PV shading fault detection and classification based on I-V curve using principal component analysis: Application to isolated PV system

Siwar Fadhel, Claude Delpha, Demba Diallo, I. Bahri, Anne Migan-Dubois, Mohamed Trabelsi, Mohamed Faouzi Mimouni

PV Shading Fault Detection and Classification Based on I-V Curve Using Principal Component Analysis: Application to Isolated PV System

S. Fadhel[1,2,4], C. Delpha[3], D. Diallo[2], I. Bahri[2], A. Migan[2], M. Trabelsi[4], M.F. Mimouni[4]

[1]ENISO, BP 264 Sousse Erriadh 4023, Univ. Sousse, Tunisia

[2]GeePs | Group of electrical engineering-Paris, CNRS, CentraleSupélec, Univ. Paris-Sud, Univ. Paris-Saclay, Sorbonne Université, 3 & 11 rue Joliot-Curie, Plateau de Moulon 91192 Gif-sur-Yvette CEDEX, France

[3]L2S | Laboratoire des Signaux et Systèmes, CNRS, CentraleSupélec, Univ. Paris-Sud, Univ. Paris-Saclay, 3 rue Joliot-Curie, Plateau de Moulon 91192 Gif-sur-Yvette CEDEX, France

[4]LASEE | Laboratoire d'Automatique, des Systèmes Électriques et d'Environnement, ENIM, 5000 rue Ibn El Jazzar, 5035 Monastir, Univ. Monastir, Tunisia

Abstract

Health monitoring and diagnosis of photovoltaic (PV) systems is becoming crucial to maximise the power production, increase the reliability and life service of PV power plants. Operating under faulty conditions, in particular under shading, PV plants have remarkable shape of current-voltage (I-V) characteristics in comparison to reference condition (healthy operation). Based on real electrical measurements (I-V), the present work aims to provide a very simple, robust and low cost Fault Detection and Classification (FDC) method for PV shading faults. At first, we extract the features for different experimental tests under healthy and shading conditions to build the database. The features are then analysed using Principal Component Analysis (PCA). The accuracy of the data classification into the PCA space is evaluated using the confusion matrix as a metric of class separability. The results using experimental data of a 250 Wp PV module are very promising with a successful classification rate higher than 97% with four different configurations. The method is also cost effective as it uses only electrical measurements that are already available. No additional sensors are required.

Keywords: PV shading faults; I-V curves; Principal Component Analysis; Fault detection; Fault classification

1. Introduction

In recent years, photovoltaic (PV) systems have received considerable attention thanks to the development of PV technologies and the growing demand for renewable energy in a wide range of applications (satellites, telecommunication, electric vehicles, homes, agriculture…).

Solar PV energy has become the third most important renewable energy after hydro and wind energy with a global installed capacity of 402 GWp by the end of 2017 (REN21, 2018).

The efficiency of PV systems is limited to 15–20% (Maghami et al., 2016). In addition, PV modules present an average performance degradation rate of 0.923% per year according to the study of Tabatabaei et al. (2017), which has been evaluated for mono-crystalline silicon (mc-Si) PV systems. More recently, for the same PV technology, Quansah et al. (2018) reported an annual degradation rate of maximum produced power of 1.54%.

PV systems are subject to various types of faults. These faults can be related to many factors such as material interactions (corrosion of connectors, yellowing, browning of encapsulation material and discoloration of busbars…) and environment factors such as soiling and shading. Soiling refers to the accumulation of snow, dirt, dust, leaves, pollen, and bird droppings on PV panels (Maghami et al., 2016). Shading may be a result of soiling or occurs due to obstructions caused by trees, buildings or chimneys. Thus, PV cells or modules may be partially or completely shaded during their operation. Shading is one of the most recurrent and damageable faults. In fact, this condition induces important degradation of PV

system performances. Partial shading can lead to more than 10-20% of annual reduction in power production in residential applications as shown by Deline et al. (2011). Moreover, the presence of localized shading on PV modules leads to an overheating of the shaded cells despite the presence of bypass diodes. Using the infrared thermography (IRT) many studies prove the presence of hot spot zones on the shaded PV cells (Basri et al., 2015; Tsanakas et al., 2016). Thus, the temperature increase in these zones leads to a thermal power dissipation (Bressan et al., 2016), reduces considerably the PV module lifetime and can damage the shaded cells (Brooks et al., 2015). The detection of such undesirable operating conditions has become mandatory for obvious safety and economic reasons.

a) Existing PV diagnosis approaches in literature

Several PV diagnosis and monitoring studies have been developed. However, the used techniques often require a relatively high cost in equipment or complexity in the diagnosis process development. In general, they are three main approaches used for PV fault diagnosis: image-based, model-based and process history-based also known as data-driven.

The common image-based PV diagnosis methods are the ElectroLuminescence (EL) and the IRT imaging under steady state conditions. These methods are becoming increasingly popular, since they offer efficient solution not only for detecting the fault occurrence within a PV plant, but also for isolating accurately the fault. Such optical inspection techniques need appropriate and expensive equipment (thermal camera, silicon charged coupled device (CCD) camera…). EL-based diagnosis method is rather efficient to indicating the existence and the location of contact failures; cell cracks and shunts, inactive PV cells or sub-strings (due to disconnection or shunted bypass diodes) and potential induced degradation (PID) with high accuracy (IEA, 2014). Nevertheless, this technique requires particular test conditions. A

camera with high resolution and a high pass filter are required. In addition, electroluminescent inspections must not be done under maximum power point (MPP) conditions; they are performed either in dark environment or after interrupting the PV system's operation. To perform the EL technique, the PV module must be supplied by a DC-current to stimulate radiative recombination in the PV cells (IEA, 2014). Thus, in the case of large PV installations, the experimental setup may become complex, costly and time-consuming. From this point of view, this technique appears more practical for small PV plants. IRT measurements are conducted outdoors and at MPP operation. The majority of faults detected by this method, which are similar to those detected by EL-imaging, have a significant effect on the defective PV module's thermal behaviour; their signature appears as marked and inhomogeneous points in the temperature distribution on the surface of the PV module. IRT-based method is fast, real time and effective to detect and exactly locate the faults thanks to the thermal signature, and without disturbing or interrupting the PV system operation. However, IRT method needs also specific conditions to be performed (sunny cloudless day, high irradiation, low ambient temperature and wind speed, accurate angle of view…) for correct and accurate temperature measurement (IEA, 2014).

Model-based approaches generally use an analytical model of the PV system to estimate the parameters, which will be compared to the measured ones obtained from real data. The generated residuals are used as fault features for diagnosis purposes. Recently, some model-based techniques rely on the PV power losses analysis. These modelling methods need knowledge of both irradiance and PV generator temperature to predict the output power of the PV system (Chouder and Silvestre, 2010; Kang et al., 2010). More recently, model-based techniques use the empirical parameters (fill factor (ff), short-circuit current (Isc), open-circuit

voltage (Voc)…) that are calculated from the shape of the current-voltage (I-V) curves (Garoudja et al., 2017; Ali et al., 2016; Spataru et al., 2015). The main advantage of these methods is that they have low hardware requirements and are applicable to a wide range of PV systems. If the designed model can capture the main physics of the system, these methods are efficient for shading detection.

Data-driven approach is based on data history, collected during operation. Fault features are extracted and analysed for fault diagnosis. Different techniques can be used ranging from signal processing to computational intelligence and machine learning. They do not require any explicit model of the process under monitoring. Among signal processing techniques, time-domain reflectometry (TDR) (Takashima et al., 2008) is used to detect and identify open-circuit faults and spread spectrum time-domain reflectometry (SSTDR) techniques are used to detect catastrophic faults, ground-faults and PV arc faults (Alam et al., 2013; Alam et al, 2014). These techniques are costly and require a specific external signal function generator. Moreover, they are not used to detect and identify shading faults. Other techniques extract the fault features from the I-V characteristic of PV module, string or array. Based on the analysis of the first and the second derivatives of I-V curves, Bressan et al. (2016) detect the activation of bypass diodes that indicate the presence of shading fault. This fault is also detected by comparing the I-V curves in normal and shaded operations as studied by El Basri et al. (2015). These methods are simple and effective to detect shading faults, but they are not able to identify and classify the type of shading patterns.

Artificial neural network (ANN) and fuzzy classifier are the most used methods for shading fault detection as described in (Dhimish et al., 2017; Spataru et al., 2015). However,

these methods suffer from several disadvantages like requiring a large amount of training data for accuracy detection, time-consuming training step and sensitivity to unbalanced weather conditions. In addition, this data is obtained for a specific PV installation. Thus, the rules are strongly tied to the system under study. Another disadvantage of these techniques is that the trained data need to be updated periodically. This is due to the high variability of operating conditions such as the environment variation or solar cells degradation and aging. This means that a trained data in low irradiance and low temperature condition for example, may misclassify the data and generate false alarms for healthy operating conditions if the irradiance and the temperature are higher.

However, taking advantage of the PV systems during operation, a huge quantity of data can be collected for analysis. Therefore, data-driven modelling is relevant, and features can be extracted then analysed for fault diagnosis purposes. In the field of features extraction techniques, Principal Component Analysis (PCA) is one of the most common multivariate statistical tools used for data representation and classification (Jollife, 2002).

PCA has been proved in several studies to be effective and powerful for the diagnosis in different applications and shows good classification performances as in many studies (Harkat et al., 2006; Harmouche et al., 2012; Gharavian et al., 2013; Harmouche et al., 2014; Adouni et al., 2015; Harmouche et al., 2015). This technique is very attractive for applications involving complex systems. To the best of our knowledge, it has not yet been used for PV systems diagnosis in such operating conditions.

b) Paper Contribution

We propose in this study to investigate the effectiveness of shading fault diagnosis using PCA for the analysis of the features extracted from real I-V curves. The proposed method is applied offline, for the case of a PV module. Based on the obtained results, the PCA's performances for fault detection and classification are discussed.

Here are the research contributions:

- For the first time, an implicit PCA model is developed for PV system fault detection and classification (FDC). This model has several advantages over the reported models in literature, such as simplicity and low training cost. Moreover, this model leads to good and clear data visualization.
- Compatibility with the existing PV systems. The FDC method can operate with any connected PV system, thanks to the integration of online I-V tracers for the new existing PV inverter technologies. It takes the advantage of available measurements in such existing systems with no additional hardware.
- In addition to its ability to discriminate the healthy data from the faulty ones, the proposed approach shows a good classification capability for the same category of fault
(shading). In fact, the different shaded configurations are well classified using the PCA algorithm.

c) Paper outline

160 The paper is organised as follows; in section 2, preliminary simulation results are presented
161 to verify the ability of using I-V curve and data processing with PCA to separate different
162 faulty conditions. In section 3, a brief description of the experimental setup is done and the
163 experimental tests in healthy and faulty operations are presented. In section 4, the fault
164 detection methodology is described and implemented. The data processing and the
165 evaluation's results are also detailed both in the training and the validation steps. Section 5
166 concludes the paper.

167 ## 2. Preliminary studies: analysis of the I-V curves for different faults
168

169 For PV systems, the degradation (or the faulty) modes are reflected differently in the I-
170 V curve, which has a particular shape under shaded condition due to the activation of bypass
171 diodes. In the following, we consider a PV module with the same specifications as the one
172 considered in this paper and a series resistance of 0.3 Ω at Standard Test Conditions (STC).
173 To show the effects of some of the PV faults on the I-V curve, we present in Fig.1 the
174 simulation results obtained under healthy conditions, partial shading and degradation of the
175 series resistance. Under shading fault, we consider that the three sub-strings receive non-
176 uniform irradiations. We can observe from Fig.1 that a degradation of the series resistance
177 mainly modifies the I-V curve in the region close to the open circuit voltage V_{oc} (Rodríguez et
178 al. 2015). The same effect can be observed in the case of potential-induced degradation (PID)
179 (Spataru et al. 2015).
180

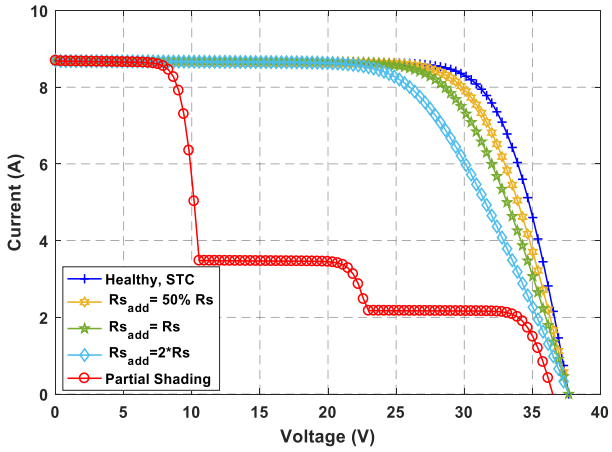


Fig. 1: PV faults' signatures on the I-V curve

From Fig.2 we can deduce that the projection of the data in the new reference frame spanned by (PC1, PC2) that the partial shading fault could be clearly separated from the degraded series resistance. However to separate the healthy case from the degraded series resistance, additional data processing or/and additional information should be done or included.

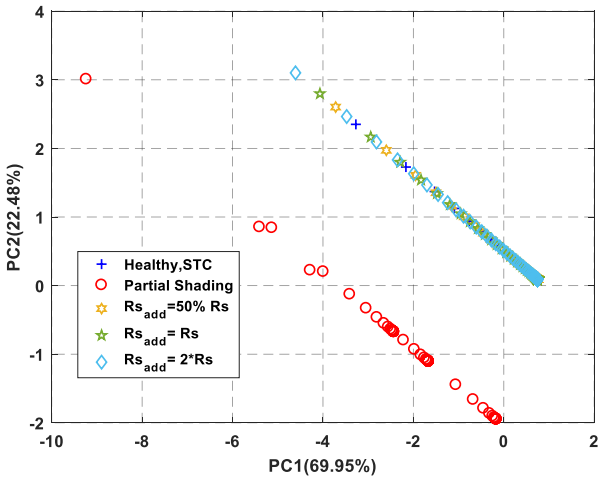


Fig. 2: PCA results in the subspace spanned by PC1 and PC2 for the simulated PV faults

In the following, we will focus on the shading fault detection with experimental data. Also more details will be provided on the features selected for PCA evaluation.

3. PV system and shading condition

The common models used to reproduce the I-V characteristic of a PV cell are based on the one diode or two diodes equivalent electrical circuit (Askarzadeh and Rezazadeh, 2012). Other models have been developed to offer a better modelling of the physical phenomena in a PV cell (Tossa et al., 2014; Bishop 1988). The classical single diode model is generally used since it is adequate at reproducing the main characteristics of a PV cell. Shading the total or the partial PV system surface is a very serious concern in such systems (Quaschning and Hanitsch, 1996; Patel and Agarwal, 2008). In order to mitigate the shading effects, PV systems are equipped with bypass diodes. These diodes become operational when the PV cells are reverse biased under shading condition. The activation of these diodes creates a short circuit of the shaded cells, which limits their reverse voltage and thus the dissipated power. In practice, a single bypass diode is usually connected across a group of 18-20 cells.

3.1. Experimental set-up description

The evaluation of the proposed fault detection technique is carried out using real data generated from the FL60-250MBP PV module. The main parameters are given in Table 1 under Standard Test Condition (STC) (1000 W/m^2, 25 °C). It is composed of 60 mc-Si based PV cells, connected in series and gathered into three sub-strings of 20 PV cells for each one. This module is equipped with three bypass diodes; each one is mounted in anti-parallel to protect a PV sub-string.

In this experiment, the I-V curves are obtained online using a variable load (Programmable DC electronic load Chroma 63600), which provides 101 data from open-circuit voltage to

short-circuit current for each I-V curve. Online measuring methods of the I-V characteristic can be also done using other real devices. In fact, their main principle is to apply a variable impedance, which changes from a very large (or small) value to a small (or large) one in order to extract voltages and currents values between open circuit voltage and short circuit current. Many examples of these methods can be found in the literature. Varying the impedance can be created by using resistive load (Van Dyk et al, 2005), charging or discharging a capacitor (Benzagmont et al., 2018; Mahmoud 2006; Muñoz and Lorenzo, 2006; Spertino et al., 2015) or using an electronic switch like MOS transistor (Kuai and Yuvarajin, 2006). The use of one device depends on the PV power and the desired accuracy of measurements. For example, the use of the capacitive load is recommended for measuring in PV installations from 2 up to 50 kWp (IEC, 1995). Considering 101 samples in this experiment is enough to sweep the I-V curve of the module under study. A reference cell (RG100 by SOLEMS) is used to measure the solar irradiance captured by the PV module area and a 4-wire Pt100 probe, bonded on the back face of the PV module, is used to measure the temperature.

A data acquisition system is installed and a computer is used for supervision and data visualization using LABVIEW®. This experimental setup is installed at the French national observatory SIRTA (Haeffelin et al., 2005). A picture and a diagram describing the instrumentation are displayed in Fig.3.

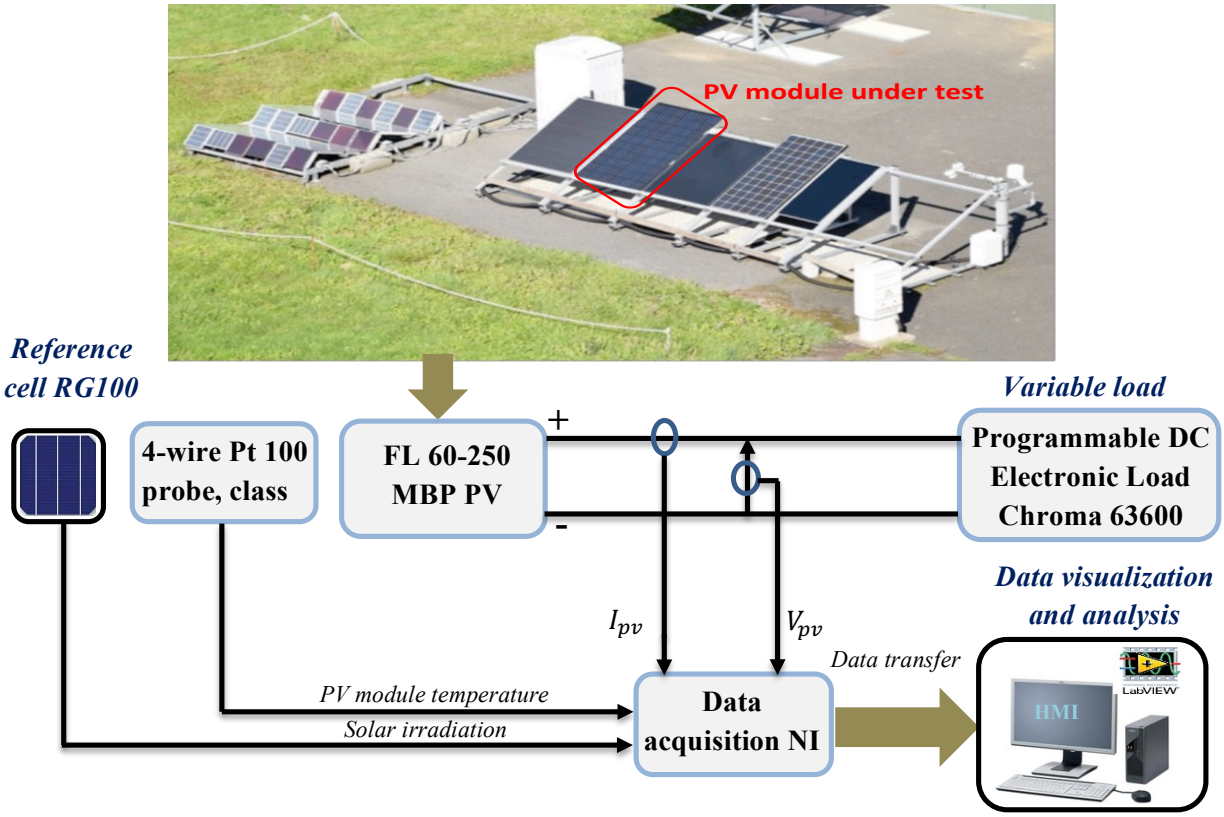


Fig. 3: Layout of the experimental setup

Table 1
PV module specifications at STC

Symbol	Quantity	Value
P_{mpp}	Maximum Power (Wp)	250
I_{mpp}	Current at P_{mpp} (A)	8.21
V_{mpp}	Voltage at P_{mpp} (V)	30.52
I_{sc}	Short-circuit Current (A)	8.64
V_{oc}	Open-circuit voltage (V)	37.67
S	Area of the module (m^2)	1.64

3.2. Data Acquisition

Five sets of experimental tests have been conducted to assess the fault detection approach. They have been realised under several operating conditions:

- One healthy mode
- Four faulty modes with different shading conditions: for each set, the shading is applied by covering the PV cells with a survival blanket.

The experimental data is redundant. We have recorded three measurements (A_m, B_m and C_m) of a complete I-V characteristic for each set of healthy and faulty tests. The collected data sets A_m and B_m will be used for the training step while C_m will be used for validation as explained in section 4. Each I-V curve is composed of 101 samples. It is recorded in one minute. 101 samples are enough for our system to sweep a complete I-V characteristic. Generally, the number of samples needed to extract these characteristics is selected according to the size of the PV system and to the mismatching conditions in order to clearly show the deviations and the inflection points on the I-V curve. Despite the short duration between measurements, the irradiation can change significantly. For each I-V curve (Fig.4 and Fig.5), the three cases A_m, B_m and C_m are drawn in blue, red and mustard lines respectively.

3.2.1. Healthy condition

Fig.4 illustrates the experimental I-V curve measured three times when the PV module operates in healthy condition and clear condition (more than 800 W/m^2). As the PV module short-circuit current is proportional to the solar irradiation, it produces less current when receiving low irradiation level. We notice that these results are consistent with the datasheet information given under STC (Table 1).

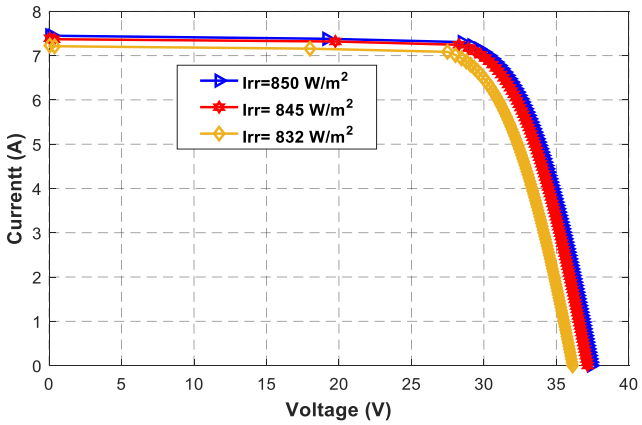
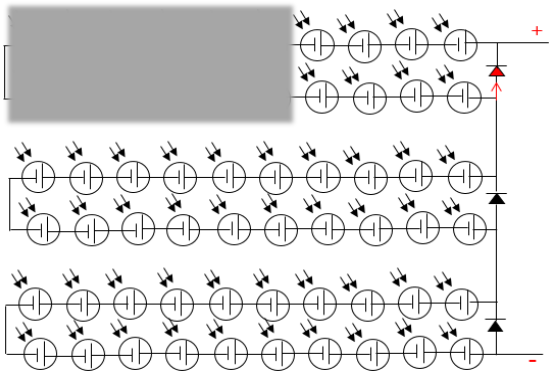


Fig. 4: Experimental I-V curve under healthy (normal) condition

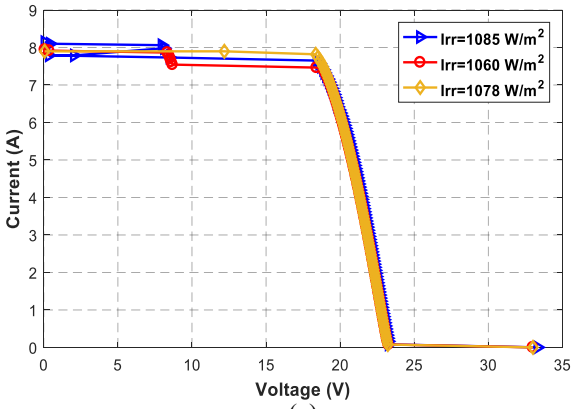
3.2.2. Shading conditions

The four configurations of shading, the active diodes and the I-V characteristics (measured three times for each configuration) are illustrated in Fig.5. In shaded conditions, all the I-V curves show multiple peaks explained by the state of the bypass diodes relative to each type of the applied shading. These peaks prove the efficiency degradation of the PV system under shading since its maximum produced power is reduced. As it is partially shaded, the PV cells of the module under test are under non-uniform irradiation. We note that for each I-V characteristic, the solar irradiation displayed in the legend is the one measured with the reference cell (RG100). The behaviour of the experimental curves differs according to the shading configuration (row level, column level and number of shaded cells) and to the environmental variations (temperature and solar irradiation). Fig.5a shows the first shading configuration; one PV sub-string is partially shaded. The shading of 12 cells leads to the activation of one bypass diode so the deactivation of the faulty sub-string. This is confirmed with the voltage steep variation due to the lost of this sub-string. (Fig.5e). For the rest of the

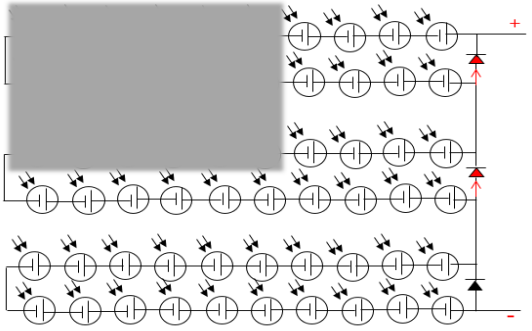
configurations, two bypass diodes are activated, as two sub-strings are partially shaded for each configuration. According to the severity of the shading fault, we have two cases: the diodes are conducting simultaneously (Fig.5c) and there is one voltage peak (Fig.5g), or they are activated one after the other (Fig.5b and Fig.5d) and we observe two peaks in the voltage.



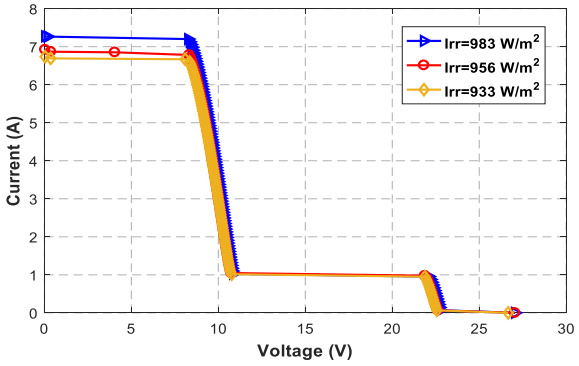
(a)



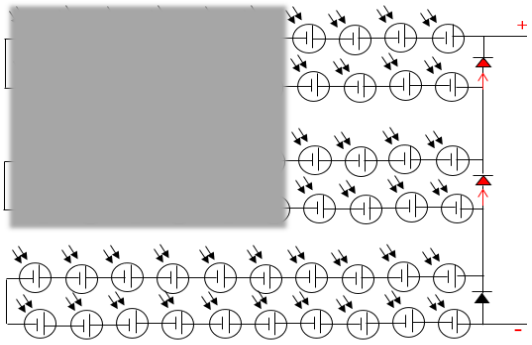
(e)

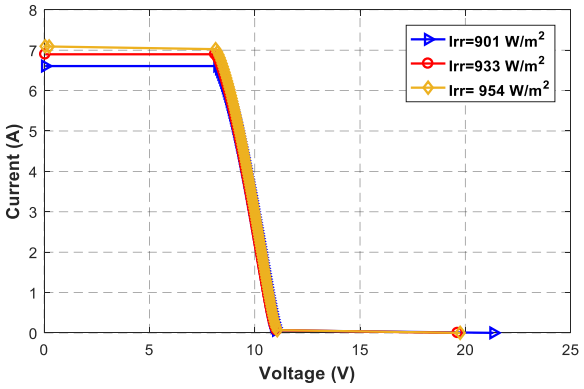


(b)



(f)





(g)

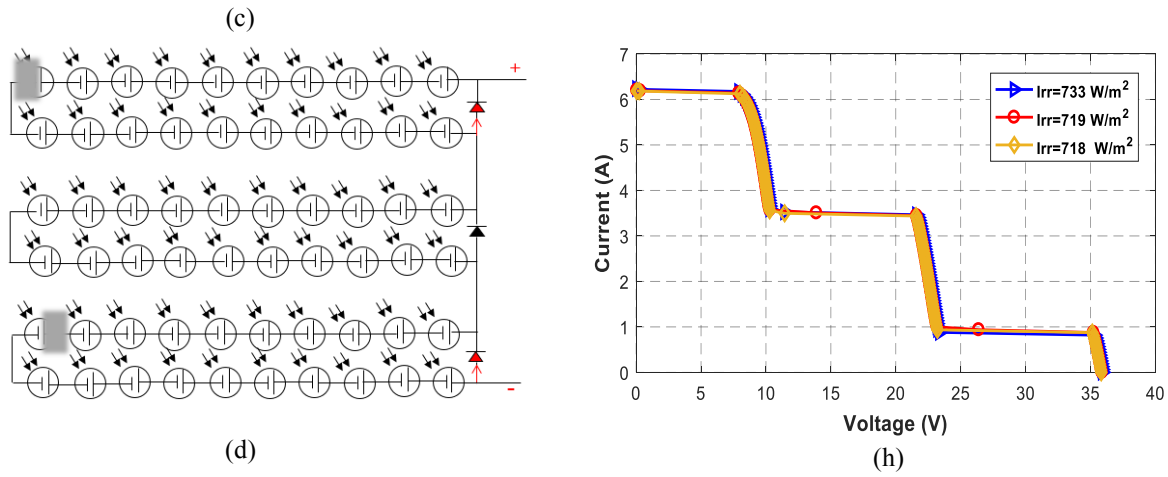


Fig. 5: Shading configurations (a) 1, (b) 2, (c) 3, (d) 4, and experimental I-V curves (e), (f), (g), (h)

3.3. Data reproducibility

For our evaluation, the I-V curve is acquired three times for each experimental test. The main idea behind the data redundancy is to create a database for the diagnosis algorithm. This database will be divided into training data and test data. The experimental tests show that the shading faults strongly modify the shape of the I-V characteristics. Therefore, we investigate in this paper the use of area under the curve (AUC) as a metric to analyse the differences between the redundant data of each experimental test. It is computed based on the PV current and PV voltage. We consider the normalized current with respect to the measured irradiation. An approximation of the AUC is expressed as following:

$$AUC = \sum_{i} \frac{1}{2}[(v_j - v_{j-1})(i_{norm,j} - i_{norm,j-1})] + i_{norm,j-1}(v_j - v_{j-1}) \quad (1)$$

$$i_{norm} = i\frac{G_{STC}}{G_{mes}} \quad (2)$$

where G_{STC} and G_{mes} are respectively the irradiations for STC and the measured one.

The standard and relative errors of the AUC computed for each test are given in Table 2. We can first notice that the highest AUC values are obtained under healthy condition. This is consistent with the I-V curves. This area mainly varies with reference to the test condition (healthy or faulty) and to the shading fault type.

The replication of the measurements shows very little deviations between the I-V data obtained for each test, which is consistent with the one-minute measurement process. In fact, the solar irradiation changes little for a maximum rate of 13 W/m^2 in healthy case and 25 W/m^2 for the first faulty configuration, 50 W/m^2 for the second, 53 W/m^2 for the third and 15 W/m^2 for the last one. The maximum relative error accounted for the healthy operation is 4.69% and 2.36% for the faulty one. According to this variability, we assume that the database is representative of all operating conditions and can be used for evaluating the fault detection.

Table 2
Evaluation of area under the curve, the standard and the relative errors for the experimental tests

		Healthy condition	**Shading configurations**			
			1	**2**	**3**	**4**
Area under the curve	A_m	299.62	153.34	83.06	71.64	157.15
	B_m	296.49	154.26	82.34	72.08	160.87
	C_m	285.56	155.04	82.16	72.10	159.84
Standard error	$\Delta A_m B_m$	3.13	0.92	0.72	0.44	3.72
	$\Delta A_m C_m$	14.06	1.70	0.90	0.46	1.03
	$\Delta B_m C_m$	10.93	0.78	0.18	0.02	2.69

		Healthy condition	Shading configurations			
			1	2	3	4
Relative error (%)	$\Delta A_m B_m/A$	1.04	0.60	0.86	0.61	2.36
	$\Delta A_m C_m/A$	4.69	1.10	1.08	0.64	0.65
	$\Delta B_m C_m/B$	3.82	0.50	0.21	0.02	1.67

4. Shading fault detection

The flowchart of the method is displayed in Fig.6. The first step consists in the modelling. The training data represents 67% of the experimental measurements determined for the healthy and all cases of PV shading faults. The second step consists in the data pre-processing. The PV power and efficiency are used as additional information to build the matrix X. Then, the logarithm function is applied to each variable of this database. Finally, the features are extracted and analysed.

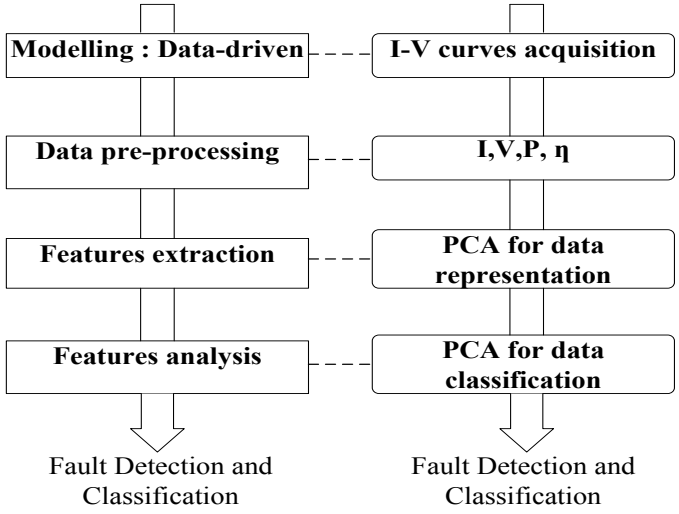


Fig. 6: Flowchart of the FDC algorithm

When dealing with I-V curves in real environmental conditions, it might be difficult to distinguish healthy operations from faulty ones. Therefore, to ensure reliable fault detection one should find an appropriate workspace in which the data's separability is highlighted. Indeed, PCA is known to be an efficient multivariate statistical tool for this purpose.

The PCA is applied to the matrix X to get the principal scores that form the PCA space for data representation and classification.

4.1. Principal Component Analysis formulation

Principal Component Analysis (PCA) is a multivariate statistical technique that seeks in the multidimensional space of system variables the most dominating dimensions to re-express the multivariate database built from a large number of measurements recorded at different times. The new dimensions are uncorrelated so the reduced subspace acts like a denoising filter and keeps the underlying "latent structures" in the data. This new subspace is denoted the principal subspace or the 'representation' subspace. Its complementary subspace into which noises and outliers are rejected is termed the residual subspace. Analytically, PCA searches orthogonal directions, which contain the maximum variance of the projections for the data set points. The PCA task is formulated by a problem of the eigenvector decomposition of the data covariance matrix (Jollife, 2002).

The data consists of measurements collected at N different sampling times of M variables. The time points represent the observations. PCA uses a linear combination of the original variables to build the new variables while keeping maximum variance information. The first

352 principal components, which span the principal subspace, are given by the first l dominant
353 eigenvectors of the data covariance matrix. They are associated to the l highest eigenvalues.
354 The last non-retained eigenvectors ($M - l$) define the residual subspace. In the representation
355 subspace containing the most significant variations, the eigenvectors are denoted loading
356 vectors and the projection of the data on these loading vectors are called principal component
357 scores. These searched directions are called principal components (PCs), each one being
358 characterised by a loading vector and a score component. The percentage of the variance of
359 data contained in each PC is expressed by its corresponding eigenvalue. Each PC is aligned in
360 a direction corresponding to the largest variance of the data, starting with the first PC. Principal
361 components are therefore ordered from the most energised associated to the highest
362 eigenvalue, to the less energised associated to the lowest eigenvalue. Based on stop criteria the
363 principal subspace is spanned with most energised PCs while the residual one is spanned with
364 the remaining PCs. For this purpose, many stopping criteria have been proposed in the
365 literature such as the cumulative percentage of total variance (CPV) (Chiang et al., 2001) and
366 minimizing the Variance of Reconstitution Error (VRE) (Qin et al., 2000). A comparison
367 between 11 methods to determine the number of most energised PCs has been given by Valle
368 et al. (1999). From the data collected in the healthy operating mode of the process, PCA is
369 applied and the loading vectors are used as references to design the model. The scores or their
370 statistical distributions can be also used to design the implicit model. When new data is
371 collected, it is projected in the subspace (principal or/and residual). The deviation from the
372 reference is then measured and analysed to assess the fault occurrence.
373 Let us consider N observations of M process variables gathered into the original data matrix
374 $X_{[N\times M]}$ given by:

$$X_{[N\times M]} = [x_1,...,x_k,...,x_M] \quad (3)$$

Where $\boldsymbol{x}_{j(j=1,...N)}$ is the j^{th} variable.

At first, it consists in centring (zero mean) and reducing (unit variance) the variables for each observation k of $\boldsymbol{x}_{j(j=1,...N)}$:

$$(x_i)_c(k) = \frac{x_i(k) - (\overline{x_i})}{(\sigma_i)} \quad (4)$$

where $(\boldsymbol{x}_j)_c$ is the centred and reduced variable, $(\overline{x_j})$ and (σ_j) are respectively the mean value and the standard deviation of $\boldsymbol{x}_{j(j=1,...M)}$.

We can therefore define the new data matrix as:

$$(X_c)_{[N\times M]} = [(x_1)_c,...,(x_k)_c,...,(x_M)_c] \quad (5)$$

The covariance matrix is then calculated as:

$$C = \frac{1}{N-1} X_c{}^T X_c \quad (6)$$

The quality of the representation for the collected measurements for fault diagnosis purposes relies on the accuracy of the PCA model. This model depends on the retained PCs to represent the data variability. Let us denote ***P*** the column matrix of loading vectors, which are arranged in the descendent order of their corresponding eigenvalues. The principal component scores are obtained by the projection of the original data centred and reduced on the new space spanned with ***P***. The matrix $T_{[N\times M]}$ of the principal component scores is defined by:

$$T_{[N\times M]} = (X_c)_{[N\times M]} P_{[M\times M]} = [t_1,...,t_k,...,t_M] \quad (7)$$

396

397 *4.2. Data pre-processing and features extraction*

398 The selection of the variables is very important to obtain the best representation and
399 discrimination of the data. In order to detect the shading fault, Fadhel et al. (2018) have used
400 the voltage, the current and the power of the PV module as variables. Thanks to PCA, they
401 have successfully distinguished the healthy data from the faulty one. However, using these
402 variables in our case has led to a severe overlapping in the space spanned with the principal
403 scores. This is due to the variation of the irradiance between two measurements for the same
404 operating condition and also to the common levels of voltage between the PV curves obtained
405 under shading faults (Fig.5). In order to have a fault diagnosis, robust to environmental
406 changes and sensitive to fault occurrence, we have:

407 ✓ normalised the PV voltage and power with respect to the PV efficiency,

408 ✓ used the log function.

409 The selection of these features allows us to use experimental measurements obtained in
410 non-controlled irradiance operating conditions without a huge influence of this environmental
411 parameter.

412 The principal component analysis is finally applied to the training data matrix
413 $X_{[1010\times3]} = \log[v/\eta \quad i \quad P/\eta]$ composed of 1010 observations of 3 variables (Fig. 7) where P
414 and η are respectively the power and the efficiency. It consists of 5 sub-matrices of 101*2
415 training observations. The efficiency is computed for each couple of observation (i,v) and is
416 expressed as follows:

$$\eta = \frac{P}{G_{mes} S} 100 \quad (8)$$

where S is the area of the PV module under test.

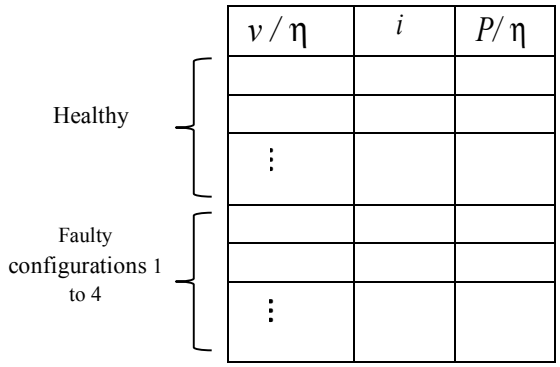


Fig. 7: Data set design for PCA analysis

The data set generated for both healthy and shading conditions during the experimental tests is composed of 1515 samples. The training step is performed with 1010 samples using the first two measurements A_m and B_m of each test. The main task of this step consists in the construction of the PCA model from the learning data set. This implicit model will be used for validation of the test data set.

4.3. Features analysis in the training step

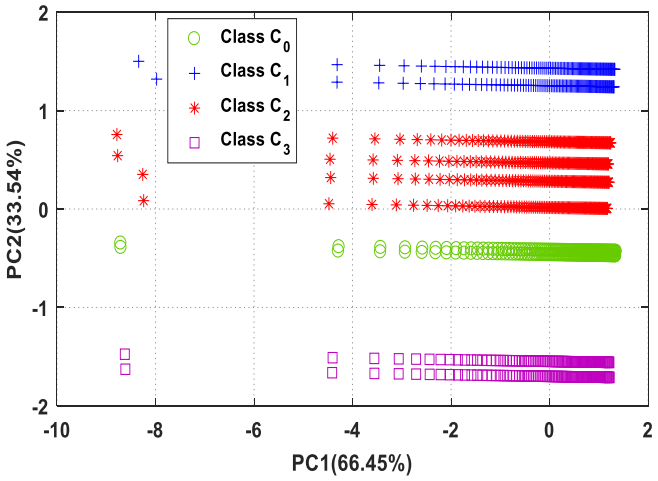
Table 3 presents the eigenvalues and their relative contributions. The first two PCs retain 99.99% of the information. Projecting the training data into the PCA subspace spanned with PC1 and PC2, PC2 and PC3, PC1 and PC3 gives the data scatter displayed in Fig.8a, Fig.8b and Fig.8c respectively. The inclusion of the three principal components in the PCA space used to project the data gives the 3D representation of Fig.8d. The representation in the subspace spanned with the first two PCs is able to detect and identify the fault. We can

observe four classes: one healthy named class C_0 and 3 faulty obtained from 4 faulty configurations as given in Table 4. Indeed, shading configurations 2 and 3 correspond to the simultaneous activation of two bypass diodes. The healthy class is well separated from the faulty classes. Those ones are distinguished with reference to the fault size and location.

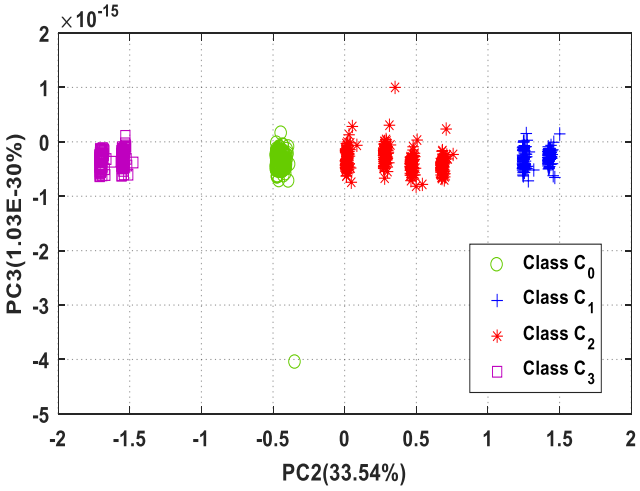
Table 3

Eigenvalues and Percentage of the Principal Component contributions

	Principal components		
	PC1	PC2	PC3
Eigenvalue	1.99	1.006	3.09 E-32
Variance (%)	66.45	33.54	1.03 E-30



(a)



(b)

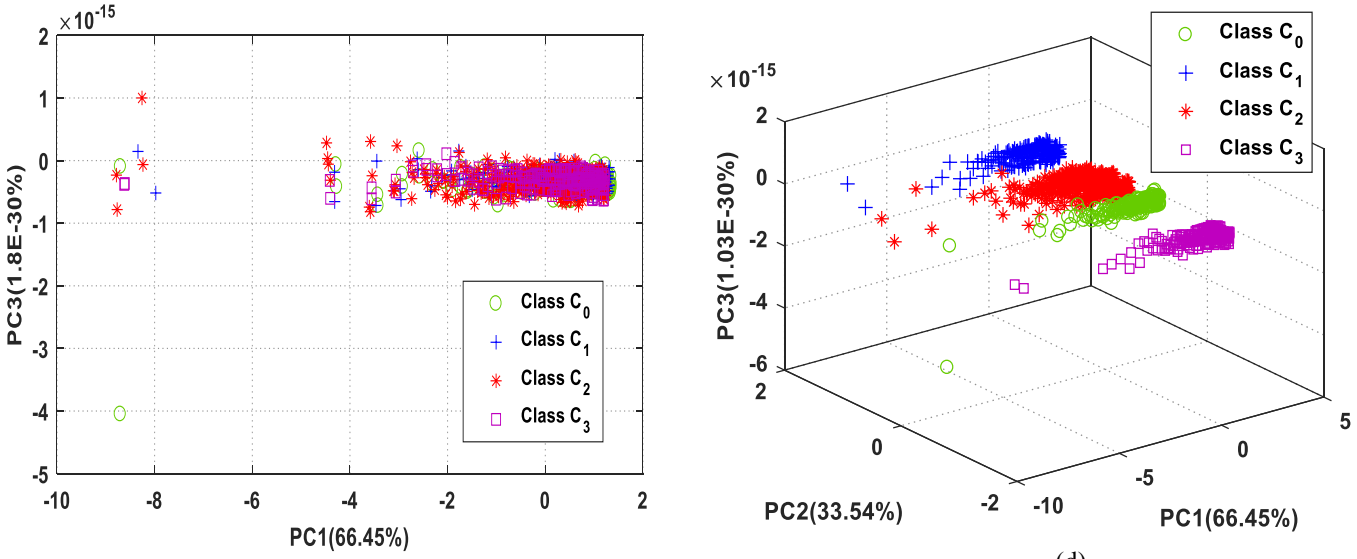


Fig. 8: PCA training data set results in the subspace spanned by PC1 and PC2 (a), PC2 and PC3 (b), PC1 and PC3 (c), and 3D PCA plot

Table 4

Classes in the PCA space

Test Condition	Class
Healthy	C_0
Shading configuration 1	C_1
Shading configurations 2 and 3	C_2
Shading configuration 4	C_3

4.4. Classification performance in the training step

The results are analysed through the confusion matrix displayed in Table 5. The columns of this table show the percentage of affectation of the observations of a class *a priori* in a class *a posteriori*. The error rates of class separability are checked by the one-leave-out cross validation method. We first compute the coordinates of the gravity centre (considered as the mean value in our case) of each class a priori in the PCA space. Then, the Euclidean distance

between these centres and each observation in the training database are evaluated. Finally, an observation is assigned to a class among the four obtained classes if it is the closest to its centre of gravity. This table shows that a few errors are found by classifying the measurements of the training data set. We found that 97.03% of the measurements in healthy condition have been classified in their a priori class C_0 and only 2.97% misclassification is found and affected to the faulty class C_2 (corresponding to three misclassified healthy measurements). We have also found that 100% of the faulty data of shading configuration 4, represented by faulty class C_3, are perfectly classified. Based on these good discrimination results, we can use the implicit model obtained from the training data and PCA for the analysis of the new observations.

Table 5
Confusion Matrix for training data set classification

Class a priori	Data Assignement a posteriori			
	Healthy Class C_0 (%)	Faulty Class C_1 (%)	Faulty Class C_2 (%)	Faulty Class C_3 (%)
Healthy Class C_0	**97.03**	0	2.97	0
Faulty Class C_1	0	**98.52**	1.48	0
Faulty Class C_2	12.62	0	**87.38**	0
Faulty Class C_3	0	0	0	**100**

4.5. Fault identification performance in the validation step

In order to evaluate the effectiveness of the PCA model for PV system Fault Detection and Classification, we have used a new data set composed of 505 observations, representing 33% of the experimental database. The test data set corresponds to the measurements C_m that were not included in the training data. These measurements are grouped in the test database

according to the three selected representative variables, v/η, i and P/η. Then they are projected into the PCA space spanned by the eigenvectors determined during the training step using (7). This projection gives the data classification shown in Fig.9. We can conclude that all the test observations are well identified and classified in the relevant group. The data dispersion obtained for the new data in the PCA space is similar to the one obtained with the training data. The faulty classes are well separated and discriminated from the healthy one.

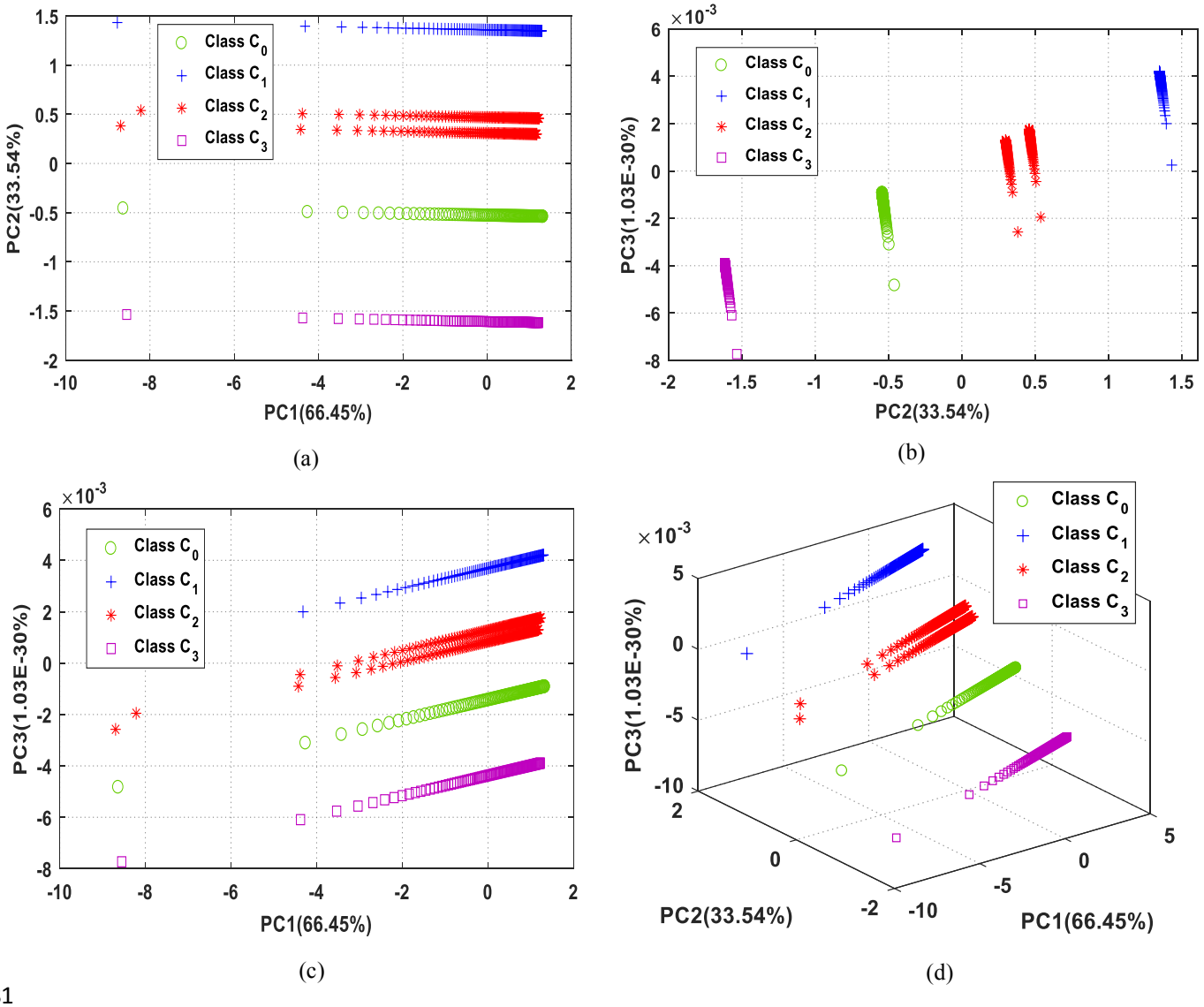


Fig. 9: PCA results for the test data set in the subspace spanned by PC1 and PC2 (a), PC2 and PC3 (b), PC1 and PC3 (c) and 3D PCA plot

With the PCA, we are able to identify successfully the test data in their corresponding groups. This performance is evaluated through the confusion matrix given in Table 6. We have succeeded to separate the four *a priori* defined classes with a minimum rate of 97.03%. Only 4/505 test samples are misclassified in two *a posteriori* groups; 3/101 for healthy class C_0 and 1/101 for faulty class C_1 representing a classification error of 2.97% and 0.99% respectively. This confirms that the developed PCA model is very efficient to detect and identify the fault.

Table 6
Confusion Matrix for test data set classification

Class a priori	Data Assignement a posteriori			
	Healthy Class C_0 (%)	Faulty Class C_1 (%)	Faulty Class C_2 (%)	Faulty Class C_3 (%)
Healthy Class C_0	**97.03**	0	2.97	0
Faulty Class C_1	0	**99.01**	0.99	0
Faulty Class C_2	12.62	0	**100**	0
Faulty Class C_3	0	0	0	**100**

5. Conclusion

In this study, a data-driven FDC approach is proposed for a PV module shading fault diagnosis. This method uses the I-V curve of the PV module generated under healthy and faulty conditions. An experimental setup has allowed the collection of data for five different operating conditions to build the database. In the pre-processing step the power P has been added to the original variables v/η and i. Then a normalisation with the efficiency has been

done to mitigate the variation of the irradiance and the logarithmic function has been introduced to make the method more sensitive to fault occurrence. Principal Component Analysis has been applied to the training database (66% of the data history). The obtained model has been used for fault detection and classification. In the training step, we have rather good performances with a minimum classification success rate of 87.38% for the 4 classes (one healthy and three faulty). During the validation step (with the remaining 33% of data history), we have obtained successful classification rate with a minimum of 97%.

This method does not depend on any particular PV size and only uses available measurements (PV current and voltage) avoiding extra hardware and costs. Furthermore, it is insensitive to the weather conditions changes (sudden variations of solar irradiation and temperature of the PV module). Based on the analysis of real PV data, the study demonstrates the feasibility and effectiveness of the PCA for the diagnosis of PV shading faults.

Acknowledgement

The authors gratefully recognize the financial support of the Ministry of Higher Education and Scientific Research in Tunisia and University of Paris Sud in France who provided the scholarship to the PhD student.

References

Adouni, A., Delpha, C., Diallo, D, Sbita, L., 2015. Voltage Dip fault detection and ddentification based on Principal Component Analysis: Application to Wind Energy Conversion System. In: IEEE 24th International Sympsium on Industrial Electronics (ISIE), pp.867-872.

Alam, M.K., Khan, F., Kerkes, D., 2014. PV Arc-fault Detection using Spread Spectrum Time Domain Reflectometry (SSTDR). In: IEEE Energy Conversion Congress and Exposition (ECCE), pp. 3294 – 3300.

Alam, M. K., Khan, F., Johnson, J., Flicker, J., 2013. PV ground-fault detection using spread spectrum time domain reflectometry (SSTDR). In: Proc. IEEE Energy Convers. Cong. Expo, pp. 1015–1020.

Ali, M., H.,Rabhi, A.H., El hajjaji, A., Tina, G.M., 2016. Real Time Fault Detection in Photovoltaic Systems. In:8th International Conference on Sustainability in Energy and Buildings, pp. 914-923.

Askarzadeh, A., Rezazadeh, A., 2012. Parameter identification for solar cell models using harmony search-based algorithms. Solar Energy 86(11), 3241-3249.

Benzagmont, A., Martire, T., Beaufils, G., Fruchier, O., Talbert, T., Gachon, D., 2018. Measurement of the I(V) characteristics of photovoltaic array by the capacitve load method for fault detection. In: IEEE International Conference on Industrial Technology (ICIT), pp.1031-1036.

Bishop, J.W., 1988. Computer Simualtion of the Effects of Electrical Mismatches in Photovoltaic Cell Interconnection Circuit, Commission of the European Communities. Joint Research Center, ESTI Project. Solar Cells 25(1), 73-89.

Bressan, M., Basri, Y.El., Galeano, A.G., Alonso, C., 2016. A shadow fault detection method based on the standard error analysis of I-V curves", Renewable Energy 99, 1181-1190.

Brooks, A.E., Cormodel, D., Cronin, A.D., Kam-Lum, E., 2015. PV System Power Loss and Module Damage due to Partial Shade and Bypass Diode Failure Depend on Cell Behavior in Reverse Bias. In: IEEE 2nd Photovoltaic Specialist Conference (PVSC), pp. 1-6.

Chiang, MS., Leo, H., Russell, E. L., Braatz, R. D.,2001. Fault Detection and Diagnosis in Industrial Systems. Springer.

Chouder, A., Silvestre, S., 2010. Automatic supervision and fault detection of PV systems based on power losses analysis. Energy Conversion and Management 51(10), 1929-1937.

Deline, C., Marion, B., Granata, J., Gonzalez, S., 2011. A Performance and Economic Analysis of Distributed Power Electronics in Photovoltaic Systems. (NREL), prepared under task No. PVD9. 1410.

Dhimish, M., Homles, V., Mehradadi, B., Dales, M., 2017. Diagnostic method for photovoltaic systems based on six layer detection algorithm. Electric Power Systems Research 151, 26-39.

El Basri, Y.El., Bressan, M., Seguier, L., Alawadhi, H., Alsonso, C., 2015. A proposed graphical electrical signatures supervision method to study PV module failures. Solar Energy 116, 247-256.

Fadhel, S., Migan, A., Delpha, C., Diallo, D., Bahri, I., Trablesi, M., Mimouni, M.F., 2018. Data-Driven Approach for Isolated PV Shading Fault Diagnosis Based on Experimental I-V Curves Analysis. In: IEEE International Conference on Industrial Technology (ICIT), pp.927-932.

Garoudja, E., Harrou, F., Sun, Y., Kara, K., Chouder, A., Silvestre, S., 2017. Statistical fault detection in photovoltaic systems. Solar Energy 150, 485-499.

Gharavian, M.H., AlmasGanj, F., Ohadi, A.R., Bafroui, H.H., 2013. Comparison of FDA-based and PCA-based features in fault diagnosis of automobile gearboxes. Neurocomputing 121, 150-159.

Haeffelin, M., Barthès, L., Bock, O., Boitel, C., Bony, S., Bouniol, D., Chepfer, H., Chiriaco, M., Cuesta, J., Delanoë, J., Drobinski,P., Dufresne, J.-L., Flamant, C., Grall, M., Hodzic, A., Hourdin, F. Lapouge, F., Lemaître, Y., Mathieu, A., Morille, Y., Naud, C., Noël, V., O'Hirok, W., Pelon, J., Pietras, C., Protat, A., Romand, B., Scialom, G., Vautard, R., 2005. SIRTA, a ground-based atmospheric observatory for cloud and aerosol research. Annales Geophysicae 23(2), 253–275.

Harmouche, J., Delpha, C., Diallo, D., 2012. Fault diagnosis and detection using Principal Component Analysis and Kullback-Leibler Divergence. In: 38th Annual Conference on IEEE Industrial Electronics Society (IECON), pp. 3907 -3912.

Harmouche, J., Delpha., C., Diallo, D., 2014. Incipient Fault Detection and Diagnosis Based on Kullback-Leibler Divergence Using Principal Component Analysis: Part I. Signal Processing 94. 278-287.

Harmouche, J., Delpha., C., Diallo, D., 2015. Incipient Fault Detection and Diagnosis Based on Kullback-Leibler Divergence Using Principal Component Analysis: Part II. Signal Processing 109. 334-344.

Harkat, M.F., Mourot, G., Ragot, J., 2006. An improved PCA scheme for sensor FDI: application to an air quality monitoring network. Process Control 16 (6), 625-634.

IEC, 1995. International Standard 61829. Crystalline silicon photovoltaic (PV) array – On-site measurement of I–V characteristics.

Jollife, I.T,2002. Principal Component Analysis, Second ed, Aberdeen U. K, Springer.

Kang, B.K., Kim, S.T. Bae, S.H., Park, J.W., 2010. Diagnosis of output power lowering in a PV array by using the Kalman-filter algorithm. IEEE Trans. Energy Convers. 27(4). 885-894.

Kontges, M., Kurts, S., Packard, C., Jahn, U., Berger, K.A., Kato, K., Frisen, T., Liu, H., Iseghem, M.V., 2014. Performance and reliablity of photovoltaic modules, Substack 3.2: Review of Failures of Photovoltaic Modules. In : External Final Report by Internationael Energy Agency (IEA) for Photovoltaic Power Systems Programme (PVPS).

584 Kuai, Y, Yuvarajan, S., 2006. An electronic load for testing photovoltaic panels. Power Sources 154, 308–313.
585 Mahmoud, M.M., 2006. Transient analysis of a PV power generator charging a capacitor for measurement of the
586 I–V characteristics. Renewable Energy 31, 2198–2206.
587 Maghami, M., R. Hizam, H., Gomes, C.,M.A. Radzi, M. I. Rezadad, S .Hajighorbani, 2016. Power loss due to
588 soiling on solar panel: A review. Renewable and Sustainable Energy Reviews 59, 1307-1316.
589 Muñoz, J., Lorenzo, E., 2006. Capacitive load based on IGBTs for on-site characterization of PV arrays. Solar
590 Energy. 80, 1489–1497.
591 Patel, H., Agarwal, V., 2008. MATLAB-Based modeling to study the effects of partial shading on pv array
592 characteristics. IEEE Transactions on Energy Conversion 23 (1), 302-310.
593 Qin, S.J., Dunia, R., 2000. Determining the number of principal components for best reconstruction. Process
594 Control 10, 245-250.
595 Quansah, D.A., Adaramola, M.S., 2018. Ageing and degradation in solar photovoltaic modules installed in
596 northern Ghana. Solar Energy 173, 834-847.
597 Quaschning, V., Hanitsch, R., 1996. Numerical simulation of current-voltage characteristics of photovoltaic
598 systems with shaded solar cells. Solar Energy 56(6), 516-520.
599 REN21, 2018. Renewables 2018 Global Status Report.
600 Rodríguez, J.D.B-, Franco, E., Petrone, G., Paja, A.R-, Spagnuolo, G., 2015. Model Based Degradation Analysis
601 Of Photovoltaic Modules Through Series Resistance Degradation Estimation. IEEE Transactions on
602 Industrial Electronics 62(11), 7256-7265.
603 Spataru, S. Sera, D. Kerkes, T. Teodorescu, R., 2015. Diagnostic method for photovoltaic systems based on light
604 I-V measurements. Solar Energy 119, 29-44.
605 Spertino, F. Ahmad, J., Ciocia, A., Leo, P.D., Murtaza, A.F, Chiaberge, M., 2015. Capacitor charging method for
606 I-V curve tracer and MPPT in photovoltaic systems. Solar Energy 119, 461-473.
607 Tabatabaei, S.A., Formolo, D., Treur, J., 2017. Analysis of performance degradation of domestic
608 monocrystalline photovoltaic systems for a real-world case. In: International Scientific Conference:
609 Environmental and Climate Technologies (CONECT), pp.121-129.
610 Takashima, T., Yamaguchi, J., Ishida, M., 2008. Fault Detection by signal response in PV module Strings. In:
611 33rd IEEE Photovoltaic Specialists Conference, pp. 1-5.
612 Tsanakas, J.A., Ha, L., Buerhop, C., 2016. Faults and infrared thermographic diagnosis in operating C-Si
613 photovoltaic modules: A review of research and future challenges. Renewable and Sustainable Energy
614 Reviews 62, 695-709.
615 Tossa, A.K., Soro, Y.M., Azoumah, Y., Yamegueu, D., 2014. A new approach to estimate the performance and
616 energy productivity of photovolatic modules in real opreating conditions. Solar Energy 110, 543-560.
617 Valle, S., Li, W., Qin, J., 1999. Comparaison of multivariate statistical procee control monitoring methods with
618 applications to the eastman challenge problem. Industrial and Chemistry Research 38, 4389-4401.
619 Van Dyk, E., Gxasheka, A., Meyer, E., 2005. Monitoring current-voltage characteristics and energy output of
620 silicon photovoltaic modules. Renew. Energy 30, 399-411.