Solar Energy 249 (2023) 139–151

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Explainable artificial intelligence for photovoltaic fault detection: A comparison of instruments

Christian Utama [a], Christian Meske [b], Johannes Schneider [c], Rutger Schlatmann [d], Carolin Ulbrich [d,*]

[a] Department of Information Systems, Freie Universität Berlin, Germany
[b] Institute of Work Science and Faculty of Mechanical Engineering, Ruhr-Universität Bochum, Germany
[c] Department of Information Systems, Universität Liechtenstein, Germany
[d] PVcomB, Helmholtz-Zentrum Berlin, Germany

ARTICLE INFO

Keywords:
Photovoltaic fault detection
Machine learning
Artificial intelligence
XAI

ABSTRACT

Faults in photovoltaic arrays are known to cause severe energy losses. Data-driven models based on machine learning have been developed to automatically detect and diagnose such faults. A majority of the models proposed in the literature are based on artificial neural networks, which unfortunately represent black-boxes, hindering user interpretation of the models' results. Since the energy sector is a critical infrastructure, the security of energy supply could be threatened by the deployment of such models. This study implements explainable artificial intelligence (XAI) techniques to extract explanations from a multi-layer perceptron (MLP) model for photovoltaic fault detection, with the aim of shedding some light on the behavior of XAI techniques in this context. Three techniques were implemented: Shapley Additive Explanations (SHAP), Anchors and Diverse Counterfactual Explanations (DiCE), each representing a distinct class of local explainability techniques used to explain predictions. For a model with 99.11% accuracy, results show that SHAP explanations are largely in line with domain knowledge, demonstrating their usefulness to generate valuable insights on model behavior which could potentially increase user trust in the model. Compared to Anchors and DiCE, SHAP demonstrated a higher degree of stability and consistency.

1. Introduction

Over the last two decades, solar energy has emerged to be one of the most important renewable energy sources. In Germany, 45.4% of the country's gross electricity consumption in 2020 was produced by renewables. Solar energy's contribution amounts to 9.2%, summing up to 50 TWh (Wirth, 2021). Solar energy is captured and converted into electricity by photovoltaic (PV) modules, which are connected to one another in series or parallel to form a PV array. As with any physical system, faults could occur during the operation of a PV array, which might cause significant energy losses. These faults could occur either on the DC side, i.e., on the modules or arrays themselves, or on the AC side, involving the inverters or connection to the electricity grid (Madeti and Singh, 2017). Firth et al. (2010) found that faults could reduce the energy generated by PV arrays by as much as 18.9%. To ensure effective and reliable operation of PV arrays, fault detection methods have been developed and widely researched in recent years.

PV fault detection methods can be classified into two main categories: electrical and non-electrical methods (Tina et al., 2016). Due to their simplicity, electrical methods are preferred, as they rely only on the parameters typically recorded during operation, e.g., current and voltage. Out of the current electrical methods, only methods based on artificial intelligence (AI), hence machine learning (ML), could be implemented without the need of a PV simulation or model. Furthermore, once trained, AI models are computationally efficient and could be embedded in microcontrollers alongside sensors to create real-time monitoring and fault detection systems based on the concept of the Internet of Things (IoT) such as the ones proposed in Suresh et al. (2018) and Mellit et al. (2020). A thorough overview of the challenges, recommendations and future directions for research at the intersection of AI and IoT in the context of PV fault detection has been conducted by Mellit and Kalogirou (2021).

However, the use of AI comes with its own shortcomings, such as the lack of accompanying physical intuition. This is exacerbated by the use of complex, so-called "black-box" AI techniques such as artificial neural network (ANN), as it is virtually impossible to explain their predictions. Unfortunately, there is known to be a trade-off between the predictive

* Corresponding author.
E-mail address: carolin.ulbrich@helmholtz-berlin.de (C. Ulbrich).

https://doi.org/10.1016/j.solener.2022.11.018
Received 17 January 2022; Received in revised form 29 July 2022; Accepted 13 November 2022
Available online 28 November 2022

power of AI models and their complexity; in general, more complex models are able to capture more complex relationships and therefore have higher predictive power. A recent survey by Li et al. (2021) found that the vast majority of recent research papers concerning AI for PV fault detection use ANN. Although such experimental systems demonstrated high accuracies, it might not suffice for real-world deployment. A study by Dietvorst et al. (2015) found that humans tend not to trust AI systems and even more so when they make inexplicable errors, a phenomenon known as algorithm aversion. In real-world settings, AI systems should ideally be able to explain their predictions so that users could decide the best course of action and develop trust in the systems.

To get around the issue of AI models' black-box nature, explainable artificial intelligence (XAI) techniques have been developed. According to DARPA, XAI aims to "enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems". (Gunning and Aha, 2019) Note that the terms XAI and interpretable machine learning are used almost interchangeably in the research community and we chose the former based on our focus on evaluating already existing and well-known XAI techniques. In this study, we focus on the generation of explanations from black-box models, namely ANN for PV fault detection. In practice, it is often desired to explain how black-box models produce their outputs, which is why we chose to implement the so-called local, model agnostic explanation techniques in XAI terminology. We are interested in exploring whether the explanations generated by these techniques would align with domain knowledge and allow the user to judge whether to trust the AI model's predictions. We believe that such sanity checks are urgently needed since multiple explanation methods have been scrutinized for failing "sanity" checks and simple robustness properties (Adebayo et al., 2018; Kindermans et al., 2019; Ghorbani et al., 2019). At the same time, explanations would allow users of the model to conduct a root cause analysis in order to handle faults as quickly as possible, hence offering valuable operational insights.

This study implements and discusses the behavior of three local, model agnostic post-hoc explanation techniques: Shapley Additive Explanations (SHAP), Anchors and Diverse Counterfactual Explanations (DiCE) to generate explanations from an ANN-based PV fault detection system. These techniques are state-of-the-art techniques from three distinct classes of post-hoc techniques: feature importance, rule-based explanations and counterfactual explanations and they were selected as such here to provide compare and contrast the three classes. SHAP was chosen for its solid foundation on coalitional game theory and its robust performance in various applications (Brito et al., 2022; Kim et al., 2021; Kuzlu et al., 2020), making it arguably the go-to feature importance XAI technique. At the time of writing, Anchors remains the only rule-based XAI technique which could be found in the literature. DiCE was chosen due to its unique capability of generating diverse and constrained counterfactuals, therefore putting a focus on actionable explanations which are operationally meaningful (Mothilal et al., 2020). To the authors' best knowledge, this is the first study that touches on the subject of XAI for PV fault detection. In contrast to existing studies implementing a mix of feature importance and feature selection algorithms to improve the performance of fault detection models (Belaout et al., 2018; Eskandari et al., 2020), here we focus not on conducting a performance-driven analysis with XAI but rather on exploring if and how XAI techniques could potentially be useful. With this study, we aim to demonstrate how XAI techniques can be used to provide additional information to the users of ANN-based PV fault detection models and shed some light on the behavior of the selected techniques in this context. The evaluation of explainability methods has been explored in other domains (Kakogeorgiou and Karantzalos, 2021; Muddamsetty et al., 2021) and we would like to call upon a similar attention in this domain, as many open questions on the topic remain (Yang et al., 2019).

The rest of the paper is organized as follows: Section 2 presents an overview of the relevant studies found in the literature, Section 3 presents the methodology, Section 4 contains the results generated and the relevant discussion is presented in Section 5. Section 6 wraps up the study by presenting the conclusions drawn from the obtained results.

2. Background and related work

The proliferation of ANN implementation for PV fault detection in the last decade was highlighted in a recent survey by Li et al. (2021). The authors divided ANN applications into three groups based on their structure: shallow neural network (SNN), deep neural network (DNN) and hybrid neural network (HNN). The types of faults which could be detected by the different architectures also differ; it was found that SNNs were mostly implemented to detect electrical and shading faults, which are also the most commonly studied fault types in the literature. Therefore, this study focuses on SNNs for the purpose of detecting electrical and shading faults using PV operating parameters in the form of tabular data.

A thorough enumeration of recent SNN studies in the context of PV fault detection could be found in Li et al. (2021), Table 1. Here, we summarize and present several of the relevant studies to give a brief overview of the current landscape as shown in Table 1. All the studies listed on the table use multi-layer perceptron (MLP), i.e., a fully-connected feedforward ANN as the ML algorithm of choice. First of all, it could be seen that all the listed studies tackle the fault detection problem differently: different types of faults are detected with different sets of input features, although in all cases the measured voltage V_{mp} and current I_{mp} at the maximum power point are considered as inputs. The irradiance measured at the plane of array G_m and the module temperature T_m are also common features, whereas some studies additionally consider the open-circuit voltage V_{oc}, short-circuit current I_{sc}, fill factor FF and in Sabri et al. (2018), where the PV array is assumed to be equipped with a battery connected to the DC side of the inverter, the battery voltage V_b and current I_b. Furthermore, it could be seen that some studies validated their proposed model with experimental data, whereas others only utilized data gathered from simulation. In general, the accuracies of the proposed models are exceptionally high: above 90% in all cases, although they are somewhat lower in cases where validation with experimental data is performed. A straightforward comparison of the proposed models is not possible owing to the diverse setup of input features and detected faults. Nevertheless, MLP has been shown to be capable of detecting faults in PV systems with satisfactory performance, hence demonstrating its merits for the use case.

In the energy sector, research on XAI is only beginning to emerge, with only a few studies currently found in the literature. Chakraborty et al. (2021) built an XGBoost model to predict the cooling load of buildings using climate data. On top of the model, SHAP was used to generate local explanations for the predicted cooling loads. Arjunan et al. (2020) proposed an extension to the Energy Star benchmarking process, an energy performance rating system (Arjunan et al., 2020). A gradient boosted trees (GBT) model was built to replace the multiple linear regression model commonly used. To account for the increased model complexity, SHAP was implemented to interpret the model's output. Shams Amiri et al. (2021) implemented an ANN to classify household transportation energy consumption using both numerical and categorical household characteristics. Local Interpretable Model-Agnostic Explanations (LIME) and Submodular Pick (SP) LIME were implemented to generate local and global explanations, respectively. In both cases, the authors found that the generated explanations are in line with domain expertise. Kuzlu et al. (2020) implemented LIME, SHAP and Explain Like I'm 5 (ELI5) to generate explanations from a random forest regressor used to predict the power generated by a PV array. Results show that the most important features identified by the three XAI techniques are largely the same.

Based on their implementation principle, XAI techniques can be classified into two categories: transparent models and post-hoc techniques. Using simple, transparent AI models enables straightforward interpretation of the models' results, while post-hoc techniques aim to extract information from black-box AI models to generate explanations. Post-hoc techniques can further be classified based on their scope

Table 1
Summary of selected related work on SNN for PV fault detection.

Ref.	Input	Faults detected	Accuracy	Type of data
Chine et al. (2016)	Ratio between measured and simulated values of: V_{oc}, I_{mp}, V_{mp}	short-circuit, connection resistance, inversed bypass diode, shunted bypass diode	90.3%	simulation and experimental
Sabri et al. (2018)	V_{mp}, I_{mp}, V_b, I_b	short-circuit, open circuit, external battery short-circuit	97.4%	simulation and experimental
Ul-Haq et al. (2020)	V_{mp}, I_{mp}, P_{mp}, G_m, T_m	short-circuit, open circuit, partial shading, multiple faults	99.6%	simulation
Pahwa et al. (2020)	V_{mp}, I_{mp}, G_m, T_m, V_{oc}, I_{sc}, FF, $P_{mp,stc}$	short-circuit, inversed bypass diode, partial shading, complete shading, bridging, temperature	99.91%	simulation
Lazzaretti et al. (2020)	V_{mp}, I_{mp}, G_m, T_m	short-circuit, open circuit, partial shading, degradation	95.45%	simulation and experimental

(model agnostic or model specific) and nature (local or global). Model-agnostic approaches could be implemented to any AI model, while model-specific approaches are tailored only to a subset of models. Local explanations aim to explain observations individually, whereas global explanation entails explaining the behavior of the entire model. In this study, we focus on the implementation of local, model-agnostic post-hoc explanation techniques in the context of ANN-based PV fault detection. Local, model-agnostic post-hoc explainability techniques for tabular data can be classified into three categories (Barredo Arrieta et al., 2020): feature importance explanations, rule-based explanations and counterfactuals. Feature importance explanations aim to measure the importance of a model's inputs to its output. Meanwhile, rule-based explanations attempt to find sufficient conditions that lead to a certain output. Counterfactuals could be considered as what-if analyses and involves the generation of new, artificial observations using the instance to be explained as a starting point and finding minimum changes to the inputs that also change the predicted output.

In the literature, the behavior of local post-hoc explainability techniques is at present largely unexplored. Tritscher et al. (2020) evaluated eight post-hoc XAI techniques using synthetic categorical tabular data. All the evaluated techniques are feature importance techniques covering perturbation-based and gradient-based approaches. Results of the study show that under the study's settings, perturbation-based methods are superior to gradient-based ones. Out of the evaluated techniques, only SHAP was able to explain non-linear functions with up to three variables. Schlegel et al. (2019) proposed a methodology to evaluate XAI methods on time series data. LIME, SHAP, Layer-wise Relevance Propagation (LRP), DeepLIFT and Saliency Maps were implemented and evaluated, with the results showing SHAP to be the most robust method. In this study, we seek to evaluate the behavior of SHAP, Anchors and DiCE on a real tabular dataset in the context of PV fault detection.

Doshi-Velez and Kim (2017) proposed a taxonomy of approaches to evaluate explanations: functionally-grounded, human-grounded and application-grounded evaluations, in increasing order of complexity. Functionally-grounded evaluation assesses explanation quality through a proxy based on some formal definition of interpretability without user study and it is the approach taken in this study. Sokol and Flach (2020) developed a framework to systematically assess explainability techniques, in which validation is a requirement alongside functional, operational, usability and safety requirements. The authors refer to the work of Herman (2017) for a validation approach on a functional level, which evaluates the stability and consistency of explanations. Stable explanations concern the provision of the same explanation given a fixed set of inputs (data and model), whereas consistent explanations imply that similar explanations should be obtained for similar data points given a fixed model. We include stability and consistency evaluations for the selected XAI techniques to better understand their behavior.

3. Methodology

An overview of the methodology is presented in Fig. 1. Firstly, the raw dataset was preprocessed to remove uninformative data points. The preprocessed dataset was then used to train and test an MLP through hyperparameter tuning and cross-validation. From the MLP's predictions, several observations were selected, for which explanations were generated using three XAI techniques, namely SHAP, Anchors and DiCE as motivated in the introduction. Finally, the stability and consistency of the generated explanations from each technique were evaluated.

The dataset used in paper was originally used by Lazzaretti et al. (2020) to study ML-based PV fault detection systems and is available online (Clayton Hilgemberg da Costa, 2020). Complete information regarding the setup of the system and the dataset generation process could be found in the original publication. The dataset contains 16 days of data of a grid-connected PV array in Curitiba, Brazil. The PV array consists of 2 strings of 8 modules each and has a peak power of 5.28 kW. The strings are connected to a central inverter and each string is monitored individually. Most faults in the dataset were artificially introduced and therefore represent the ground truth. It is important to note that the faults were only present on the second string, i.e., the first string always operated normally. There are seven variables in the dataset, six of which are measured operating parameters of the array and are used as features in the ML model. Temperature and irradiance were measured by a weather station placed adjacent to the PV array. All the features take continuous values and are summarized in Table 2. The remaining variable is the target variable, the fault class f_nv which is a categorical variable and could take any of the following values: 0 (healthy operation), 1 (short-circuit fault), 2 (degradation fault), 3 (open circuit fault) and 4 (shading fault). Short-circuit faults were introduced by connecting a cable between the terminal points of two modules, degradation faults by connecting a resistive load between two modules, open circuit faults by opening one string's main circuit breaker and shading occurred naturally. All the scripts used to generate the results of this study are written in Python and are available online.[1]

[1] https://drive.google.com/file/d/1XGY5TVfs93OaENB-FRu1idzdyiZuZQlf/view?usp=sharing

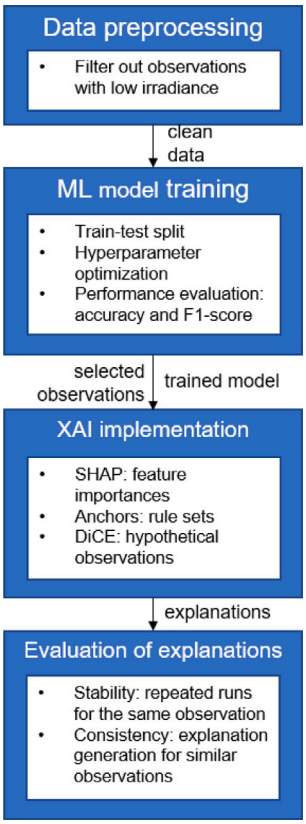

Fig. 1. Overview of the methodology.

Table 2
Features of the dataset.

Variable name	Description	Unit
$idc1$	Current at the first string	A
$idc2$	Current at the second string	A
$vdc1$	Voltage at the first string	V
$vdc2$	Voltage at the second string	V
pvt	Temperature of the array	°C
irr	Irradiance at the plane of array	$\frac{W}{m^2}$

3.1. Data preprocessing

A preprocessing step was implemented to remove uninformative observations from the initial dataset by eliminating observations with very low irradiance values, i.e., below 100 $\frac{W}{m^2}$. In the original dataset, faults were only introduced when the PV array was under high irradiance. Additionally, under very low irradiance, the PV array only produces a small amount of energy and therefore detecting faults under such condition would be less critical. Table 3 presents the number of observations belonging to each fault class before and after preprocessing. Although a significant number of observations were thrown away, it could be seen that these were mostly observations from the healthy operation class, which do not contain meaningful information and only contribute to class imbalance.

3.2. ML model training

For the purpose of PV fault detection, an MLP is built and trained using the dataset mentioned above. The problem is a multi-class classification problem with five possible classes. In this study, we conducted a grid search hyperparameter optimization to find the optimum number of hidden layers and neurons in the MLP using the ranges presented in Table 4. The maximum number of hidden layers and neurons were limited to 2 and 30, respectively, as increasing them further is not expected to improve the model's performance according to the results from existing studies (Li et al., 2021). This optimization was conducted via a 5-fold cross validation grid search covering 20 hyperparameter settings using 80% of the dataset. For every setting, 5 models were trained with different training sets consisting 64% of the dataset and 16% were used to calculate validation scores, which were then averaged to obtain the mean validation score. The remaining 20% of the dataset were used as test set to benchmark the performance of the best model from the output of the optimization. As the dataset is imbalanced, i.e., not all classes have the same number of observations, stratified sampling was utilized to ensure that the proportion of each fault class is constant in all subsets of the data (training/validation/test). In all cases, the hidden layers use Rectified Linear Unit (ReLU) as the activation function and the output layer uses the softmax activation function. Most of the other hyperparameters were kept at their default values from the scikit-learn implementation except for the maximum number of iterations and α (L2 regularization parameter), which were fixed at 500 and 0.001, respectively. Finally, the optimized model was evaluated for accuracy, precision, recall and F1-score, the formulas for which are given below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

TP and TN represent true positives and true negatives, respectively, while FP and FN denote false positives and false negatives, respectively. Since the problem in question is a multi-class classification problem, all performance metrics were calculated for each class, resulting in four scores per class.

Table 3
Summary of the observations in the dataset.

Fault class	Initial	After preprocessing
0 (healthy operation)	1,162,931	309,252
1 (short-circuit)	5,999	5,999
2 (degradation)	10,371	10,371
3 (open circuit)	6,024	6,024
4 (shading)	188,473	184,311
Total	**1,373,798**	**515,957**

Table 4
Ranges used for grid search.

No. of neurons	
First hidden layer	Second hidden layer
(15, 20, 25, 30)	(–, 15, 20, 25, 30)

3.3. XAI technique implementation

Once the MLP was trained and its predictions were obtained, explanations were generated using XAI techniques. For this purpose, one observation from each fault class was chosen for a qualitative assessment. These observations were chosen as such to represent the operation of the PV array under high irradiance and temperature. The observations should also be correctly classified by the MLP, i.e., the predicted and actual fault classes are the same. A summary of the observations can be found in Table 5.

In this study, SHAP, Anchors and DiCE were implemented. SHAP aims to explain predictions by computing the individual contributions

Table 5
Selected observations for explanation generation.

No.	$idc1$	$idc2$	$vdc1$	$vdc2$	irr	pvt	f_nv
1	7.63	7.61	271.19	270.54	878.6	43.01	0
2	7.98	7.62	260.65	201.93	836.07	47.31	1
3	8.26	7.6	253.51	242.12	842.14	49.93	2
4	8.17	0.04	264.32	2.09	856.13	43.06	3
5	8.17	4.87	263.14	306.15	853.5	43.79	4

of the features, in the form of Shapley values from coalitional game theory (Lundberg and Lee, 2017). In ML terms, the features of a model are the players, who "play" the game of reproducing the outcome of the model and Shapley values represent feature importance values. SHAP defines explanation as follows:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j, \tag{5}$$

where g is the explanation model, M is the number of features, $z'_j \in [0,1]^M$ are the simplified features, ϕ_0 is the base value of the prediction model (expected output value when no feature values are known) and ϕ_j is the Shapley value for feature j. The vector of simplified features z'_j contains only 0s and 1s, to indicate whether a certain feature is present ("playing") or absent ("not playing"), respectively. Given an observation x, its corresponding z'_j vector is a vector of all 1s, where all features are present.

The linear model is trained by minimizing the following loss function:

$$\min \ \mathcal{L}(f, g, \pi_x) = \sum_{z' \in Z} [f(h_x(z')) - g(z')]^2 \pi_x(z'), \tag{6}$$

where Z is the training data, f is the underlying black-box ML model, h_x is the mapping function which maps coalitions z's to the original input space and π_x is the SHAP kernel function. The process yields coefficients of the explanation model ϕ_j for all M features, which represent their feature importance values.

Anchors produces a set of rules which sufficiently "anchors" the prediction in the vicinity of an observation. Given an observation x, an anchor A is defined as the set of rules which applies to x as well as a fraction of at least τ of x's neighbors, for which the prediction of the underlying ML model f remains the same. τ is a precision threshold taking values of at least 0 and at most 1, where the precision is obtained by evaluating neighbors z following a certain distribution $D_x(z|A)$ using f.

In Anchors, a probabilistic definition of precision is used to construct rules in large input spaces as follows:

$$P(prec(A) \geq \tau) \geq 1 - \delta, \tag{7}$$

where δ represents the complement desired confidence level, e.g., a δ equal to 0.05 or 5% indicates a 95% confidence level.

The search for an anchor is defined by the following optimization problem:

$$\max_{A \ s.t \ (7)} v(A), \tag{8}$$

where the function v denotes the coverage function of the anchor, i.e. the probability that it applies to the sample neighbors drawn from D_x. Details on how the optimization problem could be solved are found in the original publication (Ribeiro et al., 2018). In this study, the precision threshold τ was set to 0.95.

DiCE produces counterfactuals, which are feature-perturbed versions of the original observations which result in a change of prediction. Counterfactuals are generated by solving an optimization problem, which in DiCE's case considers both the diversity of the generated counterfactuals as well as their proximity to the original observation. It is also possible to incorporate additional constraints to exclude certain features from being perturbed or to limit the range of values they could take. Starting with an initial observation x with M features and a trained ML model f, a set of k counterfactual examples $\{c_1, c_2, \ldots, c_k\}$ is generated by solving the following optimization problem:

$$\begin{aligned} \min \ \mathcal{L}(c_1, c_2, \ldots, c_k) = & \frac{1}{k} \sum_{i=1}^{k} loss(f(c_i), y) \\ & + \frac{\lambda_1}{k} \sum_{i=1}^{k} dist(c_i, x) \\ & - \lambda_2 div(c_1, \ldots, c_k), \end{aligned} \tag{9}$$

where c_i is a counterfactual, $loss$ is a metric that minimizes the distance between the underlying model f's prediction for the counterfactuals and the desired outcome/target class y, $dist$ is the distance metric representing the proximity between the counterfactuals and the original observation and div is a metric to measure the diversity of the counterfactuals. Detailed information on the implementation of $loss$, $dist$ and div in DiCE could be found in the original publication (Mothilal et al., 2020). λ_1 and λ_2 are hyperparameters that regularize the three components of the loss function and take the default values of 0.5 and 1, respectively. In this study, the features irr and pvt were excluded from the counterfactual generation process, meaning that only $idc1$, $idc2$, $vdc1$ and $vdc2$ were allowed to change. The intuition behind this is that in the context of PV fault detection, it would be useful to know what the ML model "thinks" as normal sets of operating parameters under the same external conditions to judge whether it is behaving as expected.

3.4. Evaluation of explanations

Stability and consistency were chosen as the metrics to evaluate the generated explanations. For stability evaluation, 50 runs of each technique on each observation were performed, followed up by a simple summary statistics calculation. In SHAP's case, the mean and standard deviation of the SHAP values of the features were calculated and presented in the form of box and whisker plots. DiCE evaluation follows a similar pattern, except that the deviations of the feature values from the original observation were used as the variable for summary statistics calculation. For Anchors, the frequencies of the rules generated from each run were calculated and tabulated. To evaluate consistency, 50 similar observations based on Euclidean distance were identified for each observation to obtain 250 additional observations. For each additional observation, explanations were generated using every XAI technique. Summary statistics were subsequently calculated for each original observation.

4. Results

In this section, performance of the proposed MLP for PV fault detection is presented, followed by a qualitative assessment of the explanations generated for the selected observations. Afterwards, results of the stability and consistency evaluations are presented.

4.1. ML model performance

Based on the results of hyperparameter optimization, the optimum number of neurons in the hidden layers was found to be (15, 30). Using this configuration, an optimized MLP was trained and its performance on the test set was recorded. The confusion matrix of the optimized MLP on the test set is presented in Table 6. Overall, the model recorded 99.11% accuracy across all observations and achieved an accuracy of at least 99% on all fault classes with the exception of class 4. One possible explanation for this is that shading faults on PV arrays are rather diverse in nature and could vary significantly from one occurrence to another. Precision, recall and F1-score of the model are summarized in Table 7. It could be seen that the model also performs well in terms

Table 6
Confusion matrix of the optimized MLP.

True class	Predicted class					Accuracy
	0	1	2	3	4	
0	61513	0	0	0	338	99.45%
1	0	1191	4	0	5	99.25%
2	0	1	2060	4	9	99.32%
3	0	0	0	1205	0	100%
4	526	16	19	1	36300	98.48%

Table 7
Additional performance metrics of the optimized MLP.

Class	Precision	Recall	F1-score
0	0.99	0.99	0.99
1	0.99	0.99	0.99
2	0.99	0.99	0.99
3	1	1	1
4	0.99	0.98	0.99

Table 8
Anchors explanation for the healthy observation.

Rule	Precision	Coverage
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 6.03$	0.98	0.18

of precision, recall and F1-score, suggesting that the dataset's class imbalance does not significantly impact its performance. In general, the proposed model's accuracy is in line with the values found in the literature (90 to 99%) and could therefore be considered as usable for PV fault detection.

4.2. Generated explanations

The following sections present and discuss the explanations generated by SHAP, Anchors and DiCE. SHAP and Anchors explanations are explained on a per-observation/per-class basis, while DiCE explanations are presented separately at the end.

4.2.1. Class 0 (healthy operation)

The SHAP explanation for the first observation is presented in Fig. 2. The features in red support the prediction made by the MLP, while the ones in blue oppose it. The size of the arrows represents the magnitude of the SHAP value, i.e., the feature importance value. It could be seen that the two most important features for the prediction are $idc1$ and $idc2$. This agrees with domain knowledge, which dictates that the operating current of a PV array is approximately linearly proportional to the amount of available irradiance, i.e., high irradiance implies high current under healthy operation. Since in all selected observations the PV array was exposed to high irradiance, the explanation provided by SHAP is able to capture this relationship. However, the interpretation of the third most important feature, irr, is not as straightforward. It could be argued that irr being highlighted as an important feature supports the argument that the MLP model captured the linear relationship between current and irradiance. However, since irradiance is an external variable and not an inherent operating parameter, it is difficult to judge its actual importance in the prediction. The same applies to pvt, whose SHAP value is small and therefore not shown in the figure. The sole opposing feature, $vdc1$, only has a small SHAP value and can therefore be disregarded.

Anchors explanation for the same observation is presented in Table 8. It is seen that for this observation, Anchors was able to find a rule set which captures the behavior of the MLP, recording a precision score of 0.98. The rule set can be interpreted as such: "when $vdc2$ is greater than 261.91 and less than or equal to 271.99 and $idc2$ is greater than 6.03, the PV array is most likely operating normally". Although this is certainly not a universally applicable rule, it makes sense to put this into context in the neighborhood around the prediction. For observations that are similar to this observation, i.e., under high irradiance and temperature, seeing the values of $vdc2$ and $idc2$ inside the intervals identified by Anchors should be convincing enough to show that the prediction is correct. As mentioned in Lazzaretti et al. (2020), the dataset was generated by artificially introducing faults on the second string of the array, meaning that fault signatures will only be detected on $vdc2$ and $idc2$. Anchors was able to isolate this fact, which might prove useful for model debugging.

Table 9
Anchors explanation for the short-circuit observation.

Rule	Precision	Coverage
$vdc2 \leq 261.91$ **AND** $idc2 > 3.08$ **AND** $idc1 > 7.77$ **AND** $731.25 < irr \leq 868.52$	0.30	0.01

Table 10
Anchors explanation for the degradation observation.

Rule	Precision	Coverage
$vdc2 \leq 261.91$ **AND** $idc2 > 6.03$ **AND** $vdc1 \leq 283.49$ **AND** $idc1 > 7.77$ **AND** $731.25 < irr \leq 868.52$	0.55	0.01

4.2.2. Class 1 (short-circuit)

Fig. 3 presents the SHAP explanation for the second observation, the PV array under short-circuit fault. According to the explanation, $vdc2$ has the highest positive contribution to the prediction. Its SHAP value (0.626) is significantly higher than that of the second most important feature, $idc1$ (0.177). For this observation, SHAP correctly highlights the anomalous feature, as the $vdc2$ value is significantly lower compared to healthy operation of the PV array under similar irradiance, i.e., 201.93 vs 270.54 V. This is in line with domain knowledge, in that short-circuit faults cause a number of PV modules to be bypassed and reduce the voltage generated from the entire array. The other features have significantly lower SHAP values and can therefore be regarded as unimportant to the prediction.

For the short-circuit observation, the Anchors explanation is provided in Table 9. In this case, it is evident that Anchors could not find a good rule set to approximate the behavior of the MLP judging by the low precision. Furthermore, 4 out of the 6 features were included in the rule set, producing a very specific rule set with very low coverage. This is a known issue with Anchors (Ribeiro et al., 2018), in that it sometimes produces overly specific rules. It can be concluded that for this observation, Anchors is not able to generate a good explanation.

4.2.3. Class 2 (degradation)

The SHAP explanation for the degradation observation is presented in Fig. 4. It could be seen that the most important feature for the prediction according to SHAP is $vdc2$, followed by $idc1$ and irr. Comparing $vdc2$ to $vdc1$ and $idc2$ to $idc1$, it could be seen that the operating parameters at the second string are slightly lower, although the difference on the voltage is more subtle. Degradation faults increase the internal resistance of PV modules, which in turn decrease both the optimum operating current and voltage. SHAP is able to highlight at least one of the anomalous operating parameters as a highly important feature, leading to an acceptable explanation.

Similar to the previous observation, Anchors failed to find a rule set which sufficiently explains the prediction, as shown by the result in Table 10. Again, in this case we see that the rule set generated is very specific, with low precision and coverage. For the degradation observation, no plausible explanation was found using Anchors.

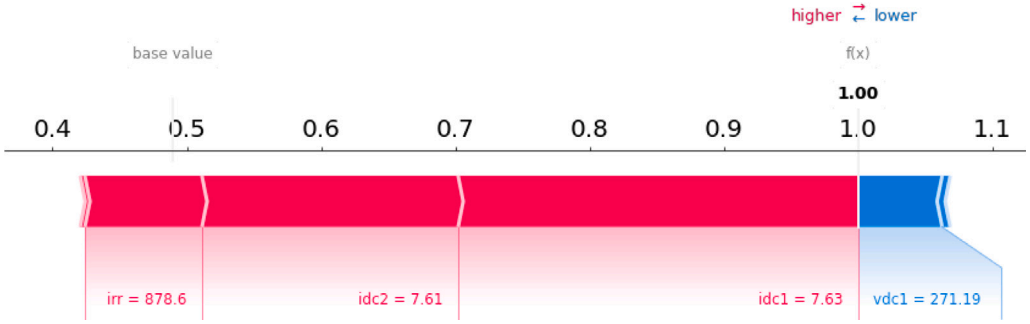

Fig. 2. SHAP explanation for the healthy observation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

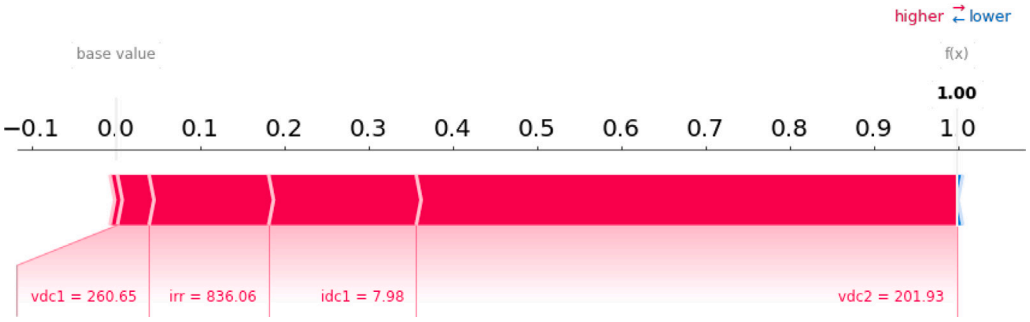

Fig. 3. SHAP explanation for the short-circuit observation.

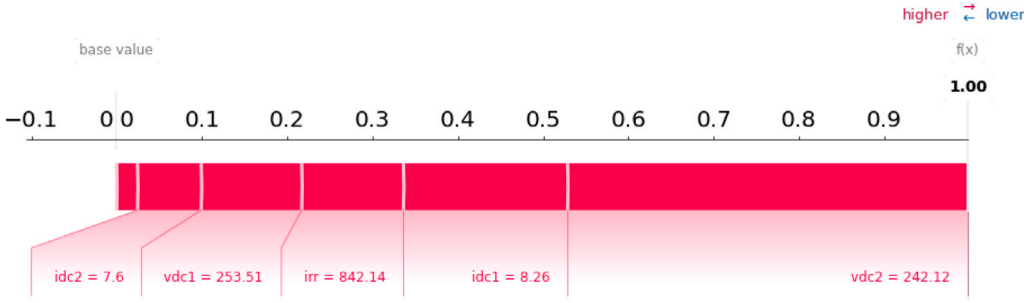

Fig. 4. SHAP explanation for the degradation observation.

Table 11
Anchors explanation for the open circuit explanation.

Rule	Precision	Coverage
$vdc2 \leq 261.91$ **AND** $idc2 \leq 3.08$ **AND** $vdc1 \leq 272.7$ **AND** $idc1 > 7.77$ **AND** $irr > 731.25$ **AND** $pvt > 39.28$	0.08	0.01

4.2.4. Class 3 (open circuit)

Fig. 5 illustrates the SHAP result for the open circuit fault. The two most important features identified are $vdc2$ and $idc2$, which take unusually low values close to zero compared to the healthy observation. According to domain knowledge, open circuit faults are caused by a loss of electrical connection between at least two PV modules in the main string, thus preventing electricity to pass and reducing the current and voltage of a PV array to zero. The result shows that SHAP correctly identified both key parameters and assigned high importance values to them. This makes for a logical and convincing explanation.

Table 11 presents the Anchors explanation for the open circuit observation. As is encountered with the short circuit and degradation observations, the generated rule set is overly specific, with all features being used to anchor the prediction. Poor precision and coverage values were also demonstrated by this huge rule set.

4.2.5. Class 4 (shading)

Fig. 6 presents the SHAP explanation for the shading observation. Compared to the other observations, it could be seen that the SHAP values for this observation are slightly more spread out, with 3 features supporting the prediction and another 3 opposing it. $idc2$ is considered the most important supporting feature, followed by $idc1$ and $vdc2$. Meanwhile, $vdc1$ is considered the most important opposing feature, albeit with a small SHAP value. For shading faults, determining the fault signature in the operating parameters is difficult since these faults occur naturally due to the presence of nearby buildings, trees, or other obstacles. Hence, they could vary in magnitude (how much irradiance is blocked by the obstacles) and scope (how many modules are shaded). Nevertheless, the features deemed important by SHAP are somewhat logical, as the values of $idc2$ and $vdc2$ do deviate a lot from the healthy observation. $idc1$, however, seems to be incorrectly identified as a supporting feature. Comparing the values of $idc1$ and $vdc1$ to those of the open circuit observation (observation 4 in Table 5), it could be seen that they are very similar and indicate healthy operation on the first string. Therefore, care needs to be taken in interpreting the SHAP explanation for shading fault.

The Anchors explanation for the shading observation is presented in Table 12. In contrary to the previous three observations, in this case Anchors managed to find a rule set which satisfactorily describes

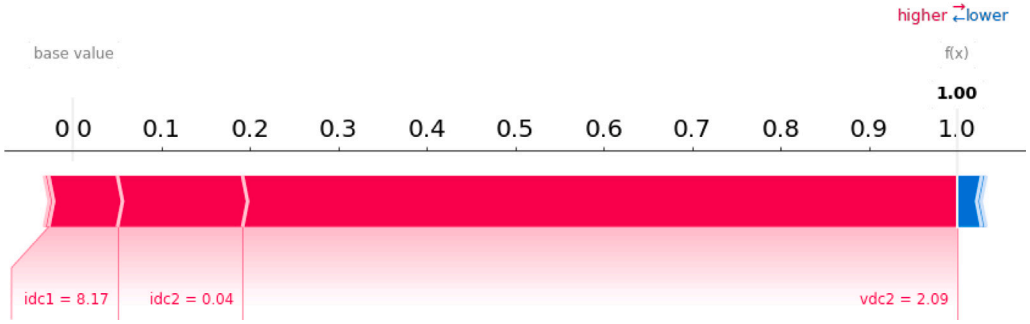

Fig. 5. SHAP explanation for the open circuit observation.

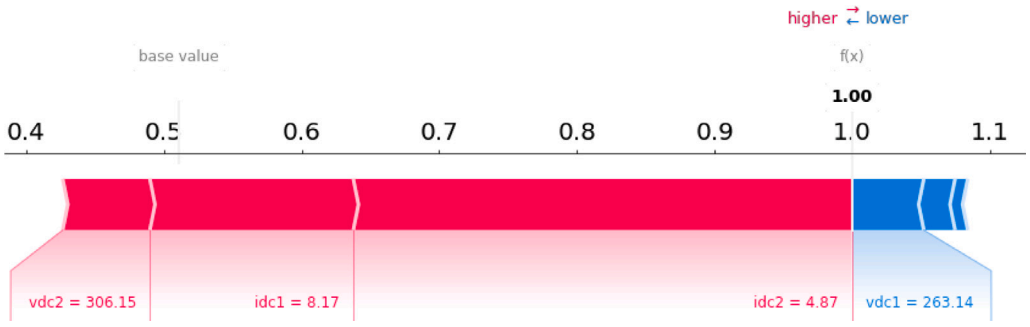

Fig. 6. SHAP explanation for the shading observation.

Table 12
Anchors explanation for the shading observation.

Rule	Precision	Coverage
$vdc2 > 285.71$ **AND** $idc2 \leq 6.03$ **AND** $idc1 > 6.54$	1	0.04

the behavior of the MLP in the vicinity of the observation. Half of the features are included in the rule set, which can be interpreted as follows: "when $vdc2$ is greater than 285.71, $idc2$ is less than or equal to 6.03 and $idc1$ is greater 6.54, the PV array is most likely operating under shading fault". Although the generated rule set might also seem somewhat specific, one possible explanation would be that this rule set only applies to one specific variation of shading faults, i.e. when only the second string is shaded and the first string is operating healthily. This is supported by the low coverage value obtained, which suggests that the rule set only applies to a very small subset of the dataset.

4.2.6. DiCE counterfactuals

A summary of selected counterfactuals for the healthy observation is presented in Table 13. The first row represents the original observation, whereas the rest are counterfactuals generated by DiCE, i.e., hypothetical observations. For all cases, one counterfactual with plausible explanation and another with meaningless explanation were selected. According to the results, it would suffice to either decrease $vdc2$ or increase $idc1$ and $vdc1$ in order to change the MLP's prediction from healthy (class 0) to short-circuit fault (class 1). While the first explanation makes sense and agrees with domain knowledge, the same does not apply to the second, as it implies that increased energy production would result in the detection of a short-circuit fault. In general, the same behavior is observed across all classes, i.e. at least one of the counterfactuals provides a reasonable explanation and another produces a confounding explanation.

4.3. Stability evaluation

The results of the stability evaluation for SHAP are presented in Fig. 7. It could be seen that, judging by the very thin spread of the

Table 13
Generated counterfactuals for the healthy observation (– indicates no change in the feature value).

No.	$idc1$	$idc2$	$vdc1$	$vdc2$	irr	pvt	f_nv
1	7.63	7.61	271.19	270.54	878.6	43.01	0
2	–	–	–	**219.37**	–	–	**1**
3	**9.11**	–	**324.92**	–	–	–	**1**
4	–	–	–	**124.3**	–	–	**2**
5	**6.71**	–	–	**124.3**	–	–	**2**
6	–	–	**305.97**	**3.45**	–	–	**3**
7	–	**1.31**	**3.26**	–	–	–	**3**
8	–	**1.43**	–	**220.91**	–	–	**4**
9	–	–	–	**322.56**	–	–	**4**

SHAP values across all classes, SHAP produces very stable explanations. In fact, the exact same SHAP values are obtained for all features in all selected observations in this study. This could be attributed to the fact that SHAP is theoretically well-grounded in game theory. Although SHAP requires the generation of artificial coalitions when calculating SHAP values, meaning that there is some stochastic element attached to the process, the same outcome will always be obtained given enough coalitions. With the default configuration of $2048 + 2n_{samples}$ coalitions, SHAP always attributes the same importance values to all features across different runs.

For Anchors, only explanations for the healthy and shading fault observations were evaluated due to the fact that Anchors did not produce valid explanations for the other observations. A summary of the Anchors explanations for stability evaluation is provided in Tables 14 and 15. It could be observed that in both cases, several different rule sets were generated across multiple runs. Although for each observation one certain rule set was generated at least 50% of the time, the generated explanations are still somewhat unstable. This is evident from the variation in the features selected in the rule sets and the threshold values for these features.

In DiCE's case, the results of the stability evaluation could be found in Fig. 8. It could be seen that for a given observation, DiCE produces diverse explanations, shown by the rather large spreads of the feature

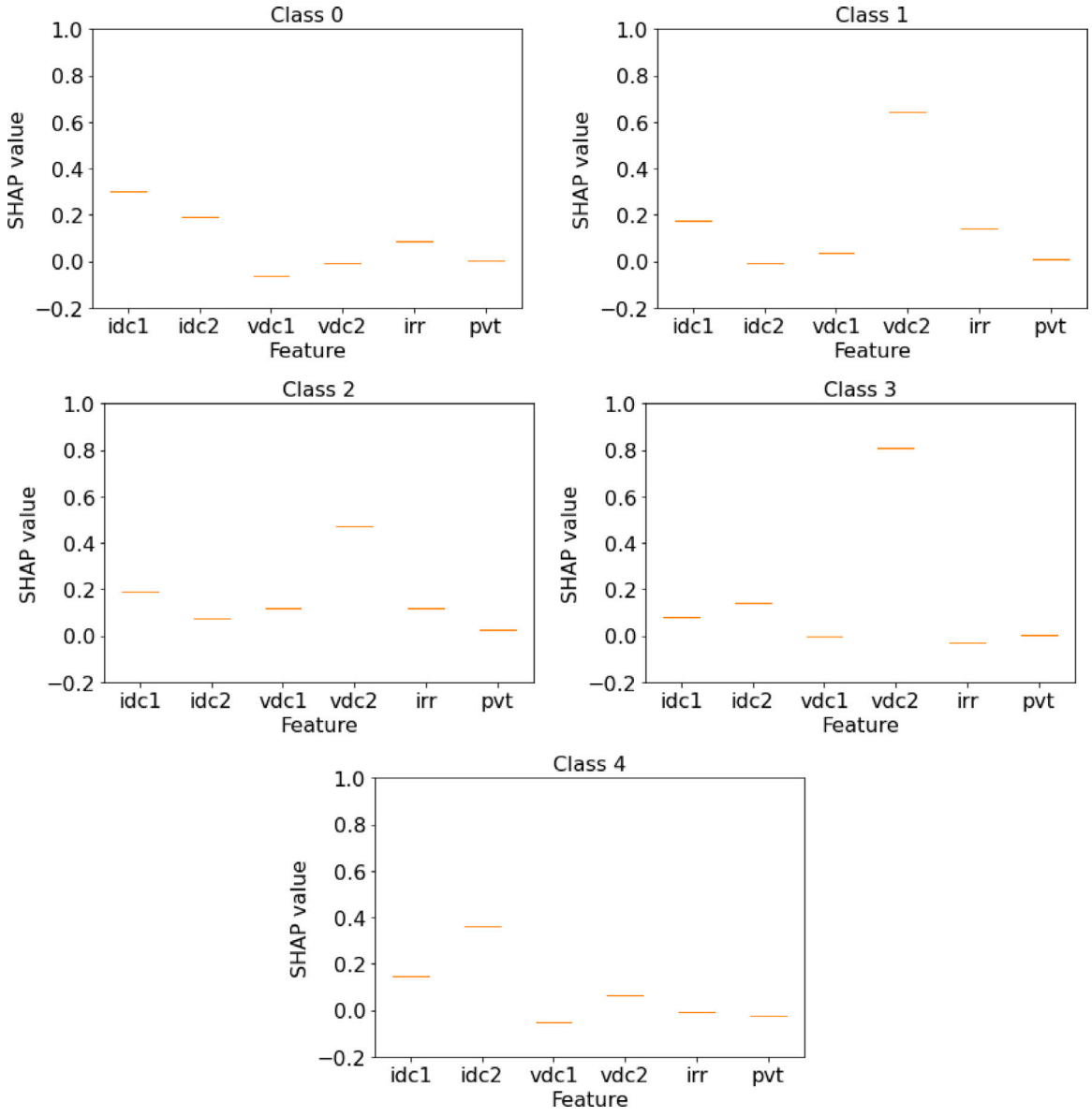

Fig. 7. Stability evaluation results for SHAP.

Table 14
Stability evaluation results for Anchors (healthy observation).

Rule	Count
$vdc2 > 261.91$ **AND** $idc2 > 7.56$ **AND** $vdc1 > 264.04$	26
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 6.03$	12
$vdc2 > 261.91$ **AND** $idc2 > 7.56$ **AND** $pvt \leq 45.59$	3
$vdc2 > 261.91$ **AND** $264.04 < vdc1 \leq 272.7$ **AND** $idc2 > 6.03$	3
$261.91 < vdc2 \leq 285.71$ **AND** $idc2 > 6.03$	2
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 7.56$	2
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 6.03$ **AND** $vdc1 \leq 272.7$	1
$vdc2 > 261.91$ **AND** $idc2 > 7.56$	1

Table 15
Stability evaluation results for Anchors (shading observation).

Rule	Count
$vdc2 > 285.71$ **AND** $idc2 \leq 6.03$ **AND** $idc1 > 6.54$	29
$vdc2 > 271.99$ **AND** $idc2 \leq 6.03$ **AND** $idc1 > 6.54$	15
$vdc2 > 261.91$ **AND** $idc2 \leq 6.03$ **AND** $idc1 > 6.54$	5
$idc2 \leq 6.03$ **AND** $irr > 731.25$ **AND** $pvt \leq 45.59$	1

value deviations. By nature, DiCE generates diverse counterfactuals, taking into account its objective of providing different ways to change the underlying ML model's prediction. Nevertheless, in some cases, the generated explanations are somewhat stable and in line with what is expected from domain knowledge (class 0 to 1 and class 0 to 2). In other cases, the generated explanations are either confounding (class 0 to 3) or downright unstable (class 0 to 4).

4.4. Consistency evaluation

The consistency evaluation results for SHAP are presented in Fig. 9. In line with the results of the stability evaluation, SHAP also performs well with regard to the consistency of its explanations. It could be observed that although there is some spread in the SHAP values obtained for the identified similar observations, it is relatively narrow and indicates consistent explanations. More importantly, for each class, almost none of the SHAP value intervals of the features overlap with one another, which means that the order of importance of the features identified by SHAP is almost always guaranteed to be the same. Therefore, it could be concluded that SHAP could be relied upon to generate similar explanations for similar observations.

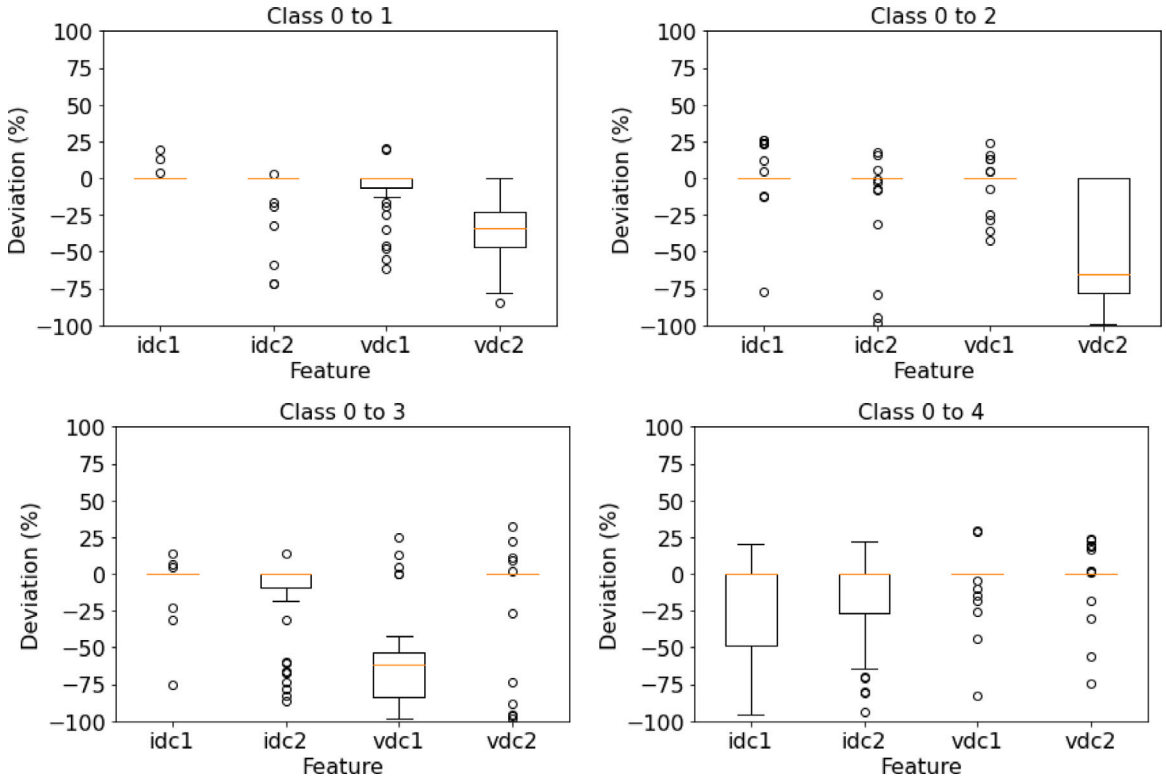

Fig. 8. Stability evaluation results for DiCE.

Table 16
Consistency evaluation results for Anchors (healthy observation).

Rule	Count
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 6.03$	49
$vdc2 > 261.91$ **AND** $idc2 > 7.56$ **AND** $vdc1 > 264.04$	1

Table 17
Consistency evaluation results for Anchors (shading observation).

Rule	Count
$vdc2 > 285.71$ **AND** $idc2 \leq 6.03$ **AND** $idc1 > 6.54$	50

Tables 16 and 17 summarize the rule sets generated by Anchors for the consistency test. It could be seen that for both the healthy and the shading fault observations, a unique rule set could be identified. In the latter's case, all similar observations yielded a single rule set, whereas in the former's case a different rule set was obtained for only a single observation.

Fig. 10 illustrates the results DiCE's consistency evaluation. It could be observed that the results largely mirror the results of the stability test shown in Fig. 8. Hence, the same conclusion could be drawn regarding DiCE's consistency, i.e., in some instances it produces sensible explanations and in other times meaningless explanations are obtained.

5. Discussion and implications

Based on the results obtained, we synthesized a summary of the strengths and limitations of the implemented XAI techniques. SHAP's main strength lies in its ability to correctly identify and attribute high importance values to relevant features for the underlying ML model's predictions, which leads to sensible explanations that agree with domain knowledge. The results of stability and consistency evaluations show that SHAP produces the most stable and consistent explanations. On the other hand, SHAP-generated explanations are of lesser value when external variables are used as features, as the interpretation of the feature importance values of such features is non-intuitive.

In Anchors' case, the generated explanations are contrastive, and thus intuitive to users. With regard to its consistency, Anchors records a relatively satisfying performance, being able to produce a single explanation for similar observations almost all the time. A somewhat lesser performance was observed in the stability evaluation, although the generated explanations are still arguably rather stable. However, implementing Anchors comes with a huge caveat, in that it sometimes produces overly specific, and thus invalid rules. Fortunately, it is generally easy to identify invalid rules by simply looking at their precision and coverage values.

In contrast to SHAP and Anchors, DiCE is the only technique evaluated to offer the possibility of constraining the generated explanations. Moreover, DiCE's way of providing counterfactuals as explanations also allows for contrastive and intuitive explanations, as it is easy to see the changes in the feature values that also lead to a change in prediction. Most importantly, DiCE offers the possibility of highlighting potential faults in the ML model's reasoning, which might be valuable for model debugging. However, it also causes DiCE to produce relatively unstable and inconsistent explanations, which is somewhat expected given the nature of DiCE's algorithm. Additionally, some of the generated counterfactuals might refer to conditions which are infeasible, leading to physically meaningless explanations. Lastly, no formal measure to determine good counterfactuals currently exists. Since multiple counterfactuals could be generated, presenting all of them might confound users, especially if some explanations contradict one another.

According to a study by Guidotti (2021), local explanation techniques fail to find good explanations when many features are relevant to the underlying ML model's predictions. In the context of PV fault detection, our results show that Anchors and DiCE demonstrate this behavior even when there are only one or two relevant features. However, SHAP was shown to consistently find good explanations, in agreement with another finding of the study which states that SHAP records the best performance and the lowest deviation with regard to the explanation qualities among the evaluated techniques. Anchors' problem of

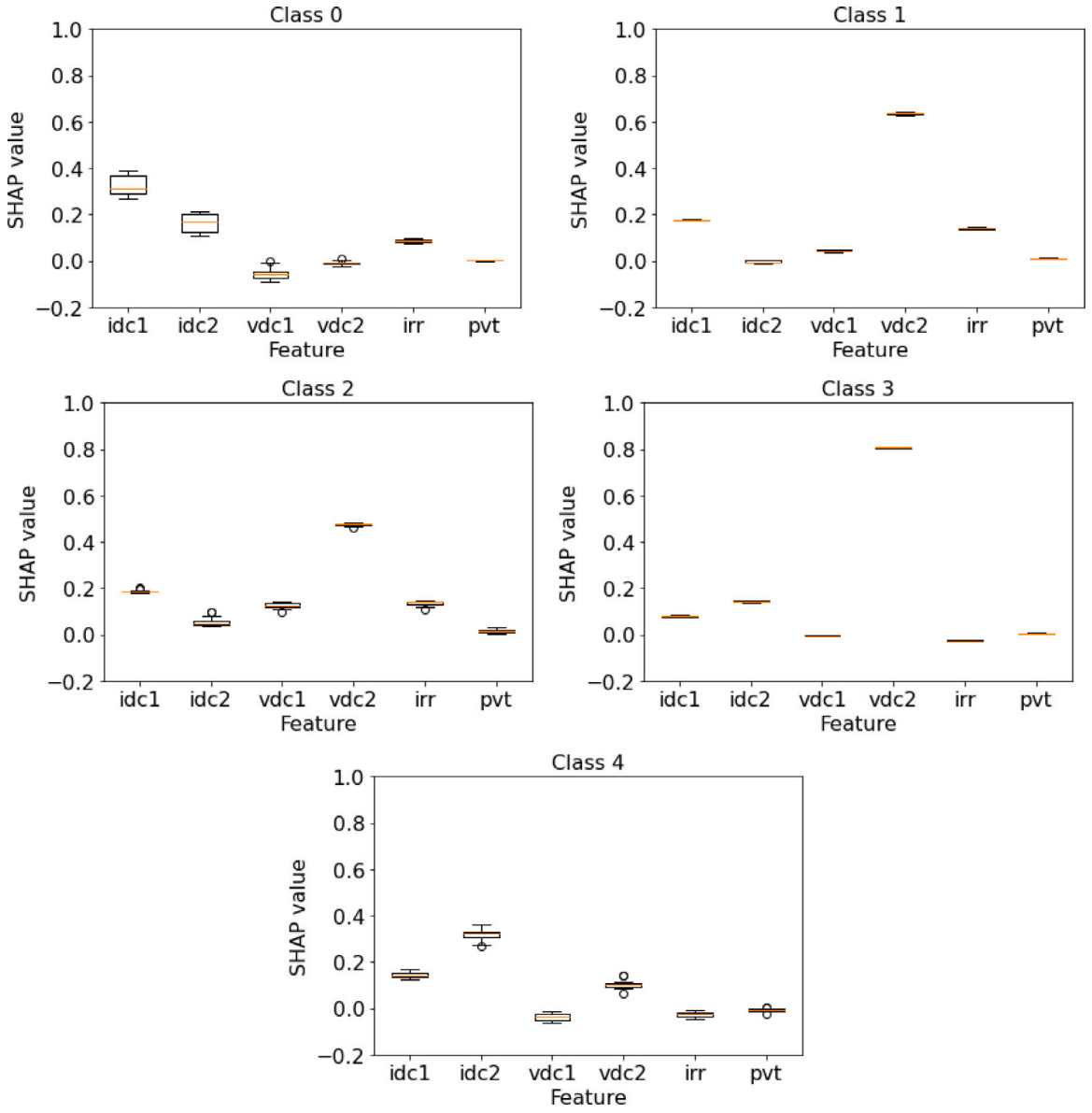

Fig. 9. Consistency evaluation results for SHAP.

producing overly specific anchors is well-documented (Ribeiro et al., 2018). It is stated that this behavior is expected for observations close to the decision boundaries of the ML model. This might be the case with the short-circuit, degradation and open circuit observations, for which no valid anchors were found. DiCE's limitation of potentially generating infeasible explanations is acknowledged by the original authors (Mothilal et al., 2020). One possible solution envisioned by the authors is to incorporate causal constraints during the generation of counterfactuals. In the context of PV fault detection, this might make DiCE a powerful tool when complemented with domain knowledge, e.g. I–V curve of the PV modules describing how the current and voltage should change alongside each other.

Although the results shown in this study are only based on a case study of a small PV system with artificially induced faults, it would be prudent to consider the potential usefulness of XAI explanations in a real-world setting. For example, considering a large system with 100 or more strings and individual string monitoring, SHAP explanations would allow the identification of the strings under fault and allow operators to act in a timely manner. As mentioned, DiCE explanations are potentially useful for model debugging and improvement, which gets harder with increasing system size. Admittedly, Anchors explanations might be too complex for a large system given the number of variables involved and they might therefore hardly be useful. In general, neither Anchors nor DiCE seems to be particularly suited to generating explanations for end-users, although DiCE might offer useful insights for ML model developers in the same vein as global feature importance techniques. Hence, in future work, we aim to further evaluate SHAP and DiCE with a dataset from a significantly larger PV system containing real (i.e., not artificially induced) faults in two different contexts: (1) understanding and trusting the AI model from the user's point of view and (2) using XAI explanations to improve model performance.

Considering the evidence gathered in this study, it could be argued that from the user's point of view, SHAP should be considered as the go-to XAI technique to generate explanations from an ANN-based PV fault detection system. This stems from the fact that in almost all of the cases, the most relevant features for the predictions were correctly identified and there is little to no variation in terms of the generated explanations, leading to straightforward interpretation. This claim has to be confirmed with a user study involving potential users of the PV

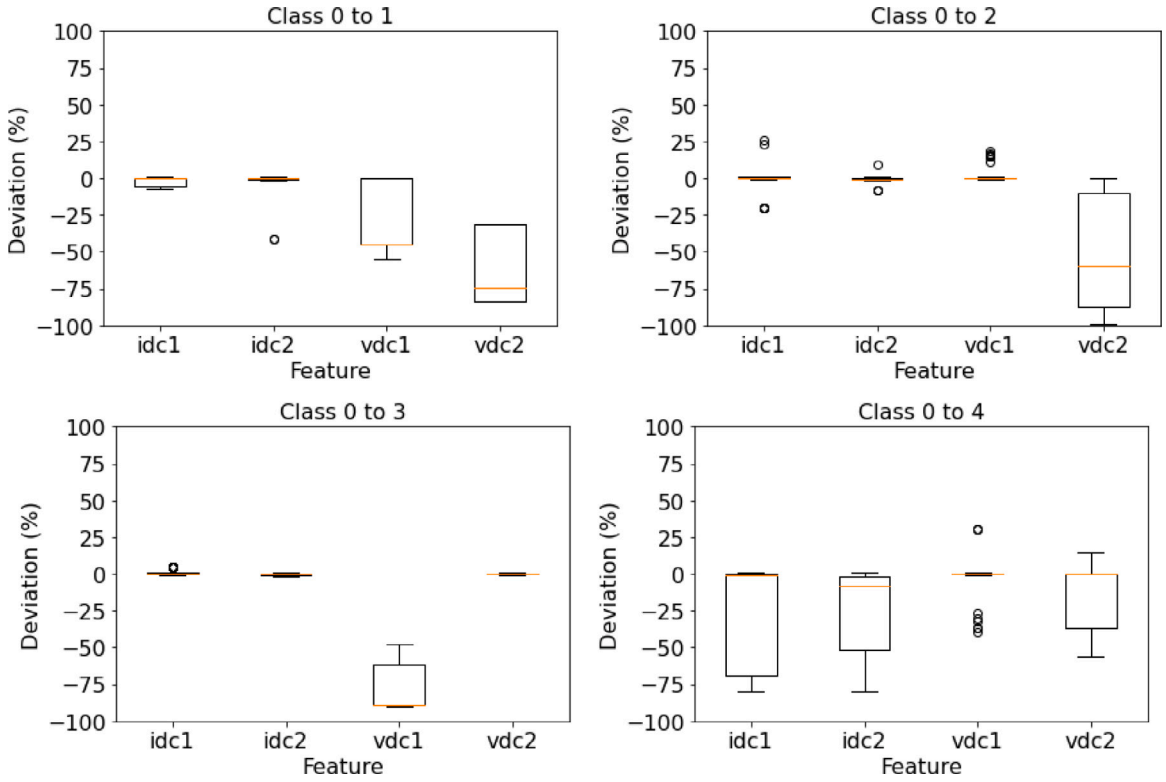

Fig. 10. Consistency evaluation results for DiCE.

fault detection system with an explanation interface, which we leave for future work.

6. Conclusions

In this study, we demonstrate how XAI instruments generate explanations in the context of PV fault detection. The behaviors of SHAP, Anchors and DiCE were explored. It is shown that in all cases, SHAP correctly attributed the prediction to the relevant features as dictated by domain knowledge. Meanwhile, Anchors was able to produce sensible rule sets only for some of the observations. DiCE, due to its diverse nature, produced highly varying explanations which might be contradictive but could also be useful for model debugging. With regard to the stability and consistency of the generated explanations, SHAP recorded the best performance, followed by Anchors and lastly DiCE. The results also show that stability and consistency are highly correlated in the context of local explanation techniques, i.e. a technique which generates stable explanations is likely to also be consistent and vice-versa.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS), Germany. This work was supported by the Helmholtz Association, Germany under the program “Energy System Design”.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. In: Neural Information Processing Systems.

Arjunan, P., Poolla, K., Miller, C., 2020. EnergyStar++: Towards more accurate and explanatory building energy benchmarking. Appl. Energy 276, 115413. http://dx.doi.org/10.1016/j.apenergy.2020.115413.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012.

Belaout, A., Krim, F., Mellit, A., Talbi, B., Arabi, A., 2018. Multiclass adaptive neuro-fuzzy classifier and feature selection techniques for photovoltaic array fault detection and classification. Renew. Energy 127, 548–558. http://dx.doi.org/10.1016/j.renene.2018.05.008.

Brito, L.C., Susto, G.A., Brito, J.N., Duarte, M.A., 2022. An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. Mech. Syst. Signal Process. 163, 108105. http://dx.doi.org/10.1016/j.ymssp.2021.108105.

Chakraborty, D., Alam, A., Chaudhuri, S., Başağaoğlu, H., Sulbaran, T., Langar, S., 2021. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. Appl. Energy 291, 116807. http://dx.doi.org/10.1016/j.apenergy.2021.116807.

Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., Massi Pavan, A., 2016. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renew. Energy 90, 501–512. http://dx.doi.org/10.1016/j.renene.2016.01.036.

Clayton Hilgemberg da Costa, 2020. Photovoltaic fault dataset. URL https://github.com/clayton-h-costa/pv_fault_dataset.

Dietvorst, B.J., Simmons, J.P., Massey, C., 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. J. Exp. Psychol. [Gen.] 144 (1), 114. http://dx.doi.org/10.1037/xge0000033.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. http://dx.doi.org/10.48550/ARXIV.1702.08608.

Eskandari, A., Milimonfared, J., Aghaei, M., Reinders, A.H., 2020. Autonomous monitoring of line-to-line faults in photovoltaic systems by feature selection and parameter optimization of support vector machine using genetic algorithms. Appl. Sci. 10 (16), 5527. http://dx.doi.org/10.3390/app10165527.

Firth, S., Lomas, K., Rees, S., 2010. A simple model of PV system performance and its use in fault detection. Sol. Energy 84 (4), 624–635. http://dx.doi.org/10.1016/j.solener.2009.08.004, International Conference CISBAT 2007.

Ghorbani, A., Abid, A., Zou, J., 2019. Interpretation of neural networks is fragile. In: AAAI Conference on Artificial Intelligence. http://dx.doi.org/10.1609/aaai.v33i01.33013681.

Guidotti, R., 2021. Evaluating local explanation methods on ground truth. Artificial Intelligence 291, 103428.

Gunning, D., Aha, D.W., 2019. DARPA's explainable artificial intelligence program. AI Mag. 40 (2).

Herman, B., 2017. The promise and peril of human evaluation for model interpretability. In: NIPS2017.

Kakogeorgiou, I., Karantzalos, K., 2021. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. Int. J. Appl. Earth Obs. Geoinf. 103, 102520. http://dx.doi.org/10.1016/j.jag.2021.102520.

Kim, D., Antariksa, G., Handayani, M.P., Lee, S., Lee, J., 2021. Explainable anomaly detection framework for maritime main engine sensor data. Sensors 21 (15), http://dx.doi.org/10.3390/s21155200.

Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B., 2019. The (un) reliability of saliency methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.

Kuzlu, M., Cali, U., Sharma, V., Güler, Ö., 2020. Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access 8, 187814–187823. http://dx.doi.org/10.1109/ACCESS.2020.3031477.

Lazzaretti, A.E., Costa, C.H.d., Rodrigues, M.P., Yamada, G.D., Lexinoski, G., Moritz, G.L., Oroski, E., Goes, R.E.d., Linhares, R.R., Stadzisz, P.C., Omori, J.S., Santos, R.B.d., 2020. A monitoring system for online fault detection and classification in photovoltaic plants. Sensors 20 (17), http://dx.doi.org/10.3390/s20174688.

Li, B., Delpha, C., Diallo, D., Migan-Dubois, A., 2021. Application of artificial neural networks to photovoltaic fault detection and diagnosis: A review. Renew. Sustain. Energy Rev. 138, 110512. http://dx.doi.org/10.1016/j.rser.2020.110512.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc..

Madeti, S.R., Singh, S., 2017. A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. Sol. Energy 158, 161–185. http://dx.doi.org/10.1016/j.solener.2017.08.069.

Mellit, A., Hamied, A., Lughi, V., Pavan, A.M., 2020. A low-cost monitoring and fault detection system for stand-alone photovoltaic systems using IoT technique. In: ELECTRIMACS 2019. Springer, pp. 349–358. http://dx.doi.org/10.1007/978-3-030-37161-6_26".

Mellit, A., Kalogirou, S., 2021. Artificial intelligence and internet of things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: Challenges, recommendations and future directions. Renew. Sustain. Energy Rev. 143, 110889. http://dx.doi.org/10.1016/j.rser.2021.110889.

Mothilal, R.K., Sharma, A., Tan, C., 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. In: FAT* '20, Association for Computing Machinery, New York, NY, USA, pp. 607–617. http://dx.doi.org/10.1145/3351095.3372850.

Muddamsetty, S.M., Jahromi, M.N., Moeslund, T.B., 2021. Expert level evaluations for explainable AI (XAI) methods in the medical domain. In: International Conference on Pattern Recognition. Springer, pp. 35–46. http://dx.doi.org/10.1007/978-3-030-68796-0_3.

Pahwa, K., Sharma, M., Saggu, M.S., Kumar Mandpura, A., 2020. Performance evaluation of machine learning techniques for fault detection and classification in PV array systems. In: 2020 7th International Conference on Signal Processing and Integrated Networks. SPIN, pp. 791–796. http://dx.doi.org/10.1109/SPIN48934.2020.9071223.

Ribeiro, M.T., Singh, S., Guestrin, C., 2018. Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (no. 1), http://dx.doi.org/10.1609/aaai.v32i1.11491.

Sabri, N., Tlemçani, A., chouder, A., 2018. Faults diagnosis in stand-alone photovoltaic system using artificial neural network. In: 2018 6th International Conference on Control Engineering & Information Technology. CEIT, pp. 1–6. http://dx.doi.org/10.1109/CEIT.2018.8751924.

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A., 2019. Towards a rigorous evaluation of XAI methods on time series. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop. ICCVW, pp. 4197–4201. http://dx.doi.org/10.1109/ICCVW.2019.00516.

Shams Amiri, S., Mottahedi, S., Lee, E.R., Hoque, S., 2021. Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption. Comput. Environ. Urban Syst. 88, 101647. http://dx.doi.org/10.1016/j.compenvurbsys.2021.101647.

Sokol, K., Flach, P., 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. In: FAT* '20, Association for Computing Machinery, New York, NY, USA, pp. 56–67. http://dx.doi.org/10.1145/3351095.3372870.

Suresh, M., Meenakumari, R., Kumar, R.A., Raja, T.A.S., Mahendran, K., Pradeep, A., 2018. Fault detection and monitoring of solar PV panels using internet of things. Int. J. Ind. Eng. 2 (6), 146–149.

Tina, G.M., Cosentino, F., Ventura, C., 2016. Monitoring and diagnostics of photovoltaic power plants. In: Renewable Energy in the Service of Mankind Vol II: Selected Topics from the World Renewable Energy Congress. WREC 2014, pp. 505–516. http://dx.doi.org/10.1007/978-3-319-18215-5_45.

Tritscher, J., Ring, M., Schlr, D., Hettinger, L., Hotho, A., 2020. Evaluation of post-hoc XAI approaches through synthetic tabular data. In: Foundations of Intelligent Systems. pp. 422–430.

Ul-Haq, A., Sindi, H.F., Gul, S., Jalal, M., 2020. Modeling and fault categorization in thin-film and crystalline PV arrays through multilayer neural network algorithm. IEEE Access 8, 102235–102255. http://dx.doi.org/10.1109/ACCESS.2020.2996969.

Wirth, H., 2021. Recent Facts about Photovoltaics in Germany. Technical Report, Fraunhofer ISE.

Yang, F., Du, M., Hu, X., 2019. Evaluating explanation without ground truth in interpretable machine learning. http://dx.doi.org/10.48550/ARXIV.1907.06831.