Conference papers

https://doi.org/10.52825/pv-symposium.v2i.2636

© Authors. This work is licensed under a Creative Commons Attribution 4.0 International License

Published: 27 Aug. 2025

Predicting the Shading of Photovoltaic Systems Using Machine Learning

Maximilian Schönau^{1,2,*} , Joseph Jachmann², Markus Panhuysen¹, Alexander Schönau³, Darwin Daume², Achim Schulze⁴, Bernd Hüttl², and Dieter Landes²

¹smartblue AG, Germany

²Coburg University of Applied Sciences, Germany

³Catholic University of Eichstätt-Ingolstadt, Germany

⁴Rosenheim Technical University of Applied Sciences, Germany

*Correspondence: Maximilian Schönau, Maximilian.Schonau@smartblue.de

Abstract. In the operation of photovoltaic power plants, precise knowledge of shading is essential in order to carry out differentiated yield and site analyses and to guarantee reliable monitoring and fault detection. A study on the causes of shading carried out with the help of GPT4-o is presented. Subsequently, an innovative approach for predicting shading using a is introduced. By combining physical and data-driven machine learning, it is possible to efficiently complete incomplete shadow analyses and eliminate erroneous data points of a physical model. The presented method utilizes data from 1380 photovoltaic devices with various shading scenarios to train an autoencoder on PV system shading. The autoencoder enables accurate prediction of shading within a detection time of only a few weeks.

Keywords: Operation and Maintenance (O&M), Digital Twin, Origins of Shadows

1. Introduction

When monitoring photovoltaic systems, shading is a critical factor that causes an average of around 7 % of power losses [1]. Shading is caused by site-specific conditions, self-shading, and temporary factors such as snow, foliage or vegetation.

Site-related and self-shadowing are often accepted to increase the degree of area utilization, yet it is crucial to precisely determine their timing and extent. This enables differentiated yield analyses, allowing operators to evaluate the efficiency and planning of their photovoltaic power plants and to optimize future projects by making adjustments such as changing row spacing or module connections [2]. Temporary shading, such as grass growth, should be monitored so that it can be corrected if necessary. Last but not least, a precise evaluation of shading plays a major role in the fault monitoring of PV power plants. Shading on strings causes short-term power losses, which can lead to false error messages to the operator if these shading periods are not excluded from the error detection [3].

Shading of PV systems is an issue, that O&M Managers often addressed in reports. Similar to our previous work [4], these reports were analyzed using *GPT4-o* by OpenAl [5], [6], [7], to get statistical information about the origin of shading issues.

The dataset consists of 5,089 medium to large-scale ground-mounted and rooftop PV plants constructed between April 1999 and January 2025. About 25 % were built before October 2013, with 75% completed by September 2020. These plants primarily feature traditional silicon module technologies representative of the past decade in Germany. Thin-Film modules and tracker systems account for less than 0.3% and 3% of installations, respectively [4].

O&M reports were anonymized and then filtered by using keyword identification of the word *shading*, which resulted in 12487 comments. Most of these comments were filtered out by *GPT4-o*, as they did not specify the reason of the shading issue. In addition to that, many reports containing the word shading did not specify physical and external shading issues, but problems like snow instead. The filtering process resulted in 530 comments from Asset Managers, who specified the origin of the shading issue of the monitored string.

After manual data inspection, the following categories were defined for the classification task:

- Trees Shading caused by trees or bushes
- Gras Shading caused by grass or other growing vegetation
- Self Shading caused by the solar panels themselves
- Buildings Shading caused by buildings or structures such as the roof, chimneys
- Others A category if none of the above fit

The full prompt used for the classification task is specified in the Appendix. 10 % of the categorized comments were randomly picked and manually reviewed, yielding no instance of misclassification.

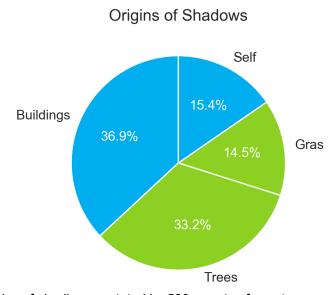


Figure 1. Origins of shading, as stated by 530 reports of asset managers in Germany.

Figure 1 displays the result of the classification task. About a third of all shading issues stem from building related shading, such as shading because of the structure of the roof of rooftop PV plants. Another third of the reported shading issues stem from trees. Self-shading and shading because of vegetation / gras accounted for 15.3 % and 14.4 %, respectively. 5 reports were classified as "Others", which account for reports of shading because of mountains and cleaning activities.

While shading issues stemming from buildings and self-shading can mostly not be modified on a PV plant, it is still important to quantify the resulting losses for accurate asset management. Around half of the shading issues however stem from vegetation, which can in many cases be removed if the power losses justify the expenditure.

The methodology may not be representative for PV plants in Germany since the comments may have been biased. Among others, biases could stem from asset managers being more aware and thus reporting some shadow origins more frequently than others. In addition to that, the PV plants monitored by smartblue may be not representative for Germany. Nevertheless, the 530 comments that could be accurately classified by *GPT4-o* do still hold relevant information about the most pressing shading issues.

2. Prediction of Shading

To detect shading, algorithms are often used that calculate the shading of a string from periodic power reductions using photovoltaic models [8]. Another method for determining shading is IV curves [9], [10], [11]. However, the measurement technology and model parameters required for this make these methods complex in practice. We therefore used a simpler and universally applicable method that only uses electrical power data from an inverter: By evaluating the photovoltaic power during clear-sky, a power reduction in direct sunlight is determined, from which the shading of a string is calculated [12]. The shading matrix η^{ij} results from the quotient of the measured power $P^{ij}_{\rm meas}$ and the modeled theoretical power $P^{ij}_{\rm gen,\,cs}$ under clear-sky [13] for all azimuth and elevation angles ij [12]:

$$\eta^{ij} = \frac{P_{\text{meas}}^{ij}}{P_{\text{gen, cs}}^{ij}} \tag{1}$$

Figure 2 shows the shading matrix of an example string at a resolution of 28×28 over the azimuth α and the elevation angle ϕ . Low values of efficiency and thus shaded points are marked with darker colors, while white light grey resembles high efficiencies and thus shadow free sun positions. The string is heavily shaded to the east and slightly shaded to the west at low sun elevations.

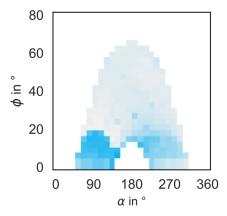


Figure 2. Exemplary shadow matrix.

A physically calculated gap-free shadow matrix such as displayed in Figure 2, requires a detection time of up to several years. In addition to that, the accuracy of the method is limited: Uncertainties in the clear-sky model and in the modeling of the photovoltaic power result in deviations, which can be recognized in Figure 2 by noise in the matrix. This can be improved with the help of machine learning.

A method for the correction and prediction of the shadow matrix is presented, which enables a detailed detection of shadows with a short detection time. A denoising CNN autoencoder is used to correct and complete incomplete shadow predictions. This is possible because shadows have cyclical patterns: For example, if the contours of a building are apparent in an incomplete shadow matrix, an intelligent model can complete these contours without needing the irradiance efficiency for all azimuth and elevation angles over a year. In addition, noise can be removed, as shadows usually have regular and uniform shapes, which can be recognized and corrected by the autoencoder.

2.1 Data

A comprehensive dataset was generated using strings for which shading was specified by the asset manager. The shading matrix was exported for 1380 strings covering the periods from January 2021 to January 2025, from which 30 datasets per string were created for training and testing. Sparse matrices were created by randomly selecting time slices ranging from 1 to 75 weeks, presenting incomplete data that the model is required to predict. To enhance dataset diversity, each matrix was inverted along the azimuth axis, effectively doubling the data. The resulting shadow matrices were then down sampled to a resolution of 28×28 values. Finally, the entire dataset was randomly split into 80% training and 20% testing sets, with a strict separation of PV strings between the two subsets to prevent overfitting. For the model training and evaluation, the dataset was filtered by using only datapoints with a simulated clear-sky power over $100 \frac{W}{kWp}$. The visualizations of the shadow matrices were done without this filter.

2.2 Clear-Sky Correction

The physical clear-sky model [12] of the modeled power $P_{\mathrm{gen, \, cs}}^{ij}$ deviates from the measured power, especially for aged strings. Thus, a clear-sky correction was incorporated into the preprocessing of the model, by correcting the aging losses of the string using the 75th percentile of the shadow matrix as an estimate of a strings aging α . This percentile estimates the normal power output of a string during clear-sky. The corrected shadow matrix $\eta_{\rm corr}^{ij}$ was thus calculated by:

$$\alpha = P_{75}(\{\eta^{ij}\}) \tag{2}$$

$$\alpha = P_{75}(\{\eta^{ij}\})$$

$$\eta_{\text{corr}}^{ij} = \begin{cases} \frac{\eta^{ij}}{\alpha} & \text{if } \frac{\eta^{ij}}{\alpha} \le 1, \\ 1 & \text{if } \frac{\eta^{ij}}{\alpha} > 1. \end{cases}$$
(3)

The correction enables using the "background color" of the shading matrix as a reference value, by which all shadow matrix points are corrected, with a limit to 100 % efficiency. This adjustment enhances the accuracy of the methodology by sharpening the contrast within the shadow matrices, facilitating more precise modeling of string power based on irradiance conditions and the refined shadow matrix.

2.3 Model

A proprietary denoising CNN autoencoder is used for shadow prediction. Figure 3 outlines the structure of the model.

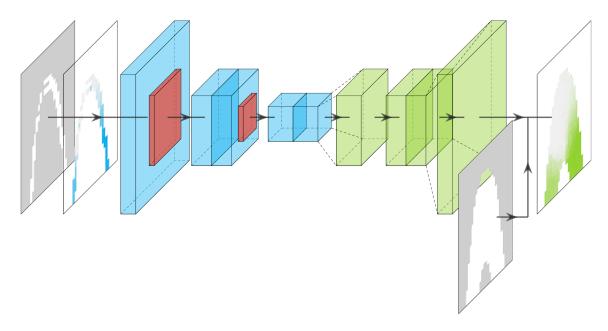


Figure 3. Schematic representation of the architecture of the denoising CNN autoencoder.

Autoencoders are neural networks that are trained to encode and then reconstruct input data by learning a representation of the data in a hidden intermediate layer [14]. This representation is usually compressed and contains meaningful features that the neural network had to learn to reconstruct the original data.

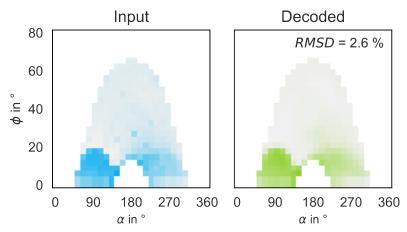


Figure 4. Test input and output of an autoencoder trained with complete shadow matrices.

Figure 4 shows a test input and reconstruction of an autoencoder that was trained with complete shadow matrices as input and output variables. The autoencoder is fed with the input data matrix (represented by the blue image), which is compressed and mapped in the intermediate dense layer. This dense layer is then reconstructed into the output data matrix (represented by the green image). Through the encoding and decoding, only the relevant shadings are transmitted in the form and characteristics, instead of a complete reconstruction of the data points. For this reason, the shading of the decoding is more structured and uniform as well as freed from noise of the original image.

Denoising autoencoders are a special form of autoencoder in which data points are specifically removed from the input data [15]. The shadow prediction masks were created by randomly selecting time intervals from 1 to 75 weeks of the full shadow matrix to create incomplete datasets. The masked shadow matrix was used as input to the model, with a Boolean matrix as the second channel, which specifies data points to be completed. Training of the model has been done on full shadow matrices.

A convolutional network (CNN) was used as the basis for the denoising autoencoder [16]; a scaled-down version of the VGG architecture [17] proved to be advantageous.

2.4 Physics-Informed Machine Learning

The autoencoder used is a hybrid model that combines knowledge-driven and data-driven aspects [18]: The completion of the shading was implemented in a data-driven manner using classical machine learning. However, the process could be optimized by incorporating the physical knowledge about the sun's course at the given location into the model.

A shading pattern, such as that of a building, should always be completed with the same learned patterns, regardless of the direction in which the building is located in relation to the PV system. It was attempted to enforce this by not requiring the model to take into account the position of the sun, but only training it by completing shadows.

This was implemented by adapting the error function L: For each training data set k, the mean squared deviation MSD from the masked and complete shading matrix $(\eta_{\mathrm{masked}}^{ij} - \eta^{ij})$ was multiplied by a sun position matrix χ^{ij} , which maps the course of the sun at the respective location in Boolean values:

$$L = MSD = \frac{1}{n_i \cdot n_j \cdot n_k} \sum_{k=1}^{n_k} \sum_{i=1}^{n_i} \left[\left(\eta_{\text{masked}}^{ijk} - \eta^{ijk} \right) \cdot \chi^{ijk} \right]^2$$
 (4)

The root mean squared error of the shading matrix was used to evaluate the model performance:

$$RMSD = \sqrt{MSD} \tag{5}$$

By adapting the error function in this way, the model can complete all shading according to the learned patterns, regardless of the actual position of the sun. The CNN is also free to complete those "shadings" that occur outside of all physical sun positions. This improves the generalization of the method, as the model is able to complete shadows independently of azimuth and elevation.

3. Results

After model training, the test results were evaluated and compared to each other using the masked RMSD. All model evaluations were done by a 5-fold cross-validation. The model was trained and tested on data over all seasons for time spans of 1 to 75 weeks, which resulted in an RMSD of 7.0 % with a standard deviation of 0.1 %.

This value displays a relatively high similarity between prediction and reference data compared to the training dataset, however it does not directly state the ability of the shadow matrix to correct the power value that a PV device should deliver. With the autoencoded shadow matrices being more refined than the original training dataset and missing outliers, there is an argument to be made that shadow predictions with high *RMSD* are in some cases better than the original train dataset, having corrected outliers (see Figure 4).

The accuracy of the shadow matrix in terms of its ability to correct the power that a PV device should generate will be investigated in future work. In this work, the *RMSD* of the predicted shadow matrix to the full matrix will be used as an indicator of model performance for different train and test datasets. In addition to the *RMSD* visual inspections were used to assess the quality of the predicted shading patterns.

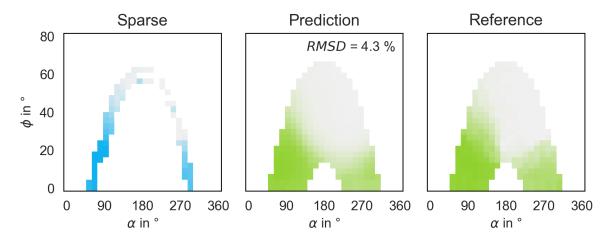


Figure 5. Sparce, Predicted and Full shadow matrix.

Figure 5 contains an incomplete shadow matrix of the test data set with an *RMSD* of 4.3 %. The reference, which took several years of electrical data from the string to generate, was predicted with a high accuracy using data points from only four weeks.

3.1 Advantages of Physics-Informed Machine Learning

The advantage of the hybrid knowledge- and data-driven model was compared with a data-driven approach, by training the model without the clear-sky correction of Chapter 2.2 and without the custom loss function of Equation (4).

Clear-Sky Corrected	Custom Loss Function	<i>RMSD</i> in %
		10.9 ± 0.2
X		9.1 ± 0.4
	X	9.0 ± 0.4
X	X	7 0 + 0 1

Table 1. Advantage of Physics-Informed Machine Learning.

Table 1 displays the models' performance without clear-sky correction and without the custom loss function. Both the clear-sky correction and the custom loss function reduces the models *RMSD* by about 30 %, which emphasizes the importance of combining knowledge-driven with data-driven methodologies.

3.2 Practical Limitations of the Shadow Prediction

The prediction of the shadow matrix is not perfect if a small amount of data is available. Figure 6 shows an incorrect prediction of the shadow matrix for such a case. In this example, the autoencoder had no information about the shadows cast when the sun was shining from the east. This led the model to incorrectly extrapolate the shading from the west to the east. Only with additional information on eastern shading after an increase in detection time does the model succeed in making an accurate shadow forecast.

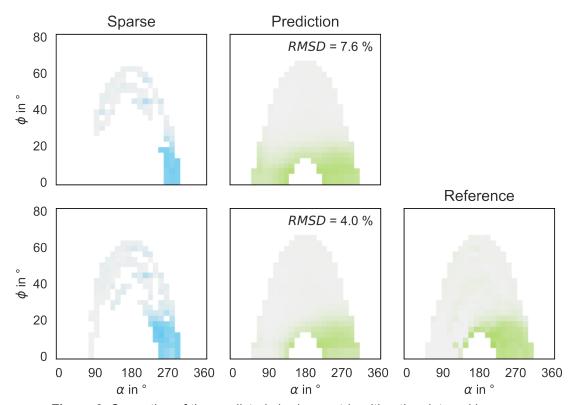


Figure 6. Correction of the predicted shadow matrix with a time interval increase.

In practice, the shadow forecast can be dynamically adapted to the available data points of the irradiance efficiency matrix, whereby the predicted shading becomes increasingly accurate as the monitoring period of a PV system increases. For most devices, it is possible to reliably forecast a complete shadow matrix within a few months.



Figure 7. Model performance when increasing the detection time.

Figure 7 displays the *RMSD* of the shadow prediction over an increase in detection time. Appendix 1 and 2 display examples of shadow predictions after 4 and after 25 Weeks. The shadow prediction is very accurate in most cases after 25 Weeks with an *RMSD* of 8 %. Notably, even at 4 weeks, the predictions are sufficiently accurate for effective asset management in many cases. This is especially important for avoidable shading such as growing grass, which must be detected as quickly as possible so that countermeasures can be taken.

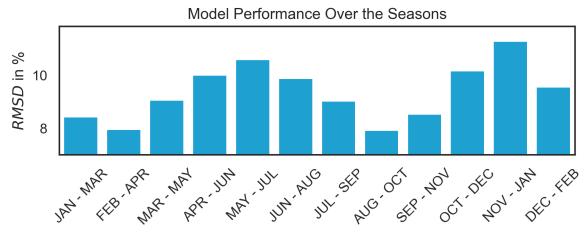


Figure 8. Model performance for different months of the year with 10 Weeks detection time.

Figure 8 displays the model performance for different months of the year, for a shadow prediction with a detection time of 10 weeks. The model's performance varies about 4 % percent for different months of the year, with the best time intervals being months with medium sun heights such as February to April and August to October. The model performs worse at solstice, with the summer solstice performing slightly better than the winter solstice, since months in the summer have more clear-sky intervals at higher sun positions, which return more information about shading patterns.

4. Conclusion

The integration of physical modeling with data-driven machine learning enables rapid and accurate prediction of shading patterns in photovoltaic systems. The hybrid approach leads to more precise shading forecasts and thus facilitates fault monitoring and system management. The shadow matrix becomes more precise with increasing monitoring time and thus improves the fault and analysis models. As a result, the required maintenance work will be reduced in future and the number of false alarms for PV systems will decrease. In addition, the methodology will enable precise calculations of yield losses due to shading.

Appendix

A.1 GPT-40 Prompt for the Classification of Shading Issues

Below the prompt that was provided to *GPT-4o* for classifying the origin of shading issues:

You are a senior photovoltaic asset manager analyzing shading issues.

IMPORTANT: The input text is in German, but you must respond in English.

TASK: Classify the comment into these categories:

- 1) General General shading mentioned without reason, or shading reason unclear
- 2) No shading Shading was not the issue
- 3) Tree Shading caused by trees or bushes

- 4) Gras Shading caused by grass or plants or other growing vegetation
- 5) Self Shading caused by the solar panels themselves
- 6) Building Shading caused by buildings or structures such as the roof, chimneys
- 8) Snow Shading caused by snow or ice
- 10) <Other> Create a new category if none of the above fit at all

CRITICAL RULES:

- Return a JSON object with a short description of the reason and the categories that apply. Example: "'{"reason": "Shading caused by trees", "categories": ["Tree"]}"
 - Try to be as specific as possible, if no reason is given, return the category "General"
 - If the shading cause is only speculated about, return the category "General"
 - Input is in German, look for words like: Verschattung, Schatten, verschattet
- High attention to context and German language nuances. If a building is mentioned, this is not automatically the cause of the shading.
 - You will be penalized for incorrect classifications or wrong JSON format

EXAMPLES:

Input: "Verschattung durch Bewuchs. Referenz ist das sanierte Kiesdach." Output: {"reason": "Shading caused by vegetation", "categories": ["Gras"]}

Input: "Die Schatten werden länger. S19.3 und S19.6 sind in er unteren Modulreihe und werden so von den vorderen Modultischen beschattet." Output: {"reason": "Shading caused by the modules themselves", "categories": ["Self"]}

Input: "Jahreszeitliche Verschattung, auf dem Dach von Garage 1 (untere Strings sind verschattet, die ganz unten am dramatischsten) Verschattung." Output: {"reason": "No reason given.", "categories": ["General"]}

A.2 Shadow Prediction after 4 Weeks

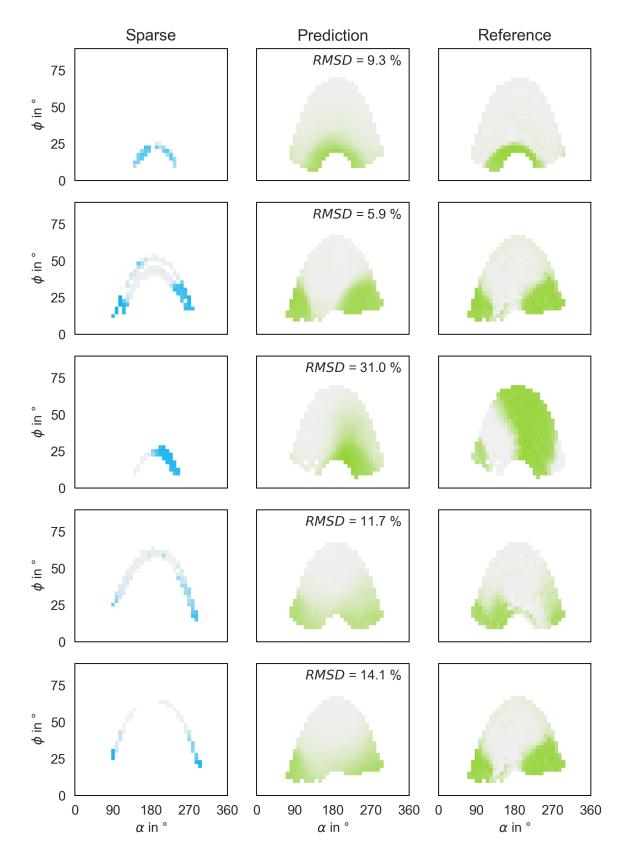


Figure 9. Shadow Prediction of the autoencoder after a detection time of 4 weeks. The dataset was filtered by using only datapoints with a simulated clear-sky power over $100 \frac{W}{kWp}$.

A.3 Shadow Prediction after 25 Weeks

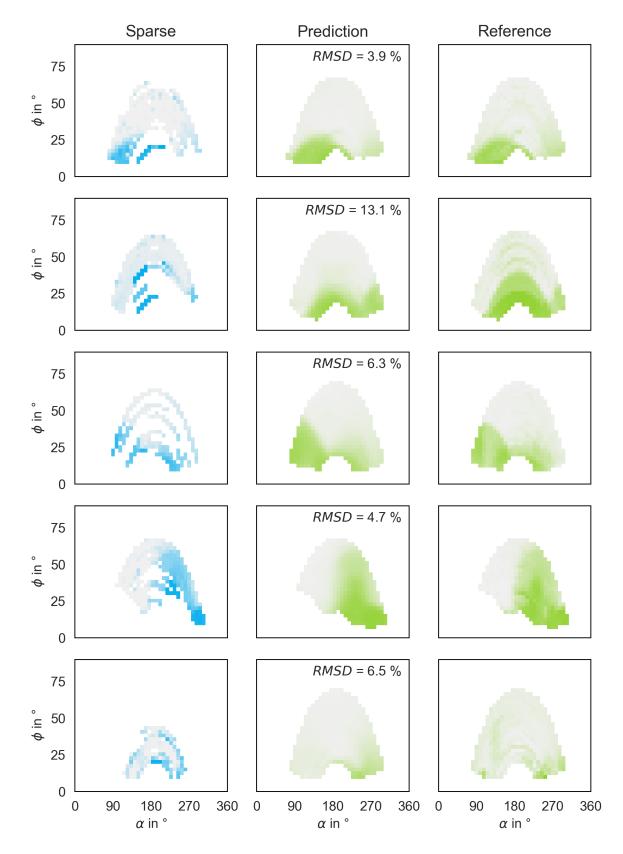


Figure 10. Shadow prediction of the autoencoder after a detection time of 25 weeks. The dataset was filtered by using only datapoints with a simulated clear-sky power over $100 \frac{w}{kWp}$.

Data availability statement

The datasets generated and analysed in this study as well as the parametrization of the methodology are proprietary to smartblue AG and are not publicly available due to competitive and confidentiality considerations. However, limited data may be made available upon request. For further information or requests for data access, please contact smartblue AG.

Author contributions

Maximilian Schönau: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft

- Joseph Jachmann: Investigation, Software, Visualization, Writing original draft
- Markus Panhuysen: Conceptualization, Data curation, Formal analysis, Methodology, Software, Supervision
- Alexander Schönau: Methodology, Software
- Darwin Daume: Methodology, Software
- Achim Schulze: Conceptualization, Data curation, Formal analysis, Methodology, Software, Supervision, Writing review & editing
- Bernd Hüttl: Funding acquisition, Project administration, Supervision, Writing review & editing
- Dieter Landes: Funding acquisition, Project administration, Supervision, Writing review & editing

Funding

The authors express their gratitude to the Bavarian Research Foundation for financial support of the project Kick-PV: Al-Based Characterization and Classification of PV Plants for Predictive Maintenance.

Acknowledgement

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG) and the Hitech Agenda Bavaria of the Free State of Bavaria.

References

- [1] S. Killinger *et al.*, "On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading," *Sol. Energy*, vol. 173, pp. 1087–1106, Oct. 2018, doi: 10.1016/j.solener.2018.08.051.
- [2] S. R. Pendem and S. Mikkili, "Modelling and performance assessment of PV array topologies under partial shading conditions to mitigate the mismatching power losses," *Sol. Energy*, vol. 160, pp. 303–321, Jan. 2018, doi: 10.1016/j.solener.2017.12.010.
- [3] M. Schönau *et al.*, "Reliable and Commercially Viable Detection of String Outages in Photovoltaic Plants," in *2024 International Conference on Renewable Energies and Smart Technologies (REST)*, Prishtina, Kosovo (UNMIK): IEEE, Jun. 2024, pp. 1–5. doi: 10.1109/REST59987.2024.10645480.
- [4] M. Schönau *et al.*, "String Outages in Photovoltaic Plants (Submitted for publication on February 28, 2025)," *Renew. Energ.*, Feb. 2025.
- [5] "Hello GPT-4o." Accessed: Feb. 06, 2025. [Online]. Available: https://openai.com/index/hello-gpt-4o/

- [6] "OpenAl Platform." Accessed: Feb. 06, 2025. [Online]. Available: https://platform.openai.com
- [7] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 01, 2023, *arXiv*: arXiv:1706.03762. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/1706.03762
- [8] S. Ghosh, J. N. Roy, and C. Chakraborty, "A model to determine soiling, shading and thermal losses from PV yield data," *Clean Energy*, vol. 6, no. 2, pp. 372–391, Apr. 2022, doi: 10.1093/ce/zkac014.
- [9] S. Sugumar, "A novel on-time partial shading detection technique for electrical reconfiguration in solar PV system," *Sol. Energy*, 2021, doi: 10.1016/j.solener.2021.07.069.
- [10] S. Fadhel, "PV shading fault detection and classification based on I-V curve using principal component analysis_ Application to isolated PV system," *Sol. Energy*, 2019, doi: 10.1016/j.solener.2018.12.048.
- [11] M. Schönau, D. Daume, B. Hüttl, and D. Landes, "Improving IV Curve Classification by Machine Learning Methods Using Deep Autoencoders," *40th Eur. Photovolt. Sol. Energy Conf. Exhib.*, pp. 020410-001-020410-004, 2023, doi: 10.4229/EUPVSEC2023/4CV.1.53.
- [12] A. Schulze, M. Panhuysen, D. Daume, and M. Schönau, "Quantitative Shade Detection for PV Systems Based on Clearsky Data," in *41th European Photovoltaic Solar Energy Conference and Exhibition*, Vienna, Sep. 2024., doi: 10.4229/EUPVSEC2024/4BV.3.29
- [13] Spatial distribution of daylight: luminance distributions of various reference skies. in Technical report / International Commission on Illumination, no. 110. Vienna: CIE Central Bureau, 1994.
- [14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, Art. no. 5786, Jul. 2006, doi: 10.1126/science.1127647.
- [15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion", doi: 10.5555/1756006.1953039.
- [16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
- [17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, *arXiv*: arXiv:1409.1556. Accessed: Sep. 18, 2024. [Online]. Available: http://arxiv.org/abs/1409.1556
- [18] M. Schönau, D. Daume, M. Panhuysen, A. Schulze, Hüttl, Bernd, and D. Landes, "Verbesserte Clear-Sky-Erkennung durch hybrides Maschinelles Lernen," presented at the RET.Con, Nordhausen, Feb. 2024. Accessed: Sep. 18, 2024. [Online]. Available: https://www.hs-nordhausen.de/fileadmin/Dateien/Veranstaltungen/RETCon 2024 Tagungsband.pdf