EPJ Photovoltaics **15**, 12 (2024)
© S. Theocharides et al., Published by EDP Sciences, 2024
https://doi.org/10.1051/epjpv/2024011

Available online at:
www.epj-pv.org

Special Issue on 'EU PVSEC 2023: State of the Art and Developments in Photovoltaics', edited by Robert Kenny and João Serra

Original Article — Open Access

PV generation forecasting utilizing a classification-only approach

Spyros Theocharides*, George Makrides and George E. Georghiou

PV Technology Laboratory, FOSS Research Centre for Sustainable Energy, Department of Electrical and Computer Engineering, University of Cyprus, Nicosia 1678, Cyprus

Received: 30 June 2023 / Accepted: 1 February 2024

Abstract. The increasing use of photovoltaic (PV) systems in electricity infrastructure poses new reliability challenges, as the supply of solar energy is primarily dependent on weather conditions. Consequently, to mitigate the issue, enhanced day-ahead PV production forecasts can be obtained by employing advanced machine learning techniques and reducing the uncertainty of solar irradiance predictions through statistical processing. The objective of this study was to present a methodology for accurately forecasting day-ahead PV production using novel machine learning techniques and a classification-only forecasting approach. Specifically, the central component of the proposed method is a classifier model based on an Extreme Gradient Boosting (XGBoost) ensemble algorithm that classifies the respective daily 30-min profiles of the forecasted global horizontal irradiance (GHI), the measured incident irradiance (G_i), and the AC power (P_{AC}) into a predetermined number of classes. The formed classifier model was used as a dictionary to designate the newly arrived forecasted GHI to a particular class and ultimately identify the corresponding forecasted PAC. The results demonstrated that the proposed forecasting solution provided forecasts with a daily normalised root mean square error (nRMSE) of 8.20% and a mean absolute percentage error (MAPE) of 6.91% over the test set period of one year, while the model's reproducibility was also evaluated and confirmed. Additionally, a comprehensive evaluation based on clear-sky index categories revealed that the model's performance was notably accurate on clear-sky days, while maintaining acceptable accuracy levels on moderate and overcast days. These findings underscore the versatility and robustness of the proposed methodology in handling diverse weather conditions and hold promise for improved PV production forecasts.

Keywords: Classification / forecasting / machine learning / photovoltaic / performance

1 Introduction

As renewable energy sources incline globally, photovoltaic (PV) technologies are emerging as the primary solution to meet rising electricity demand [1]. The production capacity of renewable energy is projected to increase by 50% by 2024, with photovoltaic systems accounting for 60% of this increase [2]. However, the integration of PV systems into the distribution system presents unique challenges that necessitate flexible power system options to ensure reliable service during rapid supply and demand fluctuations [3]. Concerns for power system stability are raised by the variability and low predictability of PV generation, as grid operators must account for the intermittent character of solar-generated electricity in generation planning and dispatch operations [3–5]. As the prevalence of distributed PV systems increases, utility grids are undertaking a transition to modern, digitally-enhanced technologies that enable the monitoring and control of distributed energy resources (DERs) to actively integrate intermittent renewable generation into network planning models and optimisation procedures. The ongoing electrification and decentralisation trends in the power sector are spurring the adoption of innovative digital tools to increase system flexibility and accommodate high penetration rates of renewable energy [6].

The forecasting of PV power generation leverages diverse methodologies, tailored to the data at hand, specific application requirements, and the prediction time frame. Intraday forecasting is indispensable for effectively managing power ramping, voltage flicker forecasts, control operations, and dispatch. Mid-term forecasting, encompassing predictions for the current day and the day ahead, aids in monitoring load consumption and production, along with regulating voltage and frequency.

In the beginning, the process of PV production forecasting was centred on the construction of best-performing-empirical models, which required a precise understanding of the features and behaviour of the system

* e-mail: theocharidis.spyros@ucy.ac.cy

[7–9]. However, due to the intricate structure of these systems, providing an accurate forecast of the amount of electricity that will be generated by PV systems has proven to be difficult. The focus of current research has changed towards predicting methodologies that are more advanced and adaptable, making use of data-driven approaches that are based on machine learning algorithms [10–14].

Recent research has demonstrated that available PV production forecasts from third-party organisations generally rely on measured resources such as weather, PV system, satellite, and sky imagery data, in addition to numerical weather prediction (NWP) models, which are primarily utilised for weather forecasting [15]. While measured data is more important for intraday forecasts (up to 6 hours), NWP models are typically utilised for projections that are longer than 6 hours in the future. Additionally, the accuracy of the forecast differs based on the area, which opens up chances for both new and established vendors. Studies have shown that correct forecasts may be obtained with low percentages of root mean square error (e.g., as low as 6%) under clear sky conditions; however, forecasts for other conditions reveal larger RMSE values ranging from 20% to 80% [16,17].

For the purpose of PV power forecasting, a great number of machine learning models have been constructed utilising a variety of methodologies. Models based on Multilayer Perceptron (MLPNN), Radial Basis Function Neural Networks (RBFNN), and Recurrent Neural Networks (RNN) have been constructed to provide site day-ahead power forecasts [18]. For intraday forecasting, several models have utilised methods such as nonlinear autoregressive exogenous neural networks (NARX), grey model (GM) linked with multilayer perceptron neural networks (MLPNN), and adaptive feed-forward back-propagation neural networks (AFFNN) [14–17]. Research has been conducted to test the accuracy of several strategies for regional forecasting. These techniques include Bagging, Random Forest, Boosting, Support Vector Machines (SVM), and Generalised Additive Models (GAM). The results of this research have shown that Random Forest is the model that performs the best [18]. For regional forecasting, Support Vector Regression (SVR) pre-processing, in conjunction with Principal Component Analysis (PCA), has been investigated [19].

Several researchers have used the same data during the training phase of their PV power forecasting models to combine weather categorization alongside machine learning techniques to increase the accuracy of their models [20]. There have also been presentations of joint models that combine the most advantageous aspects of physical models and artificial neural networks (ANNs) [20,21]. In other methods, self-organizing maps have been employed for weather classification, and the training stage of artificial neural networks (ANNs) has been improved to achieve more accurate forecasts with a smaller mean absolute percentage error (MAPE) [14]. To increase the accuracy of weather forecasts by locating commonalities across different data sets, Support Vector Machine (SVM) models that are based on daily weather classification have been created [22]. There are several studies that provide comprehensive overviews of the strategies used to anticipate PV power [23,24]. Additionally, an in-depth study for the deep learning neural network (DLNN) models were conducted by [25], explored the effectiveness of different DLNN architectures and stressed their importance.

While the results from the studies mentioned are promising, they have yet to be extensively compared with large-scale field data to ensure reliable weather forecasts that are unaffected by technological advancements and the specific climatic conditions of a region. Enhancing the accuracy of PV power forecasting necessitates the adoption of wholly data-driven methodologies capable of understanding system behaviour without the need for extensive system attributes. This is particularly critical for decentralized rooftop installations like behind-the-meter PV systems, where comprehensive metadata about the system might not always be easily accessible. In such scenarios, leveraging recent operational datasets can shed light on system behaviour, facilitating precise estimates of PV production. Furthermore, honing strategies to boost the accuracy of existing photovoltaic production forecasting models is pivotal for crafting robust and location-independent forecasting models.

This work aimed to present a methodology for accurately performing day-ahead PV production forecasting models that leverage novel machine learning techniques based on an unsupervised classification-only forecasting approach. PV generation forecasting using an unsupervised classification-only approach means using machine learning algorithms to classify a PV system's future power output into pre-defined categories instead of predicting the exact power output. This approach is usually used when the target variable has a limited range of possible values and is suitable for day-ahead forecasting, where the amount of uncertainty in the prediction is high. However, the usage of enhanced machine learning methods enables the specific methodology to provide accurate forecasts up to 24 hours ahead. In particular, the core element of the proposed method is a classifier model based on an Extreme Gradient Boosting (XGBoost) ensemble algorithm that classifies the respective daily 30-minute profiles of the forecasted global horizontal irradiance (GHI), the measured incident irradiance (G_{i}) and the AC power (P_{AC}) into a specific number of classes. The formed classifier model will be used as a dictionary to assign the newly arrived forecasted GHI into a specific class and eventually to identify the respective forecasted PAC. The model was evaluated against datasets acquired from a test site located at the University of Cyprus (UCY) and an 1 MW PV power plant in Nicosia, Cyprus.

2 Methodology

The objective of our research is to enhance the PV production forecasting through a classification-only approach, aiming to improve accuracy and reliability. The methodology proposes the development of a data-driven power output classifier model, harnessing the classification capabilities of the XGBoost ensemble algorithm [26]. The proposed model is trained on historical datasets acquired from a test-bench PV system and a

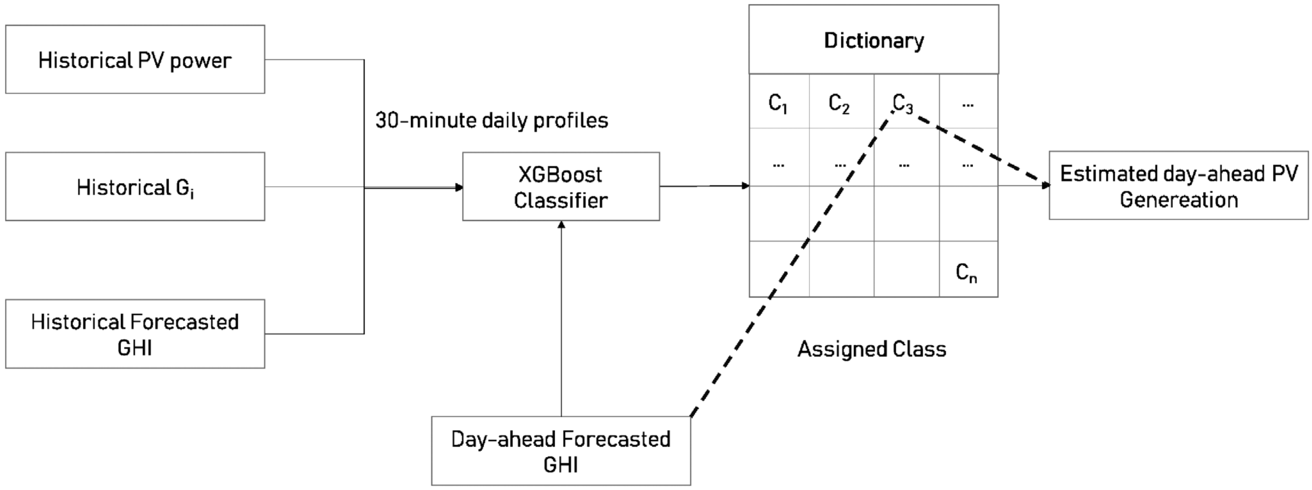

Fig. 1. Graphical representation of the proposed methodology.

utility scale system located in University of Cyprus and in Nicosia, Cyprus respectively, with a training dataset spanning 2 years (1-year for training and 1-year for testing).

The proposed methodology categorizes daily 30-min profiles based on forecasted, GHI, G_i and P_AC into distinct classes. Functioning as a dictionary, it assigns newly arrived day-ahead forecasted GHI to specific classes and predicts the corresponding forecasted P_AC. The proposed methodology internally associates the measured G_i and P_AC, facilitating the construction of predefined classes and, ultimately, the assignment of the appropriate class to the input parameter, which is the forecasted GHI. This classification-based forecasting approach enables the proposed model to deliver precise PV power generation forecasts by interpreting weather and irradiance conditions into specific classes with predefined 30-minute intervals. In addition, the dictionary serves as a lookup table enabling the model to find the representative class for a given forecasted GHI. By associating each class with an irradiance conditions, the dictionary aids in classifying the input parameter accurately.

The evaluation procedure utilizes predefined metrics, including the normalized root mean square error (nRMSE) and mean absolute percentage error (MAPE), to assess the accuracy and dependability of the forecasts. By comparing the forecasted PV power output to actual measurements, we determine the model's accuracy. Additionally, a comprehensive analysis was conducted to evaluate the model's performance under various sky conditions, employing the clearness index (k_t).

It's important to note that ramp rate analysis is not required in the proposed methodology. This is because a regression problem is addressed using a classification-only approach. The data constructed classes based on historical observations, and therefore, traditional ramp rate analysis, which is usually relevant to shorter timescales, is not a part of the proposed methodology.

Figure 1 demonstrates the architectural design of the proposed methodology.

The development of the unsupervised classification-only day-ahead PV power production forecasting model followed a structured approach, encompassing several key stages to ensure data integrity and model reliability. It commenced with a rigorous data quality assurance process, validating historical data sources, including NWP data, weather measurements, and PV power production records, which are fundamental model parameters. Subsequently, a data-driven machine learning model was tailored to discern patterns and classify daily profiles based on historical data, forming distinct classes. Further categorization based on the clearness index allowed for the classification of days according to irradiance levels, providing additional insights into weather conditions. Lastly, a comprehensive performance evaluation assessed the model's accuracy by comparing its forecasts to actual PV power production, thus confirming the reliability and effectiveness of the forecasting methodology. Together, these stages ensured the development of a robust and validated classification-based approach for day-ahead PV power production forecasting.

2.1 Data quality routine and input features

To ensure the validity of the data used for model development and performance evaluation, we implemented an initial data quality routine (DQR) on the acquired datasets. Before proceeding with further analyses, a thorough examination of the recorded data was conducted to identify and rectify inconsistencies and gaps. The developed DQR includes a range of algorithms and techniques to address various data issues, such as missing and incorrect data, duplicate records, outliers, and outages. Additionally, it incorporates data filtering to restrict measurements to daylight hours and employs data correction methods to rectify missing data points that fall below a 10% threshold. In cases where missing rates exceed 10%, we utilize data deletion. These procedures are augmented by data inference techniques, enhancing the overall data completeness and reliability. These periodic data checks are crucial for maintaining the high quality of the dataset used in constructing forecasting models [27].

2.2 XGBoost algorithm

XGBoost, is a machine learning method that is both powerful and efficient. It is a member of the gradient boosting family of algorithms. It is capable of handling a wide variety of tasks, including classification, regression, and ranking, thanks to the design of its architecture. The principle of gradient boosting is at the heart of XGBoost. Gradient boosting entails iteratively developing an ensemble of weak prediction models and merging them to produce a powerful predictive model [26].

Initially, a base learner, commonly a decision tree, is established, and the method iteratively adds more trees to the ensemble. During each iteration, XGBoost aims to minimize a specific loss function by introducing decision trees that amend the errors made by preceding trees. Gradient descent guides this process, determining the direction and magnitude of adjustments needed at each stage.

Mathematically, during training, XGBoost optimizes the objective function by minimizing the sum of the loss function and a regularization term. The regularization term controls the model's complexity, while the loss function quantifies the difference between predicted values and actual labels. Regularization in XGBoost prevents overfitting and enhances the model's generalization capability. The regularization terms are derived based on L2 norms, adjusted according to the desired level of regularization.

Furthermore, XGBoost provides a feature importance measure, derived from analysing the decision trees within the ensemble. It quantifies the overall reduction in the loss function attributable to splits on each feature across all trees, aiding in identifying the most influential features and understanding underlying patterns. XGBoost also employs an early stopping mechanism during training, halting the process if the validation error stops improving, thus preventing overfitting.

2.3 Numerical weather predictions

The Weather Research and Forecasting (WRF) model is a comprehensive atmospheric simulation system utilized for meteorological research and operational forecasting. It begins with an initialization phase where input data from observations and other meteorological models are fed into the system, and the geographical domain and resolution of the model are configured by the user. The model then proceeds to the numerical integration phase where the continuous mathematical equations representing atmospheric processes are discretized for numerical solution over the defined grid, and time-stepping is employed to solve these equations at each grid point to predict future atmospheric conditions. Within the model, physical parameterizations are used to account for processes occurring at scales smaller than the model resolution, such as cloud formation and radiation transfer. Optionally, data assimilation techniques can be employed to incorporate additional observational data to correct model errors and enhance forecast accuracy. The model generates output data representing the predicted state of the atmosphere at various future times, which may undergo post-processing to generate user-friendly forecasts or derive additional meteorological quantities. Optionally, the model's predictions can be verified against actual observations to evaluate its performance and improve future runs, and nested simulations can be employed to run a higher-resolution model within a coarser model for more detailed forecasts over a smaller area, showcasing the WRF model's flexibility and extensive suite of options that cater to a wide range of meteorological applications [28–30].

2.4 Clearness index

The clear-sky index (k_t) is used to determine how much the current weather conditions are affecting the reliability of the forecast. A measurement of how transparent the air is, the k_t is denoted in degrees. The k_t is a dimensionless quantity, and the values it can take on range from 0 to 1. When conditions are cloudless and sunny, the k_t has a high value, while when conditions are cloudy and overcast, it has a low value. The clear-sky index is used as an indicator. This index provides a ratio of the sky conditions ($k_t = 0$ corresponds to an overcast sky, while $k_t = 1$ corresponds to a clear sky).

$$k_t = \frac{GHI}{GHI_{CS}} \tag{1}$$

where GHI and GHI_{CS} is the observed and clear-sky GHI using the Ineichen–Perez clear sky model [31,32], respectively.

Specifically, the k_t was separated into three classes. In this study, three classes of the k_t were used as follows:
– Clear Sky: $k_t \geq 0.75$;
– Moderate: $k_t < 0.75$ and $k_t > 0.25$;
– Overcast: $k_t \leq 0.25$.

2.5 Experimental apparatus

2.5.1 Description of the outdoor test facility at the University of Cyprus

This research utilized the PV Technology Laboratory's outdoor testing facility at the University of Cyprus (UCY). The test facility is outfitted with a fixed plane infrastructure that enables module and system-level outdoor performance evaluations. In this investigation, a polycrystalline silicon (poly-c-Si) system was installed in a portrait orientation on aluminium mountings and positioned at the optimal annual energy yield plane-of-array (POA) angle of 27.5, which was tailored to the climate of Cyprus. In addition, the PV system was linked to a data acquisition platform that monitored and stored meteorological and operational data pertaining to the PV system. Meteorological and PV operational measuring sensors are linked to a central data acquisition system on this data acquisition platform. The system was developed in accordance with the International Electrotechnical

Fig. 2. Test-bench PV system used for the day-ahead forecasting approach.

Commission (IEC) 61724 standard [33]. The platform records global irradiance (G_{i}), relative humidity (RH), wind direction (W_{a}), wind speed (W_{S}), and ambient temperature (T_{amb}) as meteorological measurements. In addition, the operational measurements of the PV system include the current at the maximum power point (I_{mp}), the voltage at the maximum power point (V_{mp}), and the power at the maximum power point (P_{mp}), which are measured at the DC output of the PV system [34].

The test-bench PV system comprises of 5 poly-crystalline silicon (poly-c-Si) PV modules, rated at 235 $\mathrm{W_p}$ each as depicted from the manufacturer's data-sheet. The modules are connected in series to form a PV string of nominal power capacity 1.175 $\mathrm{kW_p}$ at the input of a string inverter. It is installed in an open-field mounting arrangement due South, at the optimum annual energy yield angle for Cyprus of 27.5° (see Fig. 2).

2.5.2 Description of a utility scale PV plant at the Nicosia, Cyprus

The specific system is located in Nicosia, Cyprus with the nominal power capacity 1 MW. The generated electricity is fed into the local power grid, contributing to the overall energy supply of the area. The system incorporates supporting infrastructure such as to manage the produced energy following the IEC Standards. These components ensure that the electricity produced by the PV system is compatible with grid's specifications.

2.6 Model assessment

The prediction performance accuracy was assessed based on several predefined metrics. The mean absolute percentage error ($MAPE$) which is given by [35]:

$$MAPE = \frac{100}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_{\mathrm{observed,i}} - y_{\mathrm{forecasted,i}}}{y_{\mathrm{actual,i}}} \right|. \quad (2)$$

The $nRMSE$ which is the relative $RMSE$ normalized to the nominal capacity of the PV system and defined as [35,36]:

$$nRMSE = \frac{100}{P_{\mathrm{nominal}}} \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_{\mathrm{observed,i}} - y_{\mathrm{forecasted,i}}\right)^2} \quad (3)$$

where P_{nominal} is the maximum installed capacity of the PV system.

For the aforementioned performance metrics $y_{\mathrm{observed,i}}$ and $y_{\mathrm{forecasted,}}$ is the actual and forecasted power, respectively, P_{nominal} is the nominal peak power of the PV system, $RMSE_{\mathrm{forecasted}}$ and $RMSE_{\mathrm{baseline}}$ is the $RMSE$ of the forecasted and baseline models (forecasts of the PM), respectively. Additionally, it is noteworthy that the comparisons between the forecasted and actual values were conducted at 30-minute intervals, which were then aggregated to determine the daily errors.

3 Results

3.1 Study case UCY test-facility

The proposed approach underwent its first round of validation at the UCY test-facility. The primary purpose of these first tests was to evaluate the functionality and operation of the algorithms on a PV system that was on a smaller size. Before attempting to replicate the same techniques on a bigger utility scale PV system, this served as an essential step in ensuring the dependability and efficacy of the methodology.

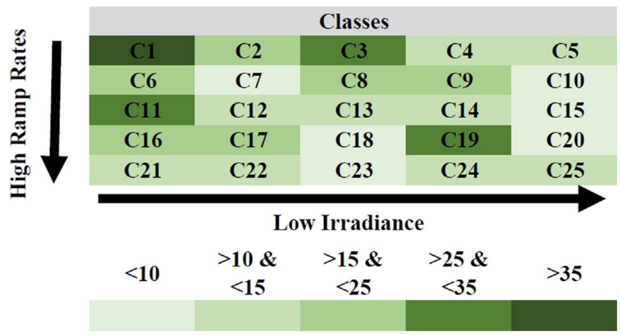

Fig. 3. Heatmap representing the number of the identify classes used as a dictionary. The colour code indicates the number days per class.

3.1.1 Identifications of the classes/dictionary implementation

As part of the process of forecasting PV generation, the approach that was proposed was implemented in the study that was carried out, and it entailed the building of 25 separate classes. These classes were represented using a visualization known as a heat map (see Fig. 3), in which each class was given a distinct colour that corresponded to the number of days that fell into that class. This heat map not only made it easier to conduct an analysis of the irradiance levels and ramping rates that are presenting the daily averages and are calculated as the average change in PV power output over the course of a day, indicating the overall variability in power generation that are related to each class, but it also provided a comprehensive overview of the distribution of days across the various classes. The color-coded representation made it possible to quickly locate groups of days that shared similar characteristics in terms of the amount of photovoltaic energy produced. In addition, the heat map made it possible to investigate the relationship between class assignment and associated irradiance levels. This inquiry shed insight into how varied weather conditions impacted the power production of the PV system. In addition, the heatmap offered information regarding the ramping rates that were present within each class. This provided a measurement of the pace at which the PV power output changed over the course of time.

3.1.2 Daily profile assessment

Additionally, a daily profile evaluation was conducted to assess the accuracy of the proposed PV production forecasting model on days. This evaluation compared the actual power output of the PV system to the two-day forecasted values. By evaluating the accuracy and dependability of the model's predictions on these designated days, insights into its performance under real-world conditions were obtained.

On the first day selected for testing, which occurred during the winter season (see Fig. 4a), the model precisely predicted the daylong power output trajectory of the PV system. The predicted values closely matched the actual power output, indicating a high degree of accuracy, especially on a day that exhibited rapid irradiance ramping rates. In general, the model was able to replicate the dynamic variations in power output associated with varying solar irradiance.

On the second summer day chosen for testing (see Fig. 4b), the model's performance was consistent and closely matched the actual power output throughout the day. The model's ability to predict PV generation patterns is validated by the significant correlation between predicted and observed power output. The model effectively accounted for variations in solar irradiance, resulting in precise predictions of power output variations. It accurately predicted the rapid increase in power production during the early morning hours, followed by a gradual decrease as the sun's intensity decreased in the late afternoon.

The daily profile evaluation of these selected days demonstrated the ability of the proposed day-ahead PV production forecasting model to precisely predict the system's power output. The congruence between predicted and actual power values showcases the model's ability to accurately forecast solar irradiance subtleties. These results provide valuable evidence of the model's performance under real-world conditions, validating its potential as a useful tool for PV generation forecasting one day in advance.

3.1.3 Overall performance evaluation

To evaluate the accuracy of the proposed PV production forecasting model for the day-ahead, a daily evaluation was conducted using two commonly used error metrics: the nRMSE and the MAPE (see Fig. 5). These metrics provided quantitative measures of the model's performance by comparing the forecasted and actual daily power output. The evaluation revealed an aggregate nRMSE of 8.20% (see Fig. 5a) and a MAPE of 6.91% (see Fig. 5b), indicating that the model's forecasts are highly correlated with the actual data sets. A low nRMSE value, such as 8.20%, indicated that the forecasts and actual measurements were in close agreement over the test set period.

Furthermore, the majority of evaluated days had error rates below 10%, showcasing that the predictions of the model were within an acceptable range and closely matched the observed power output. The low error rates demonstrated the model's ability to generate accurate forecasts, further establishing its viability for forecasting PV production one day in advance.

In addition, the MAPE value of 6.91% quantified the average percentage deviation between the forecasted and actual power output. The relatively low MAPE value showcased that the model's predictions deviated from the observed values by a small average percentage, strengthening the model's precision and indicating its capacity to generate accurate forecasts.

Moreover, a comprehensive evaluation was conducted to analyse the errors documented based on the sky conditions using the k_t index (see Tab. 1). This evaluation sought to determine how the model's performance varied depending on the weather. With a nRMSE of 5.30% and a MAPE of 4.10%, the results indicated that the fewest errors

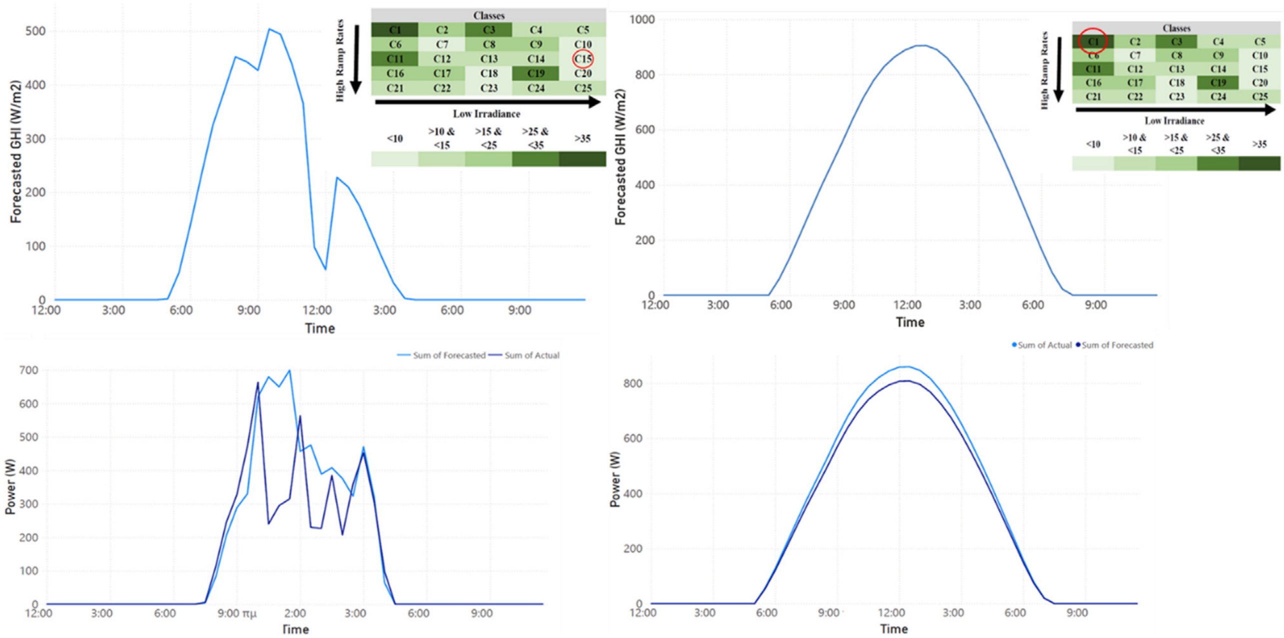

Fig. 4. Daily profile assessment of: (a) an overcasted day during the winter period and (b) a clear sky day during summer period.

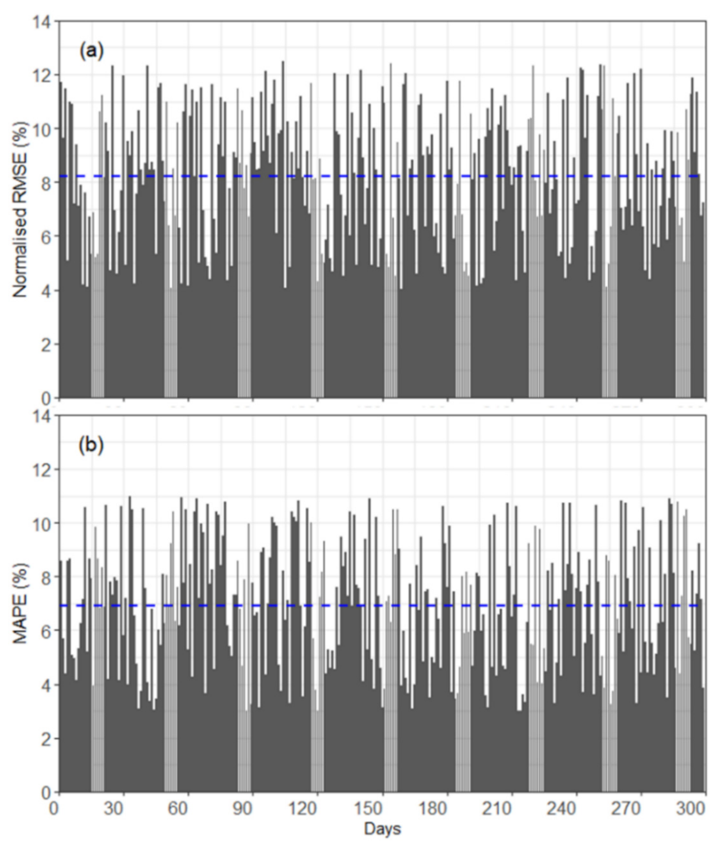

Fig. 5. Daily performance evaluation metric of: (a) nRMSE and (b) MAPE over the testing period. The blue dashed line indicates the aggregated error both for nRMSE and MAPE.

Table 1. Correlation coefficients between the input and output parameters (investigated features).

Clearness index (k_t)	Class	nRMSE (%)	MAPE (%)
Clear sky	C1–C5	5.30	4.10
Moderated	C6–C19	8.99	7.90
Overcasted	C19–C25	10.30	8.70
Overall		8.20	6.91

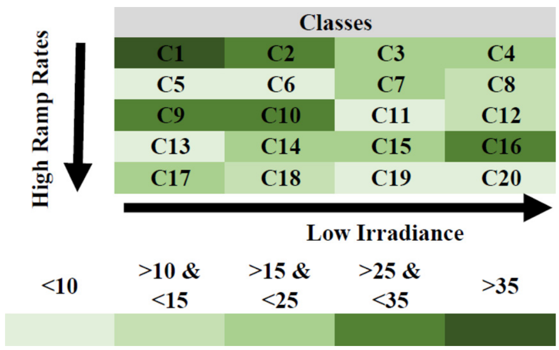

Fig. 6. Heatmap representing the number of the identify classes used as a dictionary. The colour code indicates the number days per class.

occurred on days with clear skies. When sky conditions were favourable and solar irradiance was high, these results demonstrated the model's reliability in accurately forecasting PV power output.

Upon examining the efficacy on moderate and cloudy days, the error rates were found to be slightly higher. Despite this, error rates on these days remained within an acceptable range, demonstrating the model's ability to provide reasonably accurate forecasts even under less-than-ideal weather conditions. The number of days classified as moderate or overcast was significantly lower than the number of days with clear skies. This discrepancy in the number of days further supports the conclusion that the errors were greater on those days due to more challenging weather conditions. By analysing the effect of sky conditions on forecasting errors, it became apparent that the model's performance was most accurate on days with clear skies, while remaining acceptable on days with moderate or heavy clouds. This information demonstrates the model's adaptability to diverse weather conditions and its ability to generate accurate forecasts for various scenarios. Additionally, Table 1 includes the representative classes derived by the XGBoost classifier for each sky condition, as determined by the k_t index.

3.2 Study case 1: MW PV power plant

The optimal objective of the proposed method is to develop a methodology that is simple to replicate. Consequently, if the same procedures as in Section 3.1 are followed, the proposed methodology should yield comparable results. Consequently, this segment replicates the methodology for the 1 MW utility-scale PV power plant in Nicosia, Cyprus.

3.2.1 Identifications of the classes/dictionary implementation

The specific XGBoost classification entailed the construction of 20 distinct classes. As depicted in Figure 6, these classes were visually represented using a heatmap, with each class being designated a specific colour based on the number of days falling into that class. Besides facilitating the analysis of irradiance levels and ramping rates associated with each class, the heatmap also provided a comprehensive overview of the distribution of days across the various classes.

By carefully examining the heatmap, it was possible to identify recurring patterns and trends in the data. The color-coded representation simplified the identification of groups of days sharing comparable PV energy production characteristics. Moreover, the heatmap enabled the examination of the relationship between class assignment and corresponding irradiance levels, revealing how various weather conditions affected the PV system's power output. It also provided useful information regarding the ramping rates observed within each class, signifying the rate at which the PV power output changed over time. This measure of change offered insight into the performance dynamics of the PV system.

3.2.2 Overall performance evaluation

To further investigate the accuracy of the forecasts, a contour plot analysis of the concentration of the residuals in relation to the forecasted values was conducted. This analysis sought to graphically depict the relationship between forecasted values and residuals. The contour plot revealed a high concentration of forecasts near the observed values, indicating a strong correlation between the predicted and observed power output. The compact distribution of residuals indicated that the model consistently generated predictions that closely matched the actual measurements. This result bolstered the predictability and precision of the proposed PV production forecasting model for the following day. Upon scrutinizing the contour plot, it became evident that the model captured the variations and patterns in the PV power output accurately, resulting in minimal deviations between forecasts and actual values. The close clustering of residuals around zero indicated that, on average, the forecasts were very close to the observed power output, with few significant deviations. The observed preponderance of residuals on the contour plot provided additional evidence that the model is capable of producing accurate forecasts. This concentration demonstrated the model's

Table 2. Performance evaluation.

Clear-sky index (k_t)	Class	nRMSE (%)	MAPE (%)
Clear sky	C1–C3	6.88	5.20
Moderated	C4–C17	9.50	8.30
Overcasted	C18–C20	11.05	10.10
Overall		9.14	7.86

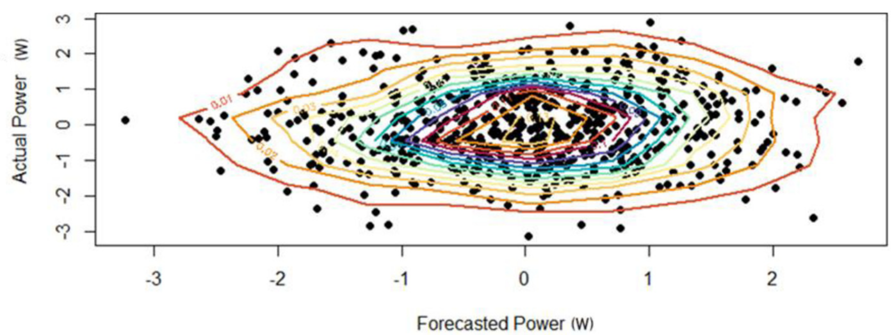

Fig. 7. Contour plot of the forecasted against the actual power to evaluate the concertation of the forecasted parameters.

ability to precisely predict the PV power output and minimize discrepancies between predictions and actual measurements (Fig. 7).

A comprehensive analysis of the errors documented based on the sky conditions using the k_t index (see Tab. 2). This evaluation sought to determine how the model's performance varied depending on the weather. The evaluation revealed a nRMSE of 6.88% and a MAPE of 5.20%, indicating that days with clear skies produced the fewest errors. These results demonstrated the model's accuracy in forecasting PV power output under favourable weather conditions and high solar irradiance. On moderate and cloudy days, however, marginally higher error rates were observed when evaluating the model's performance. Nevertheless, error rates remained within an acceptable range during these days, indicating the model's ability to provide reasonably accurate forecasts even under less-than-ideal weather conditions. It is important to observe that the number of days with moderate or overcast conditions was significantly lower than the number of days with clear skies. This variance in the number of days further supported the conclusion that errors were more prevalent on days with more difficult weather conditions. Analysing the effect of sky conditions on forecasting errors revealed that the model performed most accurately on days with clear skies, while maintaining an acceptable level of accuracy on days with moderate or significant cloud cover. This adaptability to diverse meteorological conditions demonstrated the model's ability to produce accurate forecasts for a variety of scenarios. In addition, Table 2 displays the representative classes derived by the XGBoost classifier for each sky condition based on the k_t index. The recorded overall nRMSE and MAPE were 9.14% and 7.86% respectively.

3.3 Proposed methodology replicability

The proposed forecasting method was implemented and evaluated at two distinct facilities to determine its duplicability and applicability. The overall behaviour of the methodology exhibited similar patterns and characteristics at both testing locations, indicating successful replicability.

At the first testing facility, located at the University of Cyprus, the methodology demonstrated consistent accuracy in forecasting PV power output. The observed patterns and trends in the power output data closely matched the predicted values, validating the method's ability to capture the dynamics of the evaluated PV system. These results instilled confidence in the methodology's robustness and its capability to provide accurate forecasts at this particular facility.

Similarly, at the second testing facility, the methodology exhibited comparable behaviour when replicated. The forecasts generated closely matched the actual power output, demonstrating the method's ability to account for the specific characteristics of the PV system at this facility. The reproducibility of the proposed method and its potential applicability in various contexts were bolstered by the consistent performance across different testing facilities.

The successful replication of the methodology at two separate testing facilities underscores its dependability and flexibility. The consistent behaviour observed in both instances indicates that the forecasting model and associated algorithms are adept at capturing the intricate relationships between meteorological variables, PV system parameters, and power output. This replicability provides valuable insights into the model's performance across various PV installations, contributing to its wider applicability in the field of PV generation forecasting.

4 Conclusions

The increasing integration of photovoltaic (PV) systems into electricity grids has introduced new challenges related to reliability, primarily driven by the dependence on weather conditions for solar energy generation. In response to this challenge, this study aimed to develop an accurate day-ahead PV production forecasting methodology using advanced machine learning techniques and statistical approaches to reduce solar irradiance prediction uncertainties.

PV generation forecasting plays a critical role in effectively planning, operating, and optimizing power grids, enabling utilities and system operators to efficiently schedule dispatchable energy resources. This research introduced a methodology that leverages novel machine learning techniques, specifically focusing on a classification-only forecasting approach. This approach involves categorizing future PV system power output into predefined classes rather than predicting exact power values, which is well-suited for day-ahead forecasting with inherently high prediction uncertainties. Enhanced machine learning methods were applied to extend the forecasting horizon to up to 24 h ahead.

Based on this methodology is the development of an unsupervised classifier model based on the Extreme Gradient Boosting (XGBoost) ensemble algorithm. This model classifies daily 30-minute profiles of forecasted global horizontal irradiance (GHI), measured incident irradiance (G_i), and AC power (P_{AC}) into distinct classes. The classifier model effectively acts as a dictionary, assigning newly forecasted GHI to specific classes and, subsequently, identifying the corresponding forecasted PAC.

The results demonstrated the effectiveness of this forecasting solution, achieving a daily normalized root mean square error (nRMSE) of 8.20% and a mean absolute percentage error (MAPE) of 6.91% over the test period. Furthermore, the methodology's performance under varying sky conditions, as assessed using the clearness index (kt), revealed higher accuracy during clear-sky days while maintaining errors within acceptable limits during moderate and overcast conditions.

Future work in this area may involve further refinement and optimization of the methodology, potentially incorporating additional meteorological parameters or advanced machine learning techniques to enhance forecasting accuracy under different weather scenarios. Additionally, expanding the study to a broader range of geographical locations and PV system configurations could provide valuable insights into the methodology's adaptability and generalizability. Overall, this research underscores the potential of classification-based forecasting as a valuable tool for improving day-ahead PV production forecasts, offering benefits for the reliable integration of solar energy into power grids.

Funding

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 864537, project title Flexible Energy Production, Demand and Storage-based Virtual Power Plants for Electricity Markets and Resilient DSO Operation (FEVER).

Conflicts of interest

The authors would like to declare no conflict of interest.

Data availability statement

Data availability upon request.

Author contribution statement

S.T. conceived of the presented idea, developed the theory and performed the computations and discussed the results and wrote the final manuscript. S.T. and G.M. verified the analytical methods. G.E.G supervised the project.

References

1. IEA, Electricity security in tomorrow's power systems, 2020. https://www.iea.org/articles/electricity-security-in-tomorrow-s-power-systems
2. IEA, Renewables 2019: market analysis and forecast from 2019 to 2024, 2020. https://www.iea.org/reports/renewables-2019
3. IEA, Introduction to system integration of renewables: decarbonising while meeting growing demand, 2020. 2020. https://www.iea.org/reports/introduction-to-system-integration-of-renewables
4. M. Mahoor, A. Majzoobi, A. Khodaei, Distribution asset management through coordinated microgrid scheduling, IET Smart Grid. **1**, 159 (2018)
5. H. Sadeghian, Z. Wang, A novel impact-assessment framework for distributed PV installations in low-voltage secondary networks, Renew. Energy **147**, 2179 (2020)
6. International Renewable Energy Agency, Innovation landscape for a renewable-powered future, 2019, https://www.irena.org/publications/2019/Feb/Innovation-landscape-for-a-renewable-powered-future
7. H.T.C. Pedro, C.F.M. Coimbra, Assessment of forecasting techniques for solar power production with no exogenous inputs, Sol. Energy **86**, 7 (2012). https://doi.org/10.1016/j.solener.2012.04.004
8. B. Wolff, O. Kramer, D. Heinemann, Selection of numerical weather forecast features for PV power predictions with random forests, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (2017), pp. 78–91. https://doi.org/10.1007/978-3-319-50947-1_8
9. T. Schmidt et al., Short-term solar forecasting based on sky images to enable higher PV generation in remote electricity networks, Renew. Energy Environ. Sustain. **2**, 23 (2017)
10. L. Visser, T. AlSkaif, W. van Sark, Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution, Renew. Energy **183**, 267 (2022)
11. S. Theocharides, G. Makrides, *Day-ahead solar photovoltaic forecasting for EAC supply portfolios* (Nicosia, Cyprus, 2022)
12. S. Pretto et al., A new probabilistic ensemble method for an enhanced day-ahead PV power forecast, IEEE J. Photovol. **12**, 581 (2022)

13. D. Caputo et al., Photovoltaic plants predictive model by means of ANN trained by a hybrid evolutionary algorithm, in *The 2010 International Joint Conference on Neural Networks (IJCNN),* (IEEE, 2010), pp. 1–6. https://doi.org/10.1109/IJCNN.010.5596782
14. A. Mellit, A.M. Pavan, A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy, Sol. Energy **84**, 807 (2010)
15. Y. Xue et al., Voltage stability and sensitivity analysis of grid-connected photovoltaic systems, in *2011 IEEE Power and Energy Society General Meeting*, (IEEE, 2011), pp. 1–7
16. S. Pelland et al., Photovoltaic and solar forecasting: state of the art, in *International Energy Agency: Photovoltaic Power Systems Programme*, (Report IEA PVPS T14, 2013), pp. 1–40. https://www.researchgate.net/publication/259717376_Photovoltaic_and_Solar_Forecasting_State_of_the_Art, https://doi.org/978-3-906042-13-8
17. W. Glassley et al., *California Renewable Energy Forecasting, Resource Data and Mapping* (California Energy Commission, 2011), Publication number: CEC-500-2014-026
18. A. Yona et al., Application of neural network to 24-hour-ahead generating power forecasting for PV system, in *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, (IEEE, 2008) pp.1 –6. https://doi.org/10.1109/PES.2008.4596295
19. J.G. Da Silva Fonseca Junior et al., Forecasting regional photovoltaic power generation – a comparison of strategies to obtain one-day-ahead data, Energy Procedia **57**, 1337 (2014)
20. A. Dolara et al., A physical hybrid artificial neural network for short term forecasting of PV plant power output, Energies **8**, 1138 (2015)
21. A. Gandelli et al., Hybrid model analysis and validation for PV energy production forecasting, in *2014 International Joint Conference on Neural Networks (IJCNN)*, (IEEE, 2014), pp. 1957–1962. https://doi.org/10.1109/IJCNN.2014. 6889786
22. J. Shi et al., Forecasting power output of photovoltaic system based on weather classification and support vector machine, in *2011 IEEE Industry Applications Society Annual Meeting* (IEEE, 2011) pp. 1–6. https://doi.org/10.1109/TIA.2012.2190816
23. M. Paulescu et al., Weather Modeling and Forecasting of PV Systems Operation in *Green Energy and Technology* (Springer, London, 2013). https://doi.org/10.1007/978-1-4471-4649-0
24. D. Yang et al., A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: towards carbon neutrality, Renew. Sustain. Energy Rev. **161**, 112348 (2022)
25. A. Mellit, A.M. Pavan, V. Lughi, Deep learning neural networks for short-term photovoltaic power forecasting, Renew. Energy **172**, 276 (2021)
26. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016* (ACM, 2016), pp. 785–794. https://doi.org/10.1145/2939672.2939785
27. A. Livera et al., Advanced diagnostic approach of failures for grid-connected photovoltaic (PV) systems, in *35th European Photovoltaic Solar Energy Conference (EU PVSEC*), (2018), pp. 1548–1553. https://doi.org/10.4229/35thEUPVSEC20182018-6BO.6.5
28. W.C.Skamarock-NCAR/UCAR et al., ARW modelling system userguide – V.3, p. 408, 2016. https://doi.org/10.5065/D68S4MVH
29. D. Orrell et al., Model error in weather forecasting, Nonlinear Process. Geophys. **8**, 357 (2001). https://doi.org/10.5194/npg-8-357-2001
30. P. Lynch, The origins of computer weather prediction and climate modeling, J. Comput. Phys. **227**, 3431 (2008)
31. P. Ineichen, R. Perez, A new airmass independent formulation for the Linke turbidity coefficient, Sol. Energy. **73**, 151 (2002)
32. R. Perez et al., A new operational model for satellite-derived irradiances: description and validation, Sol. Energy. **73**, 307 (2002)
33. 6. IEC, Photovoltaic system performance – Part 1: monitoring, IEC 61724-1, 2017
34. G. Makrides et al., Potential of photovoltaic systems in countries with high solar irradiation, Renew. Sustain. Energy Rev. **14**, 754 (2010)
35. C. Tofallis, A better measure of relative prediction accuracy for model selection and model estimation, J. Oper. Res. Soc. **66**, 8 (2015)
36. R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. **22**, 679 (2006)

Cite this article as: Spyros Theocharides, George Makrides, George E. Georghiou, PV generation forecasting utilizing a classification-only approach, EPJ Photovoltaics **15**, 12 (2024)