, N is the set of all features, $f(S \cup \{i\})$ is the prediction with feature i, and $f(S)$ is the prediction without feature i.

In the context of the developed models, SHAP values are used to evaluate how each feature contributes to the final predictions. This analysis includes generating summary plots and dependence plots. The summary plot aggregates the SHAP values of all features across all observations, showing both the magnitude and direction of each feature's contribution. Formally, the SHAP summary plot (SSP) is represented as:

$$SSP = \{(\phi_{i1}, x_{i1}), (\phi_{i2}, x_{i2}), \ldots, (\phi_{in}, x_{in})\}, \tag{6}$$

where ϕ_{ij} is the SHAP value for feature i and observation j, and x_{ij} is the corresponding feature value.

Dependence plots are used to visualize the interaction between two features. For a given feature i, the dependence plot (SDP) is defined as:

$$SDP = \{(x_{ij}, \phi_{ij}) \mid j = 1, \ldots, n\}. \tag{7}$$

This allows us to interpret not only the individual impact of each feature but also their interactions, highlighting regions of the feature space where certain variables have amplified or diminished effects on the prediction.

By incorporating these SHAP-based visualizations and metrics, the methodology provides a comprehensive understanding of the model's behavior, enabling users to identify critical features, mitigate biases, and refine models for improved predictive performance.

3.5 Dataset

The experiments utilized historical data on electricity production from photovoltaic panels installed at the University of Main. This dataset spanned four years (2018-2021) and included detailed weather information obtained from two sources: NASA POWER (see [10]) and World Weather Online (see [19]). The data from the University of Main provided actual energy production values, allowing for accurate forecasting under various weather conditions. All data used

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

in the experiments were preprocessed (see Section 3.1). A custom dataset was used instead of benchmark datasets to ensure alignment with real-world conditions, as it combines actual photovoltaic production data with localized weather information relevant for future applications in Poland.

Among the weather variables were features such as solar radiation intensity (I_{DIFF}), albedo (I_{ALB}), atmospheric transmissivity (K_{T}), UV index (I_{UV}), average temperature (T_{avg}), humidity (H_{rel}), visibility (V), and wind speed (v_{wind}). Table 1 provides a comprehensive overview of all features and their explanations. Each feature contributes to the analysis of energy production and the associated meteorological conditions.

4 Results

This section outlines the experimental procedure conducted to evaluate the effectiveness of the proposed photovoltaic energy production prediction system. It covers the datasets used, experiment configuration, results achieved by individual models, and analysis of the obtained results using evaluation metrics.

4.1 Experiment Configuration

We conducted experiments on a dataset split into a training set (80%), containing historical data from 2018 to 2021, and a testing set (20%). To mitigate overfitting, we applied 5-fold cross-validation. All models were trained on the same data and then evaluated on the testing set.

The models from Section 3.3 were configured as follows: 1) K-Nearest Neighbors - we set the number of neighbors to $k = 5$ and used the Euclidean distance measure; 2) Decision Tree - we set a maximum tree depth of 10 and required a minimum of 5 samples per leaf; 3) Random Forest - we used 100 trees, each with a maximum depth of 10. For each split, 80% of the input features were randomly selected.

All simulations were performed in a custom test environment implemented in Python.

4.2 Data Processing and Feature Selection

Historical data on energy production and weather conditions were divided into training and testing sets.

To identify the most relevant input variables, an automated feature selection method was applied (see Section 3.2). It removed redundant or highly correlated variables ($r > 0.8$) and retained those with the greatest impact on model performance, improving stability and explainability.

XGBoost was then used to train five models under varying conditions. This ensemble approach captured diverse patterns in the data, enhancing generalization and reducing overfitting. Each model ranked features using the F-score metric (Fig. 2).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

Table 1. Features used in the simulations (weather features and the amount of produced energy).

Feature Symbol	**Feature Name**	**Unit**
$T_{\max}$	The highest temperature	°C
$T_{\min}$	The lowest temperature	°C
T_{avg}	The mean temperature	°C
S_{total}	The total snowfall	cm
H_{sun}	The total number of sunlight hours	h
I_{UV}	Strength of sunburn-producing ultraviolet radiation	unitless
I_{moon}	The percentage of the Moon's visible disk illuminated	%
v_{wind}	The speed of the wind	km/h
0_{wind}	The meteorological wind direction	°
H_{rel}	The amount of water vapor present in air	%
V	The maximum distance one can see	km
P	The force exerted by the atmosphere at a point on the earth's surface	hPa
R_{TOA}	The total solar irradiance incident on a horizontal plane at the top of the atmosphere	W/m^2
I_{DNI}	Direct solar irradiance on a horizontal plane aligned perpendicularly to the sun	W/m^2
I_{SRF}	The all-sky rate of reflectivity of the earth's surface	W/m^2
I_{DIFF}	The diffuse solar irradiance incident on a horizontal plane at the surface of the earth	W/m^2
P_{corr}	The bias-corrected average of total surface precipitation in water mass	mm
K_{T}	A fraction representing clearness of the atmosphere	unitless
E_{prod}	The amount of energy produced by the solar plant	kWh

In model 1, K_T (clearness index) and R_{TOA} (solar irradiance at the top of atmosphere) were the top predictors, with H_{sun} and I_{DIFF} also contributing significantly. This confirmed the importance of atmospheric conditions in PV energy production.

For models 2-5, K_T, I_{DIFF} and H_{sun} consistently remained the top-ranked features, emphasizing their fundamental role across all models. However, there were notable differences in the ranking of other features. In model 2, I_{UV} (ultraviolet radiation) became more prominent, indicating the importance of ultraviolet radiation under certain weather conditions. Model 3 introduced H_{rel} (water vapor present in air) as a important feature. In model 4, I_{UV} maintained moderate importance, suggesting that ultraviolet radiation again might have secondary effects on energy production. Meanwhile, model 5 focused primarily on atmospheric transmissivity, solar irradiance and sunlight hours, with fewer additional variables contributing significantly. Despite these differences, the consistent ranking of K_T, I_{DIFF} and H_{sun} across all models confirmed their pivotal influence on the prediction of photovoltaic energy production.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

After ranking, Recursive Feature Elimination (RFE) was applied to iteratively remove less important features. This resulted in a compact, optimized feature set of four variables (Table 2).

The final selection balanced prediction accuracy, interpretability, and efficiency. Although up to 10 features were initially identified, only four were retained based on correlation filtering and F-score rankings.

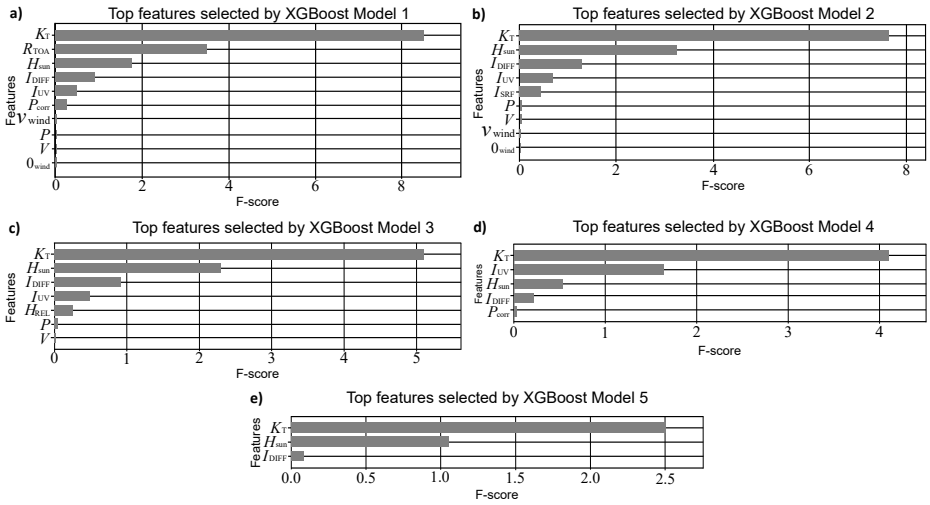

Fig. 2. Top features selected by: a) XGBoost Model 1, b) XGBoost Model 2, c) XGBoost Model 3, d) XGBoost Model 4, e) XGBoost Model 5.

Table 2. Final set of features selected by the RFE algorithm.

Feature Symbol	**Feature Name**	**Meaning for Models**
H_{sun}	The total number of sunlight hours	Medium
I_{UV}	Strength of sunburn-producing ultraviolet radiation	High
I_{DIFF}	Diffuse solar irradiance incident on a horizontal plane at the surface of the earth	High
K_{T}	A fraction representing clearness of the atmosphere	Medium

4.3 Evaluation of the Algorithm

Each model was optimized using cross-validation techniques, and the final parameters were selected based on the results of the RMSE, MAPE, and R^2 met-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

rics. The models' performances were compared, and the results are summarized in Table 3.

Table 3. Results of predictive models.

Model	RMSE	MAPE	R^2
Results for selected features			
K-Nearest Neighbors	140.56	14.50%	0.85
Decision Tree	120.45	12.34%	0.89
Random Forest	105.30	10.78%	0.92
Results for all features			
K-Nearest Neighbors	52837.74	100.57%	0.85
Decision Tree	81776.65	86.54%	0.64
Random Forest	48414.28	77.32%	0.87

The results indicate that the Random Forest model consistently outperforms the other models across both scenarios—using selected features and using all features. With the selected features, Random Forest achieves the lowest RMSE (105.30) and MAPE (10.78%) while obtaining the highest R^2 score (0.92), demonstrating its ability to generalize well and deliver accurate predictions. When all features are used, Random Forest still performs better than the other models, but its RMSE and MAPE values increase significantly (48414.28 and 77.32%, respectively), likely due to the inclusion of irrelevant or redundant features.

However, it should be noted that direct comparisons with other authors' methods are challenging due to the specificity of the datasets used in this study and the lack of publicly available data for replication and simulation. Nevertheless, we compare our results (RMSE = 105.30, MAPE = 10.78%, $R^2 = 0.92$) with the best-performing models (in the context of R^2) reported in [1] and [6]. In [6], which employs the Random Forest algorithm, the reported metrics are RMSE = 252.11, MAPE = 10.07%, and $R^2 = 0.72$. In contrast, [1] uses the Extra Trees method and achieves RMSE = 41.92, MAPE = 0.08%, and $R^2 = 0.91$. In this study, we prioritize R^2 as a key performance indicator, as it quantifies the proportion of variance in the target variable explained by the model. This is particularly important in photovoltaic energy forecasting, where capturing variability caused by dynamic environmental conditions is critical for producing reliable predictions.

Overall, these results highlight the importance of feature selection in improving model performance. Random Forest, in particular, demonstrates robustness and accuracy, making it the most suitable model for this task.

4.4 Importance of features

To analyze the impact of individual features on prediction outcomes, the SHAP tool was utilized, providing detailed insights into feature importance for the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

Table 4. SHAP values for K-Nearest Neighbors, Decision Tree and Random Forest models.

K-Nearest Neighbors model			
K_{sun}	K_{T}	I_{DIFF}	I_{UV}
2.037×10^{-10}	2.037×10^{-10}	2.037×10^{-10}	2.037×10^{-10}
1.048×10^{-10}	1.048×10^{-10}	1.048×10^{-10}	1.048×10^{-10}
0.000	0.000	0.000	0.000
1.048×10^{-10}	1.048×10^{-10}	1.048×10^{-10}	1.048×10^{-10}
1.543×10^{-10}	1.543×10^{-10}	1.543×10^{-10}	1.543×10^{-10}
-3.027×10^{-10}	-3.027×10^{-10}	-3.027×10^{-10}	-3.027×10^{-10}
-1.281×10^{-10}	-1.281×10^{-10}	-1.281×10^{-10}	-1.281×10^{-10}
4.191×10^{-10}	4.191×10^{-10}	4.191×10^{-10}	4.191×10^{-10}
-2.561×10^{-10}	-2.561×10^{-10}	-2.561×10^{-10}	-2.561×10^{-10}
1.048×10^{-10}	1.048×10^{-10}	1.048×10^{-10}	1.048×10^{-10}
Decision Tree model			
H_{sun}	K_{T}	I_{DIFF}	I_{UV}
380.920	1183.189	-3626.670	-90.425
160.514	104.241	-179.825	-12.420
433.181	1502.080	-2575.660	-21.308
380.920	1183.189	-3626.670	-90.425
160.514	104.241	-179.825	-12.420
433.181	1502.080	-2575.660	-21.308
433.181	1502.080	-2575.660	-21.308
380.920	1164.897	-3590.084	-90.425
380.920	1183.189	-3626.670	-90.425
433.181	1502.080	-2575.660	-21.308
Random Forest model			
H_{sun}	K_{T}	I_{DIFF}	I_{UV}
208.762	504.179	-878.081	14.684
-440.186	-531.175	798.061	558.297
258.659	794.225	-478.862	733.082
-573.083	-105.456	318.214	235.924
535.856	-1295.585	526.573	239.871
274.316	23.506	808.128	221.667
-260.073	-817.188	-48.584	150.036
135.853	-329.904	894.965	201.821
321.430	484.074	-870.427	14.911
248.315	49.301	799.588	221.409

K-Nearest Neighbors, Decision Tree, and Random Forest models. The SHAP summary plots (see Fig. 3) highlight I_{DIFF} and K_T as the most influential features across all models, with H_{sun} and I_{UV} having more minor roles. The Decision Tree and Random Forest models demonstrated greater sensitivity to these key features, exhibiting a wider range of SHAP values, whereas the K-Nearest Neighbors model showed smaller SHAP values, reflecting a more balanced but less feature-specific prediction pattern.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

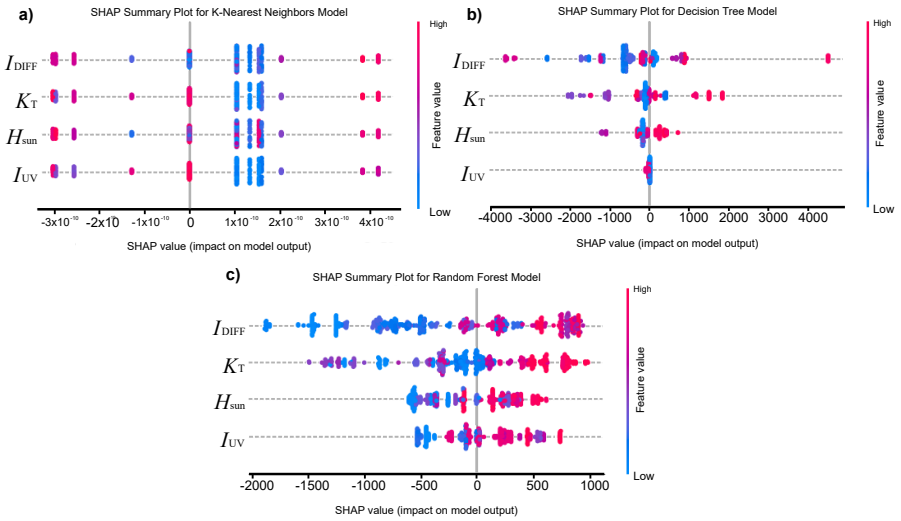

Fig. 3. SHAP summary plots for: a) K-Nearest Neighbors Model, b) Decision Tree Model, c) Random Forest Model.

The SHAP summary plots visualize feature importance across the dataset. The x-axis shows each feature's contribution to the prediction (positive or negative), while the y-axis ranks features by overall importance. Each dot represents a sample, with color indicating the feature's value (blue = low, red = high), helping to reveal how feature magnitude influences model output.

For example, I_{DIFF} (diffuse irradiance) and K_T (clearness index) typically show higher SHAP values when their raw values are high (Fig. 3b–c), confirming their strong positive impact on photovoltaic energy production. This pattern suggests that clearer atmospheric conditions and higher levels of scattered sunlight both favor energy generation, though in different ways. Importantly, these SHAP visualizations help uncover nonlinear interactions and context-dependent effects that traditional feature importance methods may overlook.

The SHAP values for the K-Nearest Neighbors model are very small—on the order of 10^{-10} or effectively zero (e.g., 2.037×10^{-10}, 0.000×10^{0}; see Table 4) - indicating that KNN does not assign significant predictive weight to any individual feature, including H_{sun}, K_T, I_{DIFF}, and I_{UV}. In contrast, the Decision Tree model exhibits large and highly variable SHAP values: K_T and H_{sun} register high positive contributions (e.g., 1183.189, 1502.080 for K_T; 380.920, 433.181 for H_{sun}), while I_{DIFF} shows strong negative influence (e.g., -3626.670, -2575.660), suggesting that increased diffuse irradiance lowers predicted output. The Random Forest model demonstrates a more moderated but still substantial range of SHAP values, with K_T and I_{DIFF} again as dominant predictors. For example, H_{sun} shows both positive and negative contributions (e.g., 208.762, -440.186), reflecting its context-dependent role. The wide vari-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

ance in K_T's SHAP values highlights its flexible influence under different atmospheric conditions, while I_{DIFF} maintains a consistently strong negative impact when values are high.

In summary, SHAP analysis provides valuable insight into both the importance and the conditional effects of key features. KNN offers limited interpretability due to negligible SHAP values, the Decision Tree model is highly sensitive to specific variables, and Random Forest delivers a more stable yet informative feature interpretation - particularly for K_T and I_{DIFF} - enhancing understanding of the factors driving photovoltaic energy predictions.

5 Conclusions

This paper presents a machine learning-based system for photovoltaic (PV) energy forecasting, integrating feature selection techniques and explainability methods to enhance prediction accuracy and model interpretability. The proposed methodology leverages Recursive Feature Elimination (RFE) and SHapley Additive exPlanations (SHAP) to refine the feature set and analyze the contribution of selected predictors. The experimental results demonstrate that an optimized feature set significantly improves model performance while maintaining computational efficiency.

The evaluation of various machine learning models, including K-Nearest Neighbors, Decision Trees, and Random Forest, highlights the high predictive capability of tree-based algorithms, particularly Random Forest, which achieved the lowest RMSE and highest R^2. Feature importance analysis confirmed that atmospheric transmissivity (K_T), diffuse solar irradiance (I_{DIFF}), and sunlight hours (H_{sun}) play a dominant role in PV energy production forecasting. The use of SHAP values provided valuable insights into how these factors influence model predictions, enhancing the transparency and interpretability of the forecasting system.

Future work will focus on developing a system to identify optimal locations for photovoltaic farms using heatmap-based feature importance analysis. By incorporating key environmental, seasonal, and temporal variables, the system will support precise long-term forecasting. The flexible framework will also allow integration of alternative algorithms, such as neural networks and gradient boosting, to enhance predictive performance.

References

1. Bakht, M.P., et al.: Advanced automated machine learning framework for photovoltaic power output prediction using environmental parameters and SHAP interpretability. Results Eng. 25, 103838 (2025). https://doi.org/10.1016/j.rineng.2024.103838
2. Bentéjac, C., Csörgő, A., Martínez-Muñoz, G.: A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54(3), 1937–1967 (2020). https://doi.org/10.1007/s10462-020-09896-5

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3

3. 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 Comput. Sci. 7, e623 (2021). https://doi.org/10.7717/peerj-cs.623
4. Costa, V.G., Pedreira, C.E.: Recent advances in decision trees: an updated survey. Artif. Intell. Rev. 56, 4765–4800 (2023). https://doi.org/10.1007/s10462-022-10275-5
5. Christen, P., Hand, D.J., Kirielle, N.: A review of the F-measure: its history, properties, criticism, and alternatives. ACM Comput. Surv. 56(3), 73:1–73:24 (2023). https://doi.org/10.1145/3606367
6. Gaboitaolelwe, J., et al.: Machine learning based solar photovoltaic power forecasting: a review and comparison. IEEE Access 11, 40820–40845 (2023). https://doi.org/10.1109/ACCESS.2023.3270041
7. Guan, L., Zou, L.: Study on solar power prediction model by random forest method based on a numerical weather prediction model. Authorea (2024). https://doi.org/10.22541/au.171037425.50188248/v1
8. Dai, G., et al.: Efficient method for photovoltaic power generation forecasting based on state space modeling and BiTCN. Sensors 24(20), 6590 (2024). https://doi.org/10.3390/s24206590
9. Gutiérrez, L., Patiño, J., Duque-Grisales, E.: A comparison of the performance of supervised learning algorithms for solar power prediction. Energies 14(15), 4424 (2021). https://doi.org/10.3390/en14154424
10. NASA: The POWER project: Solar and meteorological data sets from NASA. `https://power.larc.nasa.gov`, last accessed 18 Apr 2025.
11. Nguyen, T.N., Müsgens, F.: What drives the accuracy of PV output forecasts? Appl. Energy 323, 119603 (2022). https://doi.org/10.1016/j.apenergy.2022.119603
12. Nohara, Y., et al.: Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. Comput. Methods Programs Biomed. 214, 106584 (2022). https://doi.org/10.1016/j.cmpb.2021.106584
13. Perera, M., et al.: Multi-resolution, multi-horizon distributed solar PV power forecasting with forecast combinations. Expert Syst. Appl. 205, 117690 (2022). https://doi.org/10.1016/j.eswa.2022.117690
14. Ramezan, C.A.: Transferability of Recursive Feature Elimination (RFE)-derived feature sets for support vector machine land cover classification. Remote Sens. 14(24), 6218 (2022). https://doi.org/10.3390/rs14246218
15. Shukla, V., et al.: An explainable artificial intelligence based approach for the prediction of key performance indicators for 1 megawatt solar plant under local steppe climate conditions. Eng. Appl. Artif. Intell. 131, 107809 (2024). https://doi.org/10.1016/j.engappai.2023.107809
16. Rizk-Allah, R.M., et al.: Explainable AI and optimized solar power generation forecasting model based on environmental conditions. PLoS One 19(10), e0308002 (2024). https://doi.org/10.1371/journal.pone.0308002
17. Singh, U., et al.: Forecasting rooftop photovoltaic solar power using machine learning techniques. Energy Rep. 13, 3616–3630 (2025). https://doi.org/10.1016/j.egyr.2025.03.005
18. Peng, H., et al.: Survey on kNN. In: Proceedings of the 2nd International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA 2022), 17–19 June 2022. VDE Verlag, Berlin (2022)
19. World Weather Online: Current, forecast, and historical weather. `https://www.worldweatheronline.com`, last accessed 18 Apr 2025.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97567-7_3