ISSN 2707-1898 (print)
Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2025.01.149
Article received 21.04.2025 p.
Article accepted 01.05.2025 p.
UDC 004.94

Correspondence author
M. V. Melnyk
mykhailo.v.melnyk@lpnu.ua

R. V. Melnyk, M. V. Melnyk
Lviv Polytechnic National University, Lviv, Ukraine

A METHOD FOR FORECASTING THE ENERGY GENERATION OF A SOLAR POWER PLANT

The successful deployment of solar energy systems necessitates accurate forecasting of electricity production by photovoltaic power stations (PPS) to ensure the stable operation of power supply networks. This requirement stems from the need to maintain a real-time balance between electricity generation and consumption, which is achieved through the implementation of complex hierarchical control systems governing available energy sources. In this context, short-term forecasting of solar power generation is particularly critical, as it enables operational planning, economic dispatching, and grid stability.

This study presents the results of developing and validating forecasting methods while examining the impact of meteorological data structure and quality on prediction accuracy. Particular attention is paid to assessing the significance of various meteorological parameters using statistical correlation methods, including Pearson's linear correlation, Spearman's rank correlation, and Kendall's tau, as well as the Boruta feature selection algorithm. These methods provide complementary insights into the relevance and influence of environmental variables.

Based on the extracted significant predictors, a data-driven model using the *k*-Nearest Neighbors (kNN) algorithm was implemented. The research employed two distinct meteorological datasets, both containing environmental measurements and actual energy output data from the same photovoltaic facility. The first dataset was obtained from a weather station installed directly at the solar plant, offering high temporal and spatial precision. The second dataset was derived from open-access satellite-based weather sources linked to the plant's geographic coordinates, which are often used when on-site instrumentation is unavailable.

The results confirm that the use of on-site meteorological observations significantly improves model performance. For the kNN algorithm, the coefficient of determination (R^2) reached 0.99 using local data, compared to 0.95 with the satellite-based set. Additionally, metrics such as MAPE, MAE, and generation forecast error (PFG) support the superiority of models trained on accurate, high-resolution inputs. These findings highlight the importance of equipping solar energy facilities with dedicated meteorological sensors and integrating refined data into intelligent prediction frameworks.

Keywords: solar energy forecasting, machine learning algorithms, photovoltaic power generation, feature selection, *k*-Nearest Neighbors (kNN).

Introduction

The development of alternative energy sources has significantly transformed modern power systems, accelerating the transition toward distributed generation. The increasing share of photovoltaic power stations (PPS) in the overall energy balance introduces new challenges for managing electric grids, including autonomous microgrids, as traditional control methods were originally designed for centralized power generation. One of the key aspects of effectively managing distributed energy resources is the real-time balancing of electricity generation and consumption, which is implemented through complex hierarchical control architectures.

Accurate forecasting of solar power generation is critically important for optimizing energy system operation, reducing transmission losses, and improving the economic efficiency of photovoltaic installations. The shift to distributed generation necessitates the development of new forecasting methods to balance energy flows and ensure the stability of the power grid [1, 2].

Photovoltaic systems play a vital role in ensuring a stable energy supply; however, their efficiency is highly dependent on dynamic meteorological conditions [3]. Fluctuations in generation output present additional challenges for automated control systems that are responsible for maintaining an instantaneous energy balance. To enable effective planning, prevent overloads, and avoid energy deficits or surpluses, day-ahead forecasting of solar generation with sufficient accuracy is essential [4]. At the same time, forecasting across various time horizons (from several minutes to several days ahead) remains equally relevant [3].

This forecasting process is complicated by both the stochastic nature of solar energy availability and the inherent limitations of modern predictive models. Typically, forecasting methods rely on weather data derived from public meteorological platforms or specialized APIs (e.g., Solcast, Meteonorm, NASA POWER), or on processed results from local meteorological observations. Among the most critical challenges is the quality of meteorological input data, given

that solar power output depends on multiple interconnected variables such like solar radiation, temperature, humidity, etc. Which exhibits complex stochastic behavior. Additionally, data sources differ in both spatial and temporal resolution.

The object of research is the data acquisition and preprocessing procedures used in forecasting the output power of photovoltaic stations.

The subject of research is the application of correlation analysis methods for identifying the significance of input features, as well as machine learning methods for forecasting solar generation.

The purpose of the research is to enhance the effectiveness of short-term forecasting of PV generation under conditions of stochastic variation in meteorological parameters.

To achieve this purpose, the following research objectives were defined:

Analyze recent studies and publications on photovoltaic generation forecasting.

- Perform exploratory data analysis and preprocessing (e. g., anomaly detection, normalization).
- Conduct correlation analysis of meteorological parameters to assess their significance in generation modeling.
- Justify the application of machine learning methods for forecasting PV operational parameters.
- Develop a software application for automating data acquisition, preprocessing, and short-term power forecasting.
- Evaluate the performance and effectiveness of the developed methods and tools.

Materials and methods of research. The proposed method for forecasting the power output of a photovoltaic power station (PPS) comprises several key stages: data preprocessing, feature selection, model training, and performance evaluation.

Two meteorological datasets were used in this study. The first dataset (hereafter referred to as Dataset I) consists of observations from a meteorological station located within the Radehiv solar power facility in Lviv region, Ukraine. It includes 15-minute interval data collected from April 2024 to January 2025. The dataset contains global horizontal irradiance (GHI), global tilted irradiance (GTI), ambient air temperature, panel temperature, actual power generation, and a generation curtailment flag. The "generation curtailment" parameter indicates whether the output was intentionally limited by a dispatcher or due to technical issues. If the value is "true", those records were excluded from model training.

The second dataset (Dataset II) was obtained from the Solcast service [5], using the geographical coordinates of the solar facility. As indicated in [6], Solcast has been identified as one of the most accurate sources of historical meteorological data for solar forecasting. The dataset includes 23 parameters such as diffuse horizontal irradiance (DHI), direct normal irradiance (DNI), air temperature, humidity, atmospheric pressure, wind speed and direction, cloud cover, and others.

To calculate solar geometric parameters based on the site's coordinates, a custom Python application was developed using the pvlib library. Each dataset was extended with three derived features: solar elevation angle, solar azimuth, and the number of minutes since the start of the day.

One of the initial steps involved exploratory data preprocessing, which included data cleaning and normalization. Time series related to power generation often contain irrelevant values – particularly zero values during nighttime hours – which were excluded from further analysis. Missing data were detected and handled using statistical techniques and machine learning-based imputation methods.

The next phase focused on identifying the most influential meteorological parameters affecting solar power generation. For this purpose, the Boruta feature selection algorithm [7, 8] was used in conjunction with correlation analysis employing Pearson, Spearman, and Kendall coefficients. The forecasting model was built using the *k*-Nearest Neighbors (kNN) machine learning algorithm, applied to the processed meteorological inputs.

As part of the study, a software application was developed to automate the collection, preprocessing (including anomaly detection and normalization), and short-term forecasting of solar power output. The application was implemented in Python using the scikit-learn library for machine learning and pandas for efficient data handling and analysis.

To evaluate model performance, a statistical assessment was carried out using several metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), coefficient of determination (R^2), Power Forecasting Gap (PFG), and the Kolmogorov – Smirnov (KS) test for assessing the agreement between actual and predicted distributions.

Analysis of recent research and publications. Forecasting electricity production from photovoltaic power stations (PPS) is a key challenge in maintaining the stability of power systems with a high penetration of renewable energy sources. Given the inherent variability of meteorological conditions, high forecasting accuracy can only be achieved through effective incorporation of weather-related factors and the use of advanced prediction models.

An essential component of forecasting models is the effective selection of informative input features. In [10], a detailed correlation analysis was conducted between various meteorological and astronomical parameters and solar power generation. The authors employed the Pearson correlation coefficient to quantify the strength of the relationships. The highest correlations were recorded for global horizontal irradiance (GHI) and global tilted irradiance (GTI), with values ranging between 0.86 and 0.91, confirming their critical role in generation modeling. The significance of geometric variables – such as solar azimuth and elevation – was also confirmed, with moderate to high correlation values (0.6–0.8). The study also applied the Boruta feature selection method to automatically identify the most relevant features, which helps to reduce model complexity and prevent overfitting while preserving predictive accuracy.

Recent years have seen growing interest in combining physical models, statistical approaches, and artificial intelligence (AI)-based methods. Studies [11, 12, 13] explored a broad spectrum of solar power forecasting models across different geographic regions and forecasting horizons. These works demonstrated that AI-based models significantly outperform conventional physical and statistical methods in predictive accuracy [14].

Recent publications [10, 15] highlight the increasing application of artificial neural networks (ANNs), including multilayer perceptrons (MLP), recurrent neural networks (RNN), long short-term memory (LSTM), gated recurrent units (GRU), and convolutional neural networks (CNN). As shown in [7], LSTM and GRU models are particularly well-suited for time series forecasting due to their ability to retain long-term dependencies. Meanwhile, hybrid architectures such as CNN-LSTM combine the spatial pattern recognition capabilities of CNNs with the temporal modeling strength of LSTMs, resulting in a 15–20 % improvement in forecasting accuracy compared to individual models.

The study in [10] investigated the performance of machine learning models including XGBoost, kNN, and LSTM, as well as ensemble approaches such as model stacking. The authors observed high correlations between predicted and actual values, particularly for LSTM and XGBoost, confirming their reliability for short- and medium-term solar forecasting. However, other studies [16, 17] indicate that as the forecasting horizon increases, the prediction accuracy of both standalone and hybrid models tends to decline.

The *k*-Nearest Neighbors (kNN) method has been increasingly adopted for solar power forecasting due to its simplicity, flexibility, and effectiveness in regression tasks involving temporal dependencies. As shown in [18], kNN delivers high accuracy in short-term forecasting of solar generation based on dynamic meteorological inputs. The authors compared kNN with models such as MLP and CNN, concluding that despite its conceptual simplicity, kNN achieves comparable accuracy, particularly in cases where data volume is limited or feature structure varies over time. One of the key advantages of kNN is its non-parametric, instance-based nature – it does not require a separate training phase and stores the entire dataset, allowing the model to adapt flexibly to new input conditions without retraining. This is particularly important in forecasting contexts where weather conditions are highly unpredictable.

In conclusion, the literature review supports the feasibility and appropriateness of using the kNN method to construct a baseline model for solar generation forecasting. The results produced by such a model can serve as a reference point for comparative analysis with more complex approaches such as LSTM or XGBoost.

Research results and their discussion

Preliminary data analysis. To identify the most influential parameters affecting electricity generation and to reveal potential nonlinear dependencies among variables, a correlation analysis was performed using the Pearson, Spearman, and Kendall methods [19]. The analysis focused on quantifying the relationship between meteorological parameters and the actual power output of the photovoltaic power station.

Only the parameters with a non-negligible correlation with electricity generation were retained for further analysis. The results for the first dataset (Dataset I) are presented in Tables 1, 2.

Table 1. Correlation analysis results for the first data set

Parameter / Correlation method	Pearson	Spearman	Kendall
GHI	0.91	0.88	0.84
GTI	0.91	0.86	0.84
Panel temperature	0.81	0.71	0.57
Air temperature	0.62	0.59	0.43
Solar elevation angle	0.75	0.82	0.68
Solar azimuth angle	–0.72	–0.77	–0.62

Table 2. Correlation analysis results for the second data set

Parameter / Correlation method	Pearson	Spearman	Kendall
GTI	0.9	0.92	0.86
GHI	0.89	0.91	0.85
Clearsky GTI	0.83	0.89	0.8
Clearsky GHI	0.82	0.9	0.8
Clearsky Direct Normal Irradiance (DNI)	0.78	0.89	0.8
Direct normal irradiance (DNI)	0.78	0.75	0.64
Solar elevation angle	0.75	0.86	0.72
Clearsky Diffuse Horizontal Irradiance (DHI)	0.67	0.86	0.71
Diffuse Horizontal Irradiance (DHI)	0.57	0.86	0.71
Air temperature	0.44	0.46	0.33
Relative humidity	–0.56	–0.53	–0.38
Solar azimuth angle	–0.72	–0.82	–0.66

Based on the obtained results, the most influential parameters for forecasting solar power generation are related to solar irradiance intensity, particularly Global Horizontal Irradiance (GHI) and Global Tilted Irradiance (GTI). The correlation coefficients for these variables range from 0.85 to 0.91, depending on the method used, confirming their dominant role in generation modeling.The second most significant group comprises temporal and geometric features, including time of day, solar azimuth angle, and solar elevation angle. These parameters exhibit moderate to strong correlations (between 0.5 and 0.75) and have a substantial impact on the photovoltaic output due to their role in determining the incident angle and intensity of sunlight on the panel surface. The third group includes air temperature and relative humidity, which, while less influential individually, contribute to the refinement of prediction accuracy when incorporated into the model. Their inclusion helps capture secondary environmental effects on system performance, especially under variable atmospheric conditions.

To justify the selection of the most significant variables, the Boruta algorithm [7] was employed. Boruta is a feature selection method built upon the Random Forest classifier and is particularly robust to multicollinearity. It also accounts for complex interactions between features, thereby providing a more reliable assessment of variable importance. The results of the Boruta-based feature importance analysis are presented in Figs. 1 and 2, where the vertical axis represents the relative importance of each input variable.

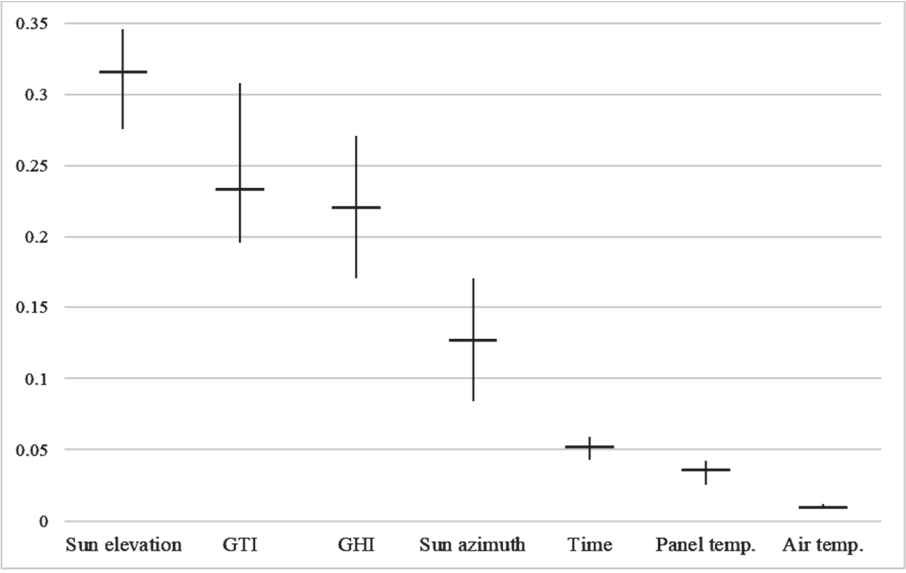

Fig. 1. Importance of features for the first dataset

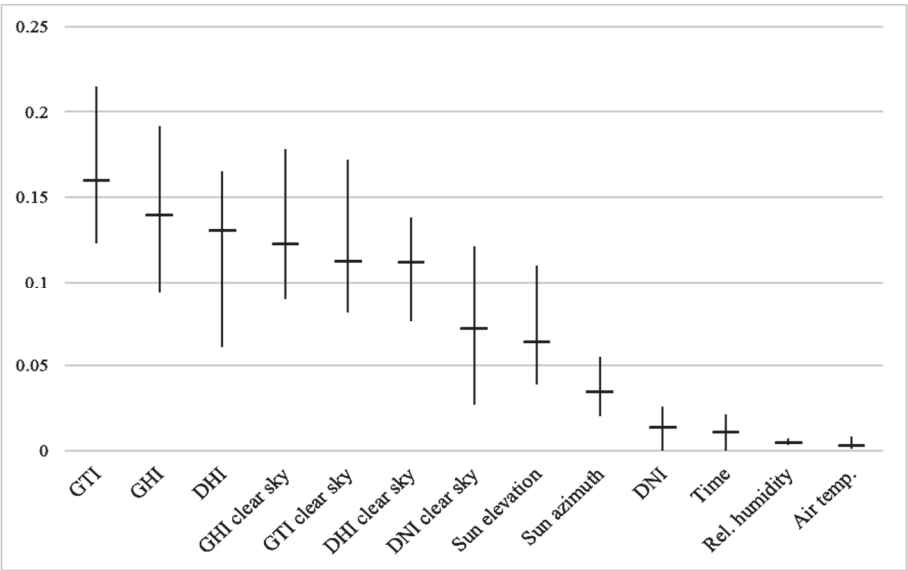

Fig. 2. Importance of features for the second dataset

Based on the obtained results, it can be concluded that parameters such as GHI, GTI, air temperature, solar elevation, and solar azimuth exert the most significant influence on electricity generation.

In Dataset II, despite containing a larger number of meteorological features, several variables were identified as low-importance and can be disregarded during the training of an optimized forecasting model, thereby simplifying the model without compromising its accuracy.

To assess the discrepancies between analogous parameters from the two meteorological data sources, a statistical comparison of GHI and GTI values was performed using cumulative distribution curves (Fig. 3). The vertical axis represents the cumulative relative frequency, indicating the proportion of observations that are less than or equal to the corresponding value on the horizontal axis.

The plot reveals a noticeable asymmetry in the distribution, with a concentration of GHI values in the mid-range between 100 and 750 W/m², which corresponds to active daytime periods under moderate to high solar radiation conditions. In contrast, values below 100 W/m² (typically associated with heavy cloud cover or twilight) and above 850 W/m² (indicating peak irradiance) are significantly less frequent.

Meanwhile, the GTI distribution exhibits a wider range, spanning from 0 to 1000 W/m², suggesting greater variability likely due to panel tilt and orientation effects (Fig. 3).

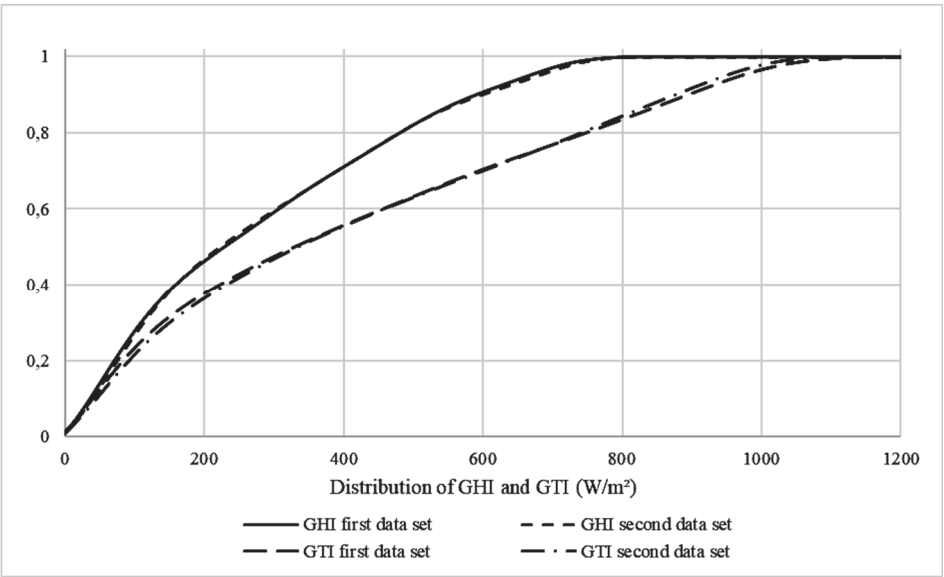

Fig. 3. Comparison of the GHI and GTI distribution for the first and second datasets

A frequency distribution of the differences between the GHI and GTI values from the two datasets was constructed (Fig. 4). The resulting histogram indicates that deviations in GHI exceeding ±150 W/m² occur in less than 0.5 % of cases (approximately 50 observations). The average error is skewed toward negative values, suggesting that GHI values in one of the datasets tend to be systematically underestimated compared to the other.

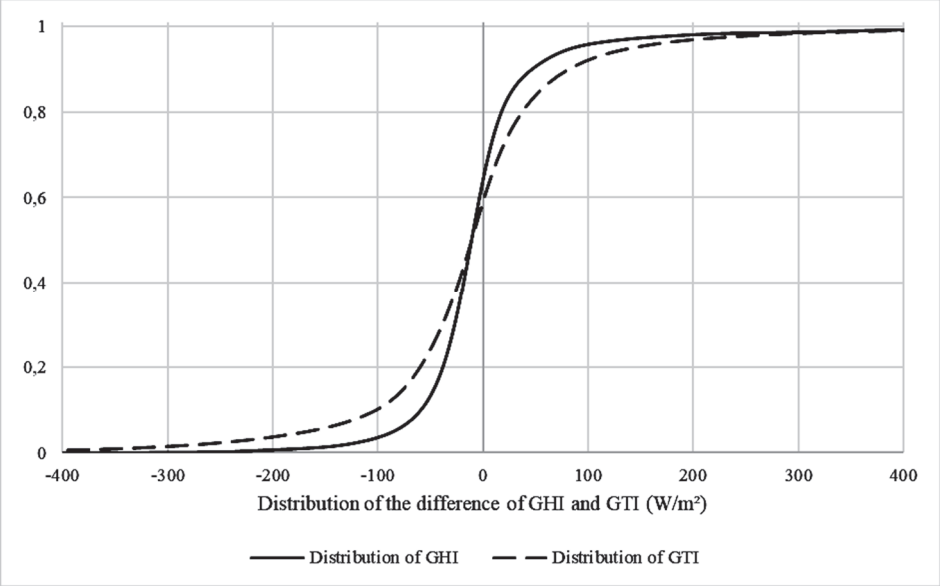

Fig. 4. Distribution of the GHI and GTI difference between the first and second datasets

The range of GTI deviations is slightly broader; however, differences greater than ±250 W/m² are observed in fewer than 4.4 % of cases (around 440 observations), indicating generally acceptable consistency between the datasets, despite localized discrepancies.

In contrast to the relatively similar frequency distributions of GHI and GTI shown in Fig. 3, the air temperature distributions for the two datasets (Fig. 5) revealed notable differences. While the frequency of temperature values in the range of [−2 °C to +16 °C] is nearly identical for both datasets, the range of [+16 °C to +42 °C] exhibits significant divergence.

Dataset I demonstrates greater temperature variability, with a higher occurrence of extreme values, particularly those exceeding 30 °C. These differences are likely attributed to the interpolation methods used in Dataset II, which may smooth out local fluctuations and reduce the representation of higher temperatures.

Data preparation and preprocessing for modeling. Data preprocessing is a critical step in constructing accurate forecasting models for photovoltaic power generation. At this stage, missing or irrelevant data values were eliminated, and the effectiveness of various normalization techniques was assessed.

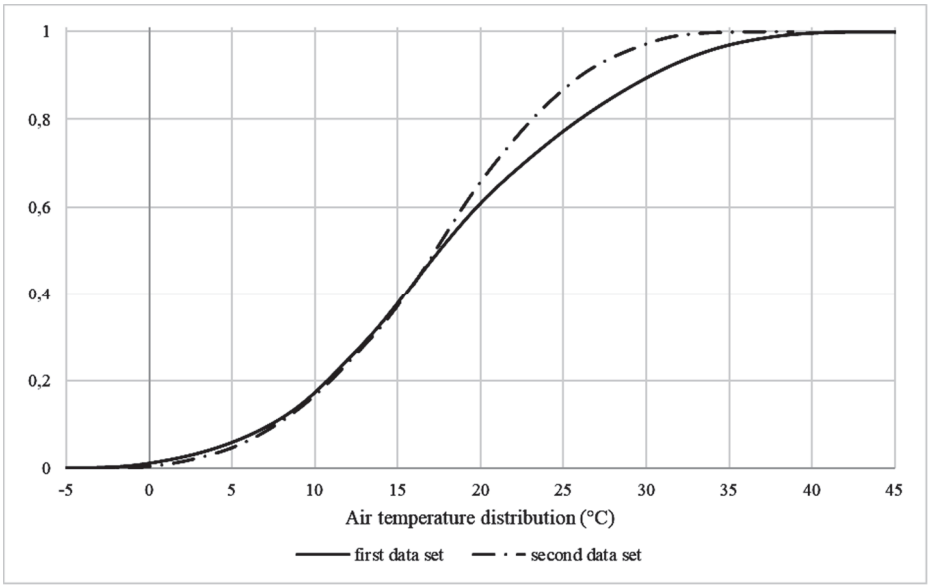

Fig. 5. Comparison of air temperature distribution for the first and second datasets

To ensure model consistency, all records with missing values were excluded from the dataset. These gaps were typically caused by technical failures of meteorological stations or data transmission errors. In addition, all records corresponding to post-sunset time intervals were removed from the analysis, as such data are not relevant for solar power prediction.

To reduce the impact of scale differences among features, several data normalization approaches were tested:

- Min-Max normalization – scales the data to a [0,1] range.
- Z-score transformation – standardizes the data to have a mean of 0 and a standard deviation of 1, using the formula:

$$X_{scaled} = \frac{X - \mu}{\sigma},$$

where X – is the original value; μ – is the sample mean, σ – is the standard deviation.

- Logarithmic transformation (Log) – reduces the impact of extreme fluctuations. Since the logarithm is undefined for negative values, a constant was added to parameters containing negative values before transformation.

However, the forecasting results indicated that data normalization did not improve prediction accuracy (Table 3). Specifically, for Dataset I, the best results were achieved using non-normalized input values, with an R^2 score of 0.9912. In contrast, applying Min-Max, logarithmic, and *Z*-score transformations led to a notable decline in accuracy, with *Z*-score standardization yielding the weakest result (R^2=0.9742).

A similar trend was observed for Dataset II, where the highest forecasting accuracy was also attained without applying any normalization (R^2=0.9489).

These findings suggest that, for the k-Nearest Neighbors (kNN) model in the context of solar power forecasting, preliminary data normalization is not essential and may, in fact, degrade model performance. The most effective results were obtained when using the original, non-normalized feature values from Dataset I.

Table 3. The impact of normalization on the results of forecasting solar power generation

Dataset I			
	MSE	MAE	R^2
No normalization	46621	97.4	0.9912
Min-Max	66497	127.7	0.9875
Z-score	136907	182.3	0.9742
Log	106252	141.3	0.9800
Dataset II			
No normalization	269322	326.9	0.9489
Min-Max	296428	336.7	0.9438
Z-score	286854	325.0	0.9456
Log	331139	361.7	0.9372

kNN model training. The kNN model was trained separately for each of the two considered datasets. An identical set of input features was selected from both datasets for model training, including Global Horizontal Irradiance (GHI), Global Tilted Irradiance (GTI), air temperature, solar elevation, solar azimuth, panel tilt angle, and time. The time feature was normalized and expressed as the number of minutes elapsed since the beginning of the day.

Each dataset was split into a training set (90 %) and a testing set (10 %). During training, the number of neighbors (n_neighbors) was set to 5, and Euclidean distance was used as the distance metric. This configuration follows both common practice in regression tasks and empirical findings. The choice of n_neighbors = 5 provides a practical trade-off between local smoothing and robustness to noise, helping the model avoid overfitting to anomalous values.

Previous studies [20, 21] have also applied the same value for n_neighbors in photovoltaic power forecasting tasks and reported consistent performance. Model evaluation was conducted on the test set, and the results are presented in Table 4.

Table 4. Results of model performance evaluation

Metric	datasets	
	Dataset I	Dataset II
MSE	46621	269322
MAE	97.4	326.9
R^2	0.9912	0.9489
PFG, %	4.1	12.9
KS (*p*-value)	0.9998	0.8532
KS (D-statistic)	0.0338	0.0619

The presented results highlight the advantages of using meteorological data obtained directly from the on-site weather station of the photovoltaic facility (Dataset I) for forecasting power generation.

Discussion of research results. A comparative analysis was carried out to evaluate the accuracy of photovoltaic power generation forecasting using two models trained on two distinct meteorological datasets obtained from different sources. The evaluation results, presented in Table 5, provide a comprehensive view of the forecasting performance and reveal noticeable differences in prediction accuracy. These findings are supported both by numerical metrics and by subsequent visualizations of the model outputs.

The model trained on Dataset I demonstrated superior predictive performance, as evidenced by lower values of Mean Squared Error (MSE) and Mean Absolute Error (MAE). In contrast, the corresponding error metrics for Dataset II were notably higher, indicating lower forecast accuracy. Additionally, the Power Forecasting Gap (PFG) was higher for the second dataset (13.2 % vs. 14.2 %), further confirming the discrepancy in model performance.

Table. 5. Results of the assessment of the efficiency of generation forecasting

Metric	datasets	
	Dataset I	Dataset II
MSE	209271	270469
MAE	311.5	336.2
R^2	0.9582	0.9459
PFG, %	13.2	14.2
KS (*p*-value)	0.0335	0.0126
KS (*D*-statistic)	0.0771	0.0859

Nevertheless, the coefficient of determination (R^2) remained high for both models – 0.9582 for Dataset I and 0.9459 for Dataset II – though it should be noted that this metric is sensitive to the range of predicted values and may not fully capture forecasting precision. The decrease in accuracy observed across other metrics supports the hypothesis that Dataset II may contain less relevant or noisier data.

Overall, the results validate the effectiveness of the applied methods for feature selection, correlation analysis, and data preprocessing. In combination with the kNN algorithm, they enabled the development of a reliable short-term forecasting system for photovoltaic power generation.

To enhance the interpretability of model performance under varying solar conditions, the forecasted and actual power generation profiles were visualized separately for periods with high and low solar irradiance (Fig. 6 and 7). This approach allows for a clearer understanding of model behavior in response to different meteorological scenarios.

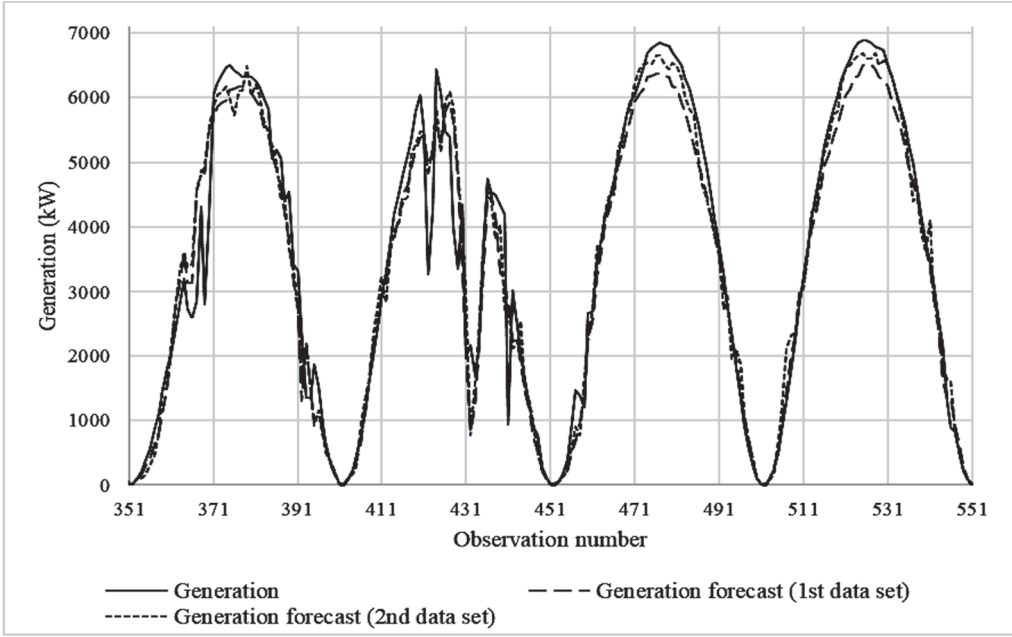

Fig. 6. Actual and forecasted solar electricity generation during periods of high insolation

During sunny days, the model trained on Dataset I more accurately replicates the daily generation profile, effectively capturing both morning ramp-ups and peak outputs (Fig. 6). This outcome confirms the model's ability to utilize stable correlations between generation and key meteorological parameters under favorable conditions.

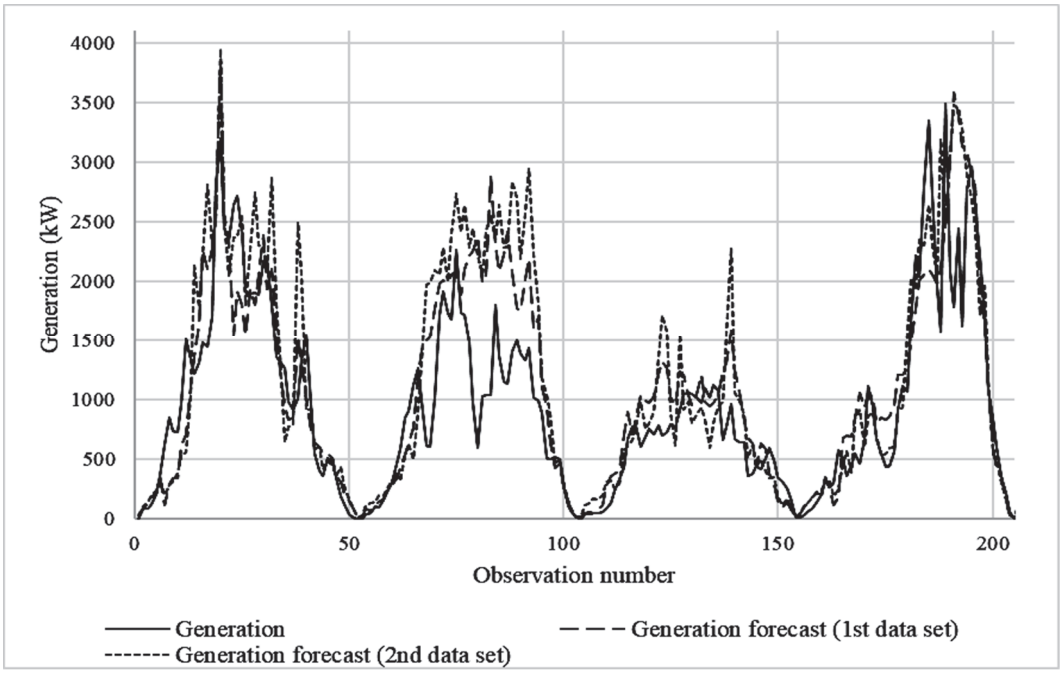

Fig. 7. Actual and forecasted solar electricity generation during periods of low insolation

In contrast, during low-irradiance periods (Fig. 7), a decrease in forecasting accuracy is observed for both models. This is a common challenge caused by the stochastic impact of cloud cover. The deviations are particularly pronounced in the model based on Dataset II, highlighting the importance of selecting high-quality predictive features – especially those that account for rapid weather fluctuations.

To visually compare the forecasting accuracy obtained using two different meteorological datasets, a scatter plot was constructed (Fig. 8), showing the relationship between the predicted and actual solar power generation values for both datasets. Each point on the graph represents a 15-minute observation, with the actual generation plotted along the *x*-axis and the predicted value along the *y*-axis.

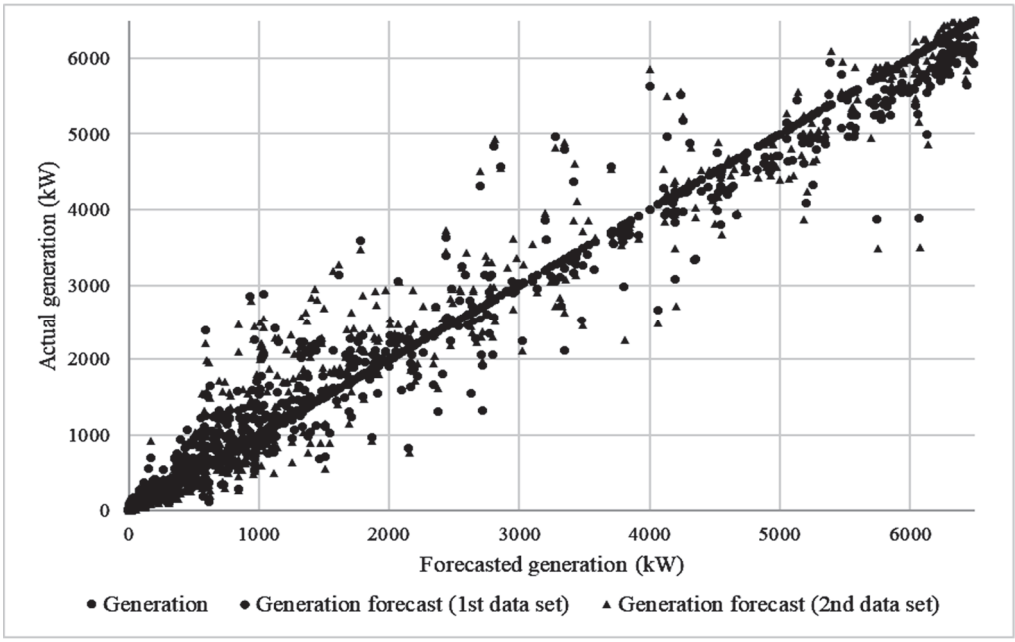

Fig. 8. Variance of predicted solar power generation values relative to actual data for two datasets

Data points concentrated along the line of perfect prediction (diagonal $y=x$) indicate a high degree of accuracy. The plot clearly shows that predictions generated from Dataset I are more closely clustered around this diagonal, reflecting a stronger correspondence with actual values. In contrast, the predictions based on Dataset II exhibit greater dispersion, indicating higher variability and lower accuracy in reproducing real power outputs.

This visualization complements the preceding quantitative error analysis and illustrates that the model trained on Dataset I is not only more accurate on average, but also more consistent at the level of individual observations.

In addition to numerical metrics, the distribution density of absolute forecasting errors was analyzed (Fig. 9) to visually assess the concentration of deviations between predicted and actual power generation. The resulting plot shows that the model built on Dataset I exhibits a narrower peak centered near zero error, indicating greater prediction stability and lower deviation levels.

By contrast, the model trained on Dataset II produces a broader and more dispersed error distribution, reflecting greater variability and more frequent large deviations. This aspect is particularly important in practical applications, as substantial forecasting errors can lead to imbalances in the power system.

These findings further emphasize the importance of high-quality feature selection and clean input data in photovoltaic power forecasting tasks. Moreover, the results of this study allow for the formulation of its scientific novelty and practical relevance.

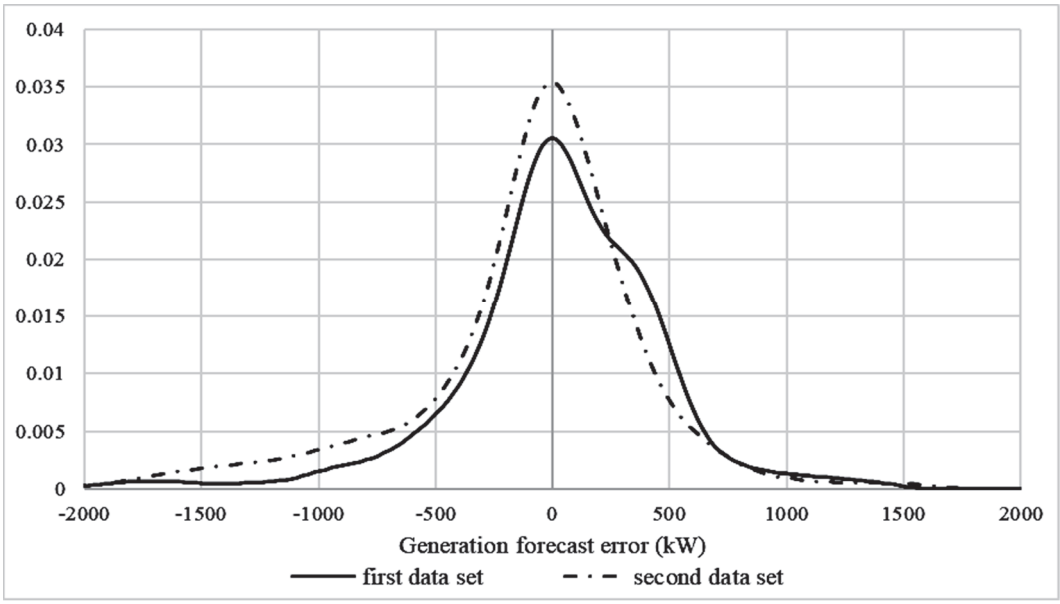

Fig. 9. Density distribution of solar power generation forecast error for two independent datasets

The scientific novelty of the research results is a novel method for forecasting the power output of photovoltaic systems was developed based on the *k*-Nearest Neighbors (kNN) model, with an enhanced approach to feature selection and data preprocessing, resulting in a prediction accuracy of R^2=0.96

The study further advanced the methodology for identifying significant input features in photovoltaic power forecasting by integrating correlation analysis techniques (Pearson, Spearman, and Kendall) with the Boruta algorithm. This combined approach enabled a reduction in the number of input parameters while maintaining high forecasting accuracy.

A comparative statistical analysis of meteorological parameters from different data sources for the same PV site location was also conducted, providing insight into their variability and influence on model performance.

*The practical significance of the research results is t*he proposed approach enables enhanced accuracy in short-term forecasting of photovoltaic (PV) power generation, which is critically important for the stable operation of distributed energy systems and for maintaining the balance between energy production and consumption.

The kNN model with a basic configuration (n_neighbors = 5) demonstrated high predictive accuracy (R^2=0.9582) when trained on a well-preprocessed meteorological dataset. Its simplicity and computational efficiency – achieved without the need for complex neural network architectures – make it an attractive solution for practical deployment, especially in environments with limited computational resources.

The findings can be integrated into energy management systems for estimating next-day PV output and can be adapted to different geographical locations with minimal model adjustment.

The results have practical value for PV system operators, SCADA system developers, and weather service providers involved in building analytical modules for power forecasting and decision support.

Conclusions

This study addressed the challenge of improving the accuracy of solar power generation forecasting by conducting a comparative analysis of two meteorological data sources and identifying the most relevant parameters influencing forecasting performance. A customized data preprocessing and kNN-based modeling approach was developed, allowing for the identification and quantification of the influence of key meteorological variables – including GHI, GTI, solar azimuth, solar elevation, and time – through the application of the Boruta algorithm and correlation analysis (Pearson, Spearman, and Kendall).

In the case of the kNN model, higher forecast accuracy was achieved without prior normalization of the input

features (R^2=0.96), confirming the effectiveness of the selected feature set. The study demonstrated the critical importance of solar irradiance parameters (GHI, GTI) along with temporal and geometric features (solar azimuth, elevation, time), which consistently emerged as significant across both datasets and align with findings from previous research [9].

The results confirm the effectiveness of kNN as a simple yet robust method for short-term solar power forecasting, provided that relevant features are carefully selected and input data is appropriately processed.

References

[1] Pelland, S., Remund, J., Kleissl, J., Oozeki, T., & De Brabandere, K. (2013). *Photovoltaic and solar forecasting: State of the art* (Report IEA-PVPST14-01:2013). IEA PVPS. ISBN 978-3-906042-13-8.

[2] Tsmots, I. G., Tesliuk, V. M., Podolsky, M. R., & Dubuk, V. I. (2020). Tools of visualization of power balances and analytical support of energy efficiency management of region. *Ukrainian Journal of Information Technology*, 2(1), 01–07. https://doi.org/10.23939/ujit2020.02.001

[3] Kuznetsov, M. P., & Lysenko, O. V. (2017). Capabilities of short-term solar energy forecasting. *Renewable Energy*, 2017(1), 25–32. http://nbuv.gov.ua/UJRN/vien_2017_1_6

[4] Matushkin, D. S., Bosak, A. V., & Kulakovskyi, L. Y. (2020). Analysis of factors for predicting electricity generation by solar power plants. *Energetics: Economics, Technology, Ecology,* 4 (2020), 62. https://doi.org/10.20535/1813-5420. 4.2020.233597

[5] Solcast Toolkit (n. d.). Retrieved April 2025, from https://toolkit.solcast.com.au/

[6] Ashan, D. K., & Geekiayange, V. (2021). Reliability comparison of weather data of PVGIS, NREL and Solcast for PV solar energy generation forecasting. Unpublished report. https://doi.org/10.13140/RG.2.2.14453.50400

[7] Kursa, Miron & Rudnicki, Witold. (2010). Feature Selection with Boruta Package. *Journal of Statistical Software*, 36, 1–13. 10.18637/jss.v036.i11.

[8] Alresheedi, A., A., & Al-Hagery, M., A. (2020). Hybrid artificial neural networks with Boruta algorithm for prediction of global solar radiation: Case study in Saudi Arabia. *International Journal of Computer Science and Network*, 9(2), (April). ISSN (Online): 2277-5420.

[9] Rinchi, B., Ayadi, O., Al-Dahidi, S., & Dababseh, R. (2024). A universal tool for estimating monthly solar radiation on tilted surfaces from horizontal measurements: A machine learning approach. *Energy Conversion and Management*, 314, 118703. https://doi.org/10.1016/j.enconman.2024.118703

[10] Shakhovska, N., Medykovskyi, M., Gurbych, O., Mamchur, M., & Melnyk, M. (2024). Enhancing solar energy production forecasting using advanced machine learning and deep learning techniques: A comprehensive study on the impact of meteorological data. *Computers, Materials & Continua*, 81(2), 3147–3163. https://doi.org/10.32604/cmc.2024.056542

[11] Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124, 109792. https://doi.org/ 10.1016/j.rser.2020.109792

[12] Meenal, R., Binu, D., Ramya, K. C., Michael, P. A., Kumar, K. V., Rajasekaran, E., & Sangeetha, B. (2022). Weather forecasting for renewable energy system: A review. *Archives of Computational Methods in Engineering*, 29(5), 2875–2891. https://doi.org/10.1007/s11831-021-09695-3

[13] Zhuravel, I. M., Onyshko, V. R., Zhuravel, Yu. I., & Ambroziak, K. A. (2024). Quantitative assessment of the visual quality of digital images based on the laws of human visual perception. *Ukrainian Journal of Information Tecnology*, 6(1), 17–25. https://doi.org/10.23939/ujit2024.01.017

[14] Li, B., & Zhang, J. (2020). A review on the integration of probabilistic solar forecasting in power systems. *Solar Energy*, 210, 68–86. https://doi.org/10.1016/j.solener.2020.07.066

[15] Asghar, R., Fulginei, F., R., Quercio, M., & Mahrouch, A. (2024). Artificial neural networks for photovoltaic power forecasting: A review of five promising models. *IEEE Access*, 12, 90461–90485. https://doi.org/10.1109/ACCESS.2024.3420693

[16] Zhou, H., Liu, Q., Yan, K., & Du, Y. (2021). Deep learning enhanced solar energy forecasting with AI-driven IoT. *Wireless Communications and Mobile Computing*, 2021, 1–11. https://doi.org/10.1155/2021/9249387

[17] Li, G., Xie, S., Wang, B., Xin, J., Li, Y., & Du, S. (2020). Photovoltaic power forecasting with a hybrid deep learning approach. *IEEE Access*, 8, 175871–175880. https://doi.org/10.1109/ACCESS.2020.3025860

[18] Obileke, K. I. K. (2024). Short-term forecasting of photovoltaic power using multilayer perceptron neural network, convolutional neural network, and k-nearest neighbors' algorithms. *Optics*, 5(2), 293–309. https://doi.org/10.3390/opt5020021

[19] Perehuda, O. V., Kapustian, O. A., & Kuryilko, O. B. (2022). Statistical data processing: Educational manual. Electronic edition, 103 p. Retrieved from [https://www.mechmat.univ.kiev.ua/wp-content/uploads/2022/02/navch_pos_perehuda.pdf]

[20] Ramli, Nor Azuana & Abdul Hamid, Mohd Fairuz & Azhan, Nurul Hanis. (2019). Solar Power Generation Prediction by using k-Nearest Neighbor Method. AIP Conference Proceedings, 2129. 10.1063/1.5118124.

[21] Mas'ud, Abdullahi. (2021). Comparison of three machine learning models for the prediction of hourly PV output power in Saudi Arabia. *Ain Shams Engineering Journal*, 13, 101648. 10.1016/j.asej.2021.11.017.

Р. В. Мельник, М. В. Мельник

Національний університет "Львівська політехніка", Львів, Україна

МЕТОД ПРОГНОЗУВАННЯ ОБСЯГІВ ГЕНЕРАЦІЇ ЕНЕРГІЇ СОНЯЧНОЮ ЕЛЕКТРОСТАНЦІЄЮ

Успішне використання сонячної енергетики зумовлює необхідність точного прогнозування виробництва електроенергії сонячними електростанціями (СЕС) для стабільного функціонування систем електропостачання. Це пов'язано з необхідністю підтримання миттєвого балансу виробництва і споживання електричної енергії, який забезпечується реалізацією складних ієрархічних систем управління наявними джерелами енергії. Особливо актуальна можливість короткочасного прогнозування виробництва енергії СЕС. У статті наведено результати розроблення методів прогнозування та дослідження їх ефективності, а також вплив структури і якості метеоданих на результати прогнозування. Для оцінювання важливості метеорологічних параметрів використано статистичні методи кореляцій Пірсона, Спірмена та Кендала та метод Борута (Boruta). На основі визначених значущих змінних побудовано модель з використанням методу *k*-найближчих сусідів (kNN). Дослідження виконано для двох незалежних наборів метеорологічних даних, що містять інформацію про зміну параметрів навколишнього середовища та фактичну генерацію енергії СЕС. Перший набір складається із даних, отриманих від метеостанції об'єкта генерації. Другий – з даних відкритого джерела для географічних координат об'єкта генерації. Одержані результати свідчать, що використання метеоданих із метеорологічної станції об'єкта генерації та їх опрацювання розробленим методом дає змогу підвищити точність навчання моделі kNN. Значення коефіцієнта детермінації (R^2) для першого і другого наборів даних однакові – 0,99 і 0,95. А це, відповідно, підвищує точність прогнозування потужності генерації СЕС, обґрунтовує доцільність введення до структури об'єктів генерації сучасних засобів вимірювання та реєстрації метеорологічних параметрів.

Ключові слова: прогнозування, машинне навчання, сонячна електростанція, аналіз кореляції, метод *k*-найближчих сусідів.

Інформація про авторів:

Мельник Роман Володимирович, д-р філософії (Ph. D.), асистент, кафедра автоматизованих систем управління. **Email:** roman.v.melnyk@lpnu.ua; https://orcid.org/0000-0003-0619-1613

Мельник Михайло Васильович, аспірант, кафедра автоматизованих систем управління. **Email:** mykhailo.v.melnyk@lpnu.ua; https://orcid.org/0000-0003-0339-3711

Цитування за ДСТУ: Мельник Р. В., Мельник М. В. Метод прогнозування обсягів генерації енергії сонячною електростанцією. *Український журнал інформаційних технологій*. 2025, т. 7, № 1. С. 149–159.

Citation APA: Melnyk, R. V., & Melnyk, M. V. (2025). A method for forecasting the energy generation of a solar power plant. *Ukrainian Journal of Information Technology*, 7(1), 149–159. https://doi.org/10.23939/ujit2025.01.149