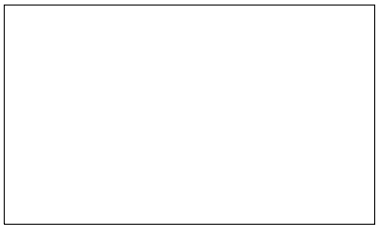
Feature extraction and health status prediction in PV systems

Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Corinne Alonso, Marko Pavlov

Graphical Abstract

Feature extraction and health status prediction in PV systems

Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Corinnne Alonso, Marko Pavlov



Highlights

Feature extraction and health status prediction in PV systems

Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Corinnne Alonso, Marko Pavlov

- Research highlight 1
- Research highlight 2

Feature extraction and health status prediction in PV systems

Edgar Hernando Sepúlveda Oviedo[a,b], Louise Travé-Massuyès[a], Audine Subias[a], Corinnne Alonso[a], Marko Pavlov[b]

[a]*LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France*
[b]*Feedgy, Paris, France*

Abstract

Diagnosis aims at predicting the health status of components and systems. In photovoltaic (PV) systems, it is vital to guarantee energy production and extend the useful life of PV power plants. Multiple prediction and classification algorithms have been proposed for this purpose in the literature. The accuracy of these algorithms depends directly on the quality of the data with which they are adjusted or trained, i.e., the features. In this paper, an innovative approach for prediction of the health status in PV systems is proposed, which includes a feature selection stage. This approach first discriminates severely affected PV panels using basic electrical features. In a second step, it discriminates the other faulty panels using more elaborated time-frequency features and selecting the most relevant features through correlation and variance analysis. Finally, the approach predicts the health status of PV panels using a nonlinear regression method named partial least squares. This later is then combined to linear discriminant analysis and compared. The approach is validated with real current data from a PV plant composed of 12 PV panels with a power between 205 and 240 Wp in three health states (broken glass, healthy, big snail snails). The results obtained show that the proposed approach efficiently predicts the three health states. It determines the level of degradation of the panels, which indicates priorities to corrective and predictive maintenance actions. Furthermore, it is cost-effective since it uses only electrical measurements that are already available in standard PV data acquisition systems. Above all, the approach is generic and it can be easily extrapolated to other diagnosis problems in other domains.

Keywords: photovoltaic system, fault classification, feature extraction, wavelet transform, multiresolution signal decomposition, unsupervised

learning, fault diagnosis, fault detection, partial least squares, hierarchical clustering, dynamic time warping
PACS: 0000, 1111
2000 MSC: 0000, 1111

1. Introduction

The increase in the demand for electricity, the price of products derived from oil or gas and environmental pollution have driven an increase in the use of renewable energies [1]. Among the most widely used renewable energy resources [2, 3], abundant and clean [4] is photovoltaic (PV) energy used in both small installations and large-scale power plants. As a consequence of the increase in the use of PV energy, improvements in the efficiency of PV cells and the extraction of maximum power, increased the efficiency, stability, reliability and robustness of photovoltaic systems [5]. Along the same lines, these advances revealed the need to improve PV plant supervision systems due to the appearance of recurring undetectable faults [6]. It is necessary that when a fault occurs in the system, it can be detected and classified as quickly as possible and thus carry out predictive or corrective maintenance of the PV system in a timely manner [7].

Faults in PV systems can be caused by aspects such as the useful life of the components, increases in temperature during their operation, external factors (environmental and non-environmental) or interactions between materials [8]. In addition, depending on their location, these faults can affect the AC or DC side of the PV system. On the DC side, some common faults are partial shadowing, hot spots, bypass diode fault, module cracks, faults in the Maximum Power Point Tracking (MPPT) algorithm, laminate discoloration, isolation of cell parts by cracks, delamination, arc faults, among others [9–18]. On the AC side, inverter faults can be found, such as open circuit of switches, short circuit of switches, filter fault and gate fault in the inverter, among others [19–23]. In addition, faults in the inverter protection systems can also occur [14].

Fault diagnosis requires data acquisition and an efficient feature extraction system for diagnosis. However, at present, there is little discussion about experimental data acquisition and processing, although it is a tedious task to design a data acquisition system aimed at diagnosing faults [17]. The lack of consensus in data acquisition may have its origin in the fact that faults are in

many cases temporal due to changes in weather conditions or automatic corrective actions by inverters or optimizers [14]. Furthermore, fault diagnosis in photovoltaic (PV) systems becomes even more complicated in scenarios of low irradiation [24]; when arc impedance may not draw high enough current; when faults occur in less than a second [25]; in presence of the MPPT device that optimizes the output power of a PV array [26]. In this type of scenario, the protection devices fail to activate and even the PV characteristics of a PV panel without faults can be similar to those of a faulty PV panel [27]. This is why faults can go for hours without being detected and not only degrade the state of the PV panel but also cause it to catch fire and pose a danger to human security [28–30]. For all these reasons, early diagnosis of faults in PV systems can sometimes be a real challenge [15] and likewise, an attractive research area that is in full development [14].

To address the issues discussed above, and as a contribution to effective fault diagnosis in PV systems, this paper proposes a new approach to feature extraction and health status prediction in PV systems, based on a commercial Tigo data acquisition platform. Our approach is based on five stages: *1)* data acquisition and pre-processing; *2)* Dynamic Time Warping hierarchical clustering; *3)* feature extraction; *4)* feature selection and *5)* health status prediction.

The contents of the paper are as follows. Section 2 explains related work and describes the proposed approach. Section 3 explains *data acquisition and pre-processing*. Section 4 presents *Dynamic Time Warping based hierarchical clustering*. *Feature extraction* and *feature selection* are presented in Section 5 and Section 6 respectively. Section 7 is dedicated to the PV panels *health status prediction*. Finally, section 8 provides a discussion on the results and conclusions.

2. Related work and description of the approach

In recent years, multiple fault detection and health prediction techniques have been proposed for PV plants. Many are based on the principle of Model-Based Difference Measurement (MBDM), where measured parameters are compared with those predicted by a statistical model [31, 32]. However, these models may be difficult to train and update. Image analysis based approaches have also been presented [33, 34]. Although these methods are efficient for detecting faults such as hot spots, they are poor for detecting faults without thermal expression and are also costly to implement. Another

widely used method for failure analysis is based on visual inspection [35], however this type of method has a high component of subjectivity, high cost and long detection time on a large scale.

For these reasons, methods based on artificial intelligence algorithms have become popular [36, 37]. Some based on clustering methods such as Fuzzy C-means [38] at the solar cell level or hierarchical clustering at the solar panel level have been proposed [38, 39]. In the same way, approaches based on Support Vector Machines (SVM) [40], kernel extreme learning machine [41], decision tree [36], neural networks [37, 42], Local outlier factor [43], Naive Bayes Classifier [44], among others [45–47]. In [48] an approach is proposed explicitly on the production current signal of PV strings. Other more advanced methods combine more than one technique as a hybrid diagnosis technique that takes advantage of each of the methods and significantly improves detection results. For example, [24] proposes a method Based on Multiresolution Signal Decomposition and Two-Stage Support Vector Machines. It is interesting to note that in these works, no special attention is paid to the process of feature extraction for training and likewise these approaches have not been tested on faulty PV panels whose fault signature is similar to that of healthy panels. The complexity of fault diagnosis has even generated methodologies based on semantics [49].

The contribution of this article is precisely to provide a solution to these two last points, as discussed below. First, the panel string current must be captured. Once this data acquisition step is achieved, it is necessary to carry out the extraction of features that allow to discriminate the different classes (health states) of PV panels. To discriminate between these classes, some works propose analyzing the similarity of the signals using elastic metrics [50]. Among these, Dynamic Time Warping (DTW) is one of the algorithms for measuring similarity between two temporal signals that is widely used in clustering and classification [51]. DTW allows to determine the similarity even between out-of-phase signals [52]. DTW is used in conjunction with Hierarchical Clustering (HC) to group signals hierarchically [53, 54], assuming that each observation is a group and the pairs of groups are merged as they move up the hierarchy.

Other slightly more in-depth analyzes propose the use of signal processing and decomposition techniques. Such techniques are also carried out in the PV domain [55]. Signal decomposition techniques such as continuous Fourier transform (FT) and discrete Fourier transform (DFT) are proposed for fault detection. However, these transforms only provide information about fre-

quency. In [56] the authors propose the Fourier Transform with a window that provides both time and frequency information. However, the fixed window selection may not always be efficient for detecting critical non-stationary disturbances such as three-phase faults and short circuits that are associated with transients [2]. Alternatively, in recent years the wavelet transform (WT) started to gain popularity [57–60] due to its multiple resolution time frequency analysis. This type of decomposition shows better identification characteristics of all types of faults in PV systems, as long as the presence of noise in the signal is avoided [2]. Based on the WT, different modifications are proposed such as: the Multiresolution Signal Decomposition (MSD) that apply wavelet decomposition in an iterative way [61], the Slantlet transform [62] which is based on a modified discrete wavelet transform with two zero moments and modified temporal localization and the Wavelet Packet Transform (WPT) which performs an iterative decomposition on the high and low frequency coefficients [18, 63].

Following the decomposition of the signal, the extraction of features is carried out. These features increase the variance between the different classes [64]. Features such as mean, variance, skewness, kurtosis and entropy are suggested for troubleshooting PV systems [14, 15, 64, 65]. Each of these features has a better or worse performance depending on the type of fault to be analyzed. Generally these features are used as input for different fault classification methods [66–69]. However, due to the high dimensionality (high number of features), the computational cost of this classification is very high and there is a high possibility of including irrelevant or redundant information [70, 71]. Therefore, dimensionality reduction methods are proposed to reduce the high dimensionality of features and irrelevant or redundant information with minimal loss of information. This dimensionality reduction is a feature selection step, creating a compressed version of the original feature matrix F_* [72].

For dimensionality reduction or feature selection, visual analysis techniques such as scatter plots of features [73] or the parallel coordinate plot [74, 75] are used individually. However, the identification of relationships between the variables requires a high component of human work and is also a process with low repeatability, since it is subject to the user criteria. On the other hand, methods based on feature correlation are used, such as the *Pearson's correlation matrix* [76]. Although these methods systematically reduce dimensionality, they can still select features with irrelevant or redundant information. Others dimensionality reduction algorithms can be used

such as Principal component analysis (PCA) [15, 64, 76–81], Isometric Mapping (ISOMAP) [82]; Locally Linear Embedding (LLE) [77, 83], Singular-Value Decomposition (SVD) [84]; among many others [79, 85–94] to select features. However, these algorithms, when applied directly to the set of features, compress the irrelevant or redundant information and the correlated features in the same way. For this reason, the approach presented in this article proposes to perform feature extraction and subsequent feature selection as a combination of these methods.

Once the features have been selected, they are generally used to predict or classify the current health status of the individuals [8, 95–97].

On the basis of all these elements, this article presents a new approach for feature extraction and health status prediction in PV. Figure 1 illustrates the five stages of proposal. The first stage performs automatic *data acquisition and pre-processing*. The second stage applies *Hierarchical Clustering* (HC) [98] to the time series issued from the captured signals, for which the time series similarity index of *Dynamic Time Warping* (DTW) [51, 52] is used as distance. This stage performs a coarse grain discrimination, aiming to separate the PV panels in two groups, those whose production is heavily affected (cluster A) and the others (cluster B). The third stage is concerned with *feature extraction*. It is intended to be carried out only on cluster B to achieve refined discrimination. This stage leverages signal decomposition with the *Discrete Wavelet Transform* (DWT) [24, 99] to generate a set of features. The fourth stage called *feature selection* uses correlation and variance analysis to select the appropriate features. Finally, the fifth phase performs the health status prediction of the PV system by two methods. It first uses the *Partial Least Squares* (PLS) algorithm as a prediction method based on a regression model. Then, this phase uses the PLS latent components, obtained as a product of the dimensionality reduction of PLS, as input to the *Linear Discriminant Analysis* (LDA) algorithm to evaluate the results of the prediction with the PLS algorithm and to perform an alternative classification of the health status of the PV panels.

3. Data acquisition and pre-processing

The proposed approach is evaluated on real data from a PV plant located in the LAAS-CNRS laboratory in Toulouse, France. This platform consists of $n = 12$ PV panels with reference *SLK60P6L* from Siliken California with a power between 205 and $240Wp$. The main parameters of these PV panels are

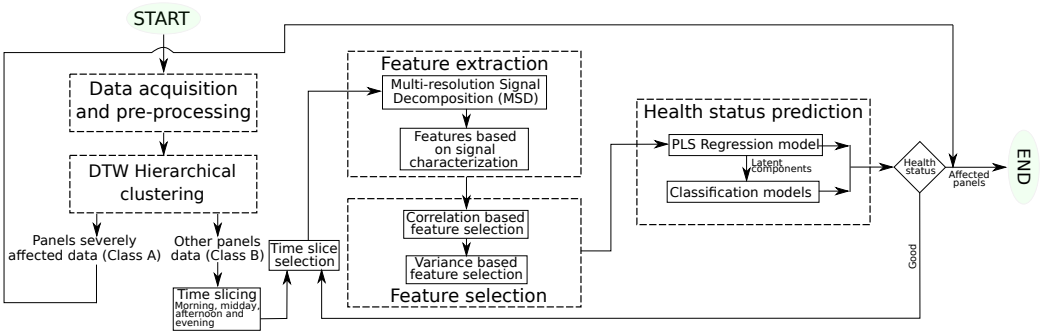


Figure 1: The five stages of the proposed approach. *i)* Data acquisition and preprocessing; *ii)* DTW Hierarchical clustering; *iii)* Feature extraction; *iv)* Feature selection; and *v)* Health status prediction.

Symbol	Quantity	Value
P_{MPP}	Maximum Power (W_p)	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 Current(A)	37.67
S	Area of the module(m^2)	1.64

Table 1: PV module specifications at STC.

given in Table 1 under standard test conditions (STC) ($1000\,W/m^2$, $25°C$). Each PV panel is composed of 60 poly-crystalline silicon cells grouped into 3 sub-strings of 20 cells.

In this study, the current signal I_i for a PV panel PV_i, $i = 1, \ldots, n$, is obtained using a commercial TIGO[1] monitoring platform for PV plants. Each PV panel of the experimental platform is equipped with its own TIGO data acquisition system. All TIGO data acquisitions are piloted by a TIGO reference TS4-R-O optimizer. This TIGO Optimizer is an MPPT device that individually controls each PV panel to achieve maximum performance. To do this, the optimizer constantly monitors the maximum power point MPP. This platform allows to acquire the signals of current, voltage, and power at the Maximum Power Point MPP.

The study is conducted using only the data from the current signal, that

[1]For more information on the TIGO platform please visit here

make up a current database of 12 panels, with signals captured in parallel with a sampling time of one minute for 13 hours from 7:00 a.m. to 8:00 p.m. on June 25, 2020. For a PV panel PV_i, the data takes the form of a time series denoted by $I_{i\{1:n_I\}} = \{i_{i,t_1}, \ldots, i_{i,t_{n_I}}\}$, where n_I is the number of samples of the $i-th$ time series that has a sampling period of one minute and $t_i, i = 1..n_I$, is the date of the sample. The analysis is carried out in a time window of one day. However, it is possible to use the same methodology on different time slices. The data downloaded from the application programme interface (API) of TIGO is not directly ready for signal decomposition and extraction of features. It contains missing or null values that can influence the performance of the next steps of the algorithm. Data cleansing is an elementary phase that must precede all other phases of the algorithm.

The first set of missing or null data is found at the beginning and end of the data set; these data are captured under low irradiation conditions. Because of this specific location in the data set, it is not possible to use conventional methods to impute data such as the arithmetic average. To solve this, it is necessary to verify the voltage and power to identify the real time interval where the PV plant is producing. All data outside this production range is then trimmed. The second group of missing or null data occurs within the PV plant production range. After verifying the records in the Tigo data set and identifying that there are no consecutive null or missing values, mean value, as given in equation (1), is used to replace the missing current values $i_{i,t}$ as:

$$i_{i,t} = \frac{i_{i,t-1} + i_{i,t+1}}{2}, \tag{1}$$

Although the data cleansing process is not complex, it is very efficient and it is an indispensable tool to eliminate most defects that affect the performance of the decomposition algorithms to be applied further. Figure 2 presents the PV panel current behaviors over one day for different health statuses after data cleaning.

The blue color corresponds to the PV panels with a broken glass fault, the yellow color corresponds to the healthy PV panels and the red color to the big snail trail fault. The big snail trail represents corrosion of the sheet of the encapsulation surface and although it does not significantly decrease the performance of the PV panels, it can be the cause of fractures or micro cracks in the modules that reduce the production of a PV panel. As shown in Figure 2, the behavior of the PV panels with a big snail trail is very similar

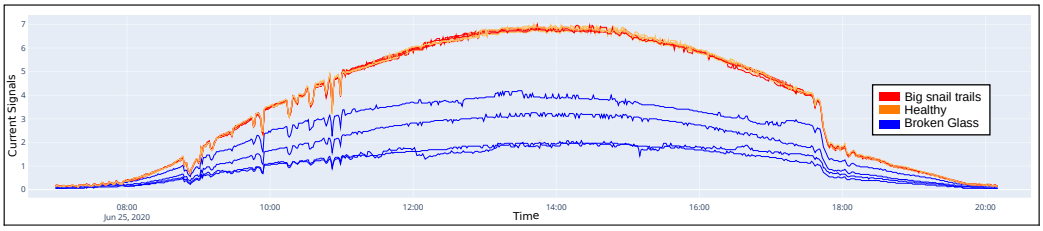


Figure 2: Behavior of the current over one day for different health statuses: healthy (yellow), broken glass (blue), and big snail trails (red) for a period of 13 hours every minute.

to that of healthy PV panels.

4. DTW Hierarchical clustering

In this stage, Hierarchical clustering (HC) is used to construct the two clusters A and B allowing to separate the panels severely affected data from the other panels data (cf. Figure 1) based on the similarity of the current time series $I_{i\{1:n_I\}}$ of the different PV panels PV_i, $i = 1, \ldots, n$. The time series similarity index is taken as the Dynamic time warping (DTW) index due to its well-known performance [50, 100].

In the following subsections, HC and DTW are presented for generic time series that are then instantiated to the current times series $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$ of our case study.

4.1. Dynamic time warping

DTW is a well-known technique that is based on the principle of dynamic programming to deform two temporal sequences in a non-linear way and find optimal alignments between them [101, 102]. To measure the similarity between two time series $S_{\{1:\eta_s\}}$ and $T_{\{1:\eta_t\}}$ the matrix of distances D of dimensions $(\eta_s \times \eta_t)$ is built. Each entry $d(i, j)$ corresponds to a local distance between S and T given by the Euclidean distance between s_i, $i = 1, \ldots, \eta_s$ and t_j, $j = 1, \ldots, \eta_t$.

A valid warping path $W_k = \{w_{k,1}, \ldots, w_{k,\eta_{W_k}}\}$, where η_{W_k} is the number of elements of the path W_k in matrix D, is defined using the above distances and satisfying the three following constraints:

1. Endpoint constraints: $w_{k,1} = d(1, 1)$ and $w_{k,\eta_{W_k}} = d(\eta_s, \eta_t)$.

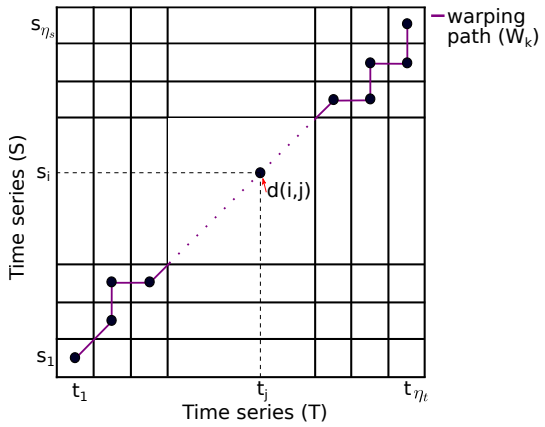


Figure 3: Example of warping path in the distance matrix $\mathbf{D}$. Each entry $d(i,j)$ represents a local distance between the time series S and T given by the euclidean distance between each point s_i, and t_j.

2. Monotonicity constraint: If $w_{k,\alpha+1} = d(i,j)$ and $w_{k,\alpha} = d(i^{'},j^{'})$, then $i \geq i^{'}$ and $j \geq j^{'}$, $\forall \alpha = 1,\ldots,\eta_{W_k}$
3. Continuity constraint: If $w_{k,\alpha+1} = d(i,j)$ and $w_{k,\alpha} = d(i^{'},j^{'})$, then $i \leq i^{'}+1$ and $j \leq j^{'}+1$, $\forall \alpha = 1,\ldots,\eta_{W_k}$

Let us define $\mathbb{W}$ as the set of valid warping paths and $W_k^{\oplus}$ as the sum of elements of a valid warping path W_k, i.e., $W_k^{\oplus} = \sum_{p=1}^{\eta_{W_k}} w_{k,p}$. Therefore, the DTW (S,T) distance is given by the minimum warping path among all valid paths in D:

$$\text{DTW}\ (S,T) = \min_{W_k \in \mathbb{W}} W_k^{\oplus}. \tag{2}$$

A more detailed description of DTW is presented in [54, 100]. Figure 3 illustrates the principle of DTW .

The results of DTW are used as input to a hierarchical clustering algorithm.

4.2. Hierarchical clustering

Agglomerative hierarchical clustering (AHC) is a well-known method that allows several individuals to be grouped into clusters according to the degree of similarity between the individuals. For this, the algorithm uses a degree of similarity between individuals and groups, and between groups [102]. Then in each iteration, the groups with the shortest distance are merged into a

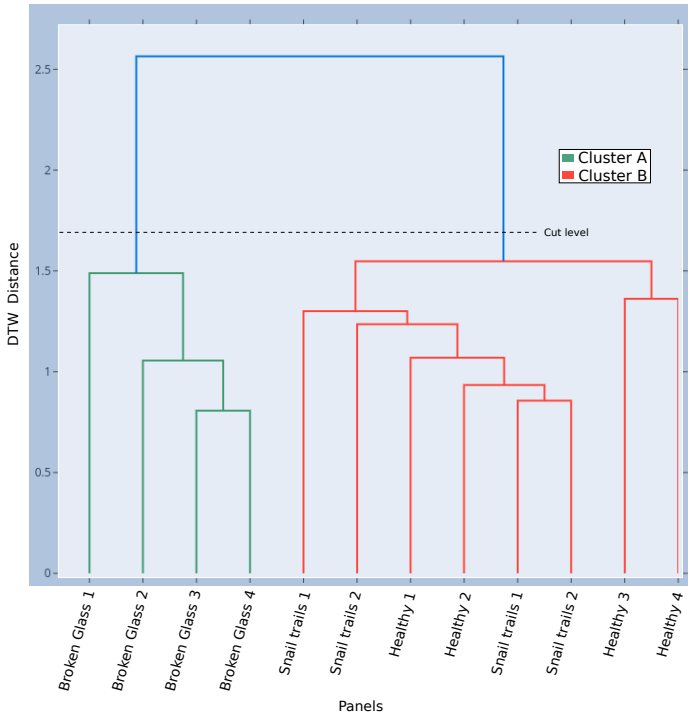


Figure 4: Dendrogram of the agglomerative hierarchical clustering of current signals. Cluster A in green groups the severely affected panels (broken glass). Cluster B of color B groups the healthy panels and those with big snail trails.

single cluster [103, 104] from bottom to top in the hierarchical grouping. This process continues until reaching the final condition [105, 106]. The result of the clustering is generally presented in the form of a tree called dendrogram [54]. The final clustering of the AHC depends on the level at which the dendrogram is cut [98].

This algorithm is applied to the time series of the current $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$. The degree of similarity is given by the DTW. The result of the hierarchical clustering on the current signals of the PV panels is presented in Figure 4.

As shown in Figure 4, the PV panels are grouped into two large clusters A (green color) and B (red color). Since the group of PV panels from cluster A is easily discriminable, the detailed analysis of the third stage is applied only on the PV panels of cluster B. In order to analyze in detail, the behavior of the PV panels of cluster B under the different irradiation conditions of the day, the signals are divided into 4 slices called: morning, midday, afternoon

and evening.

The features extraction is carried out on each of these slices. In this article, it is assumed that if the classes are discriminated in at least one of the slices, the algorithm is efficient and it is possible to detect anomalies between the PV panels of group B. In order to explain and illustrate our approach, feature extraction is explained and illustrated using the midday slice as an example.

5. Feature extraction

This third stage is based on Multiresolution Signal Decomposition, followed by the extraction of statistical features as proposed in [14, 15, 64, 65].

5.1. Multi-resolution signal decomposition

Every faulty condition in a PV system is associated with a change in the output current. These changes are reflected as variations in the waveform of the output signal compared to a healthy PV panel. Some of these changes are visible in the frequency domain and others in the time domain. In order to analyze these changes simultaneously (time - frequency), Multi-resolution Signal Decomposition is used. The Multiresolution Signal Decomposition is based on the discrete wavelet transform (DWT) that can decompose a signal into levels with different time and frequency resolutions using the wavelet transform iteratively [17].

In the following, DWT is presented in a generic form. In our case study, it is applied to the current times series $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$.

DWT is a signal processing technique (linear transformation) like the Fourier transform [25]. Some of the differences between these two techniques can be read in [107]. DWT decomposes the input signal into a variable frequency range that depends on the *mother wavelet* selected as the decomposition pattern [2]. The input signal is decomposed into approximate and detailed coefficients that correspond to the high and low frequency components respectively. DWT is known for its properties to simultaneously analyze frequency and time [108–110]. As mentioned in [25], the wavelet transformation with the proper *mother wavelet* is a useful tool for fault detection and feature extraction. For this reason DWT is widely used in this field [2, 111–113].

Wavelet decomposition uses a *mother wavelet* that decomposes the signal into a set of oscillatory functions called wavelets. Each of these *mother wavelets* is a signal in time that captures a specific frequency band [114, 115]. There are different well-known families of *discrete mother wavelets* such as: Harr, Meyer, Bior, Daubechies, Rbio, Coiflet and Symmlet. Which are composed of 1,1,15,38,15,17 and 19 *mother wavelets* respectively. Each of these *mother wavelets* has a different computational calculation speed and decomposition quality depending on the particular application.

Some *mother wavelets* are particularly used in the PV domain, for example: Sym8 [61] from the Symmlet family, Harr [18], and db1, db3 - db5, db8, db9, [2, 14, 15, 17, 25, 61, 64, 65] from the Daubechies family. Each of these wavelet families is defined according to Equation (3) [65, 116].

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \tag{3}$$

where a is the scale or dilation factor, b is the shifting factor, t refers to the timestamp of the input signal, and ψ is defined as the *mother wavelet* [116]. To restrict the values of a and b to discrete values, these factors are defined according to Equations (4) and (5) [24, 116].

$$a = a_0^{-(m_x/2)}, \tag{4}$$

$$b = n_x b_0 a_0^{m_x}, \tag{5}$$

where m_x and n_x range over $\mathbb{Z}$ and $a_0 > 1$ and $b_0 > 0$ are fixed [116]. The DWT of the discrete signal $X_{\{1:n_X\}}$ that uses the *mother wavelet* $\psi_{a,b}(t)$ of Equation (3) is described in Equation (6)[24, 64, 65, 116]:

$$\text{DWT } (a,b) = \frac{1}{\sqrt{a}} \sum_{1:n_X} X(t)\psi\left(\frac{t-b}{a}\right), \tag{6}$$

For the decomposition of the signal, it is necessary to select the appropriate *mother wavelet* $\psi_{a,b}(t)$. Some works proposed complex algorithms for the optimal selection of the *mother wavelet* [117]. In this article, the mother wavelet selection follows the work of Wang et al. [17] that aims at detecting faults in PV systems with wavelet transform. According to Wang et, al., [17] the selected wavelet must comply:

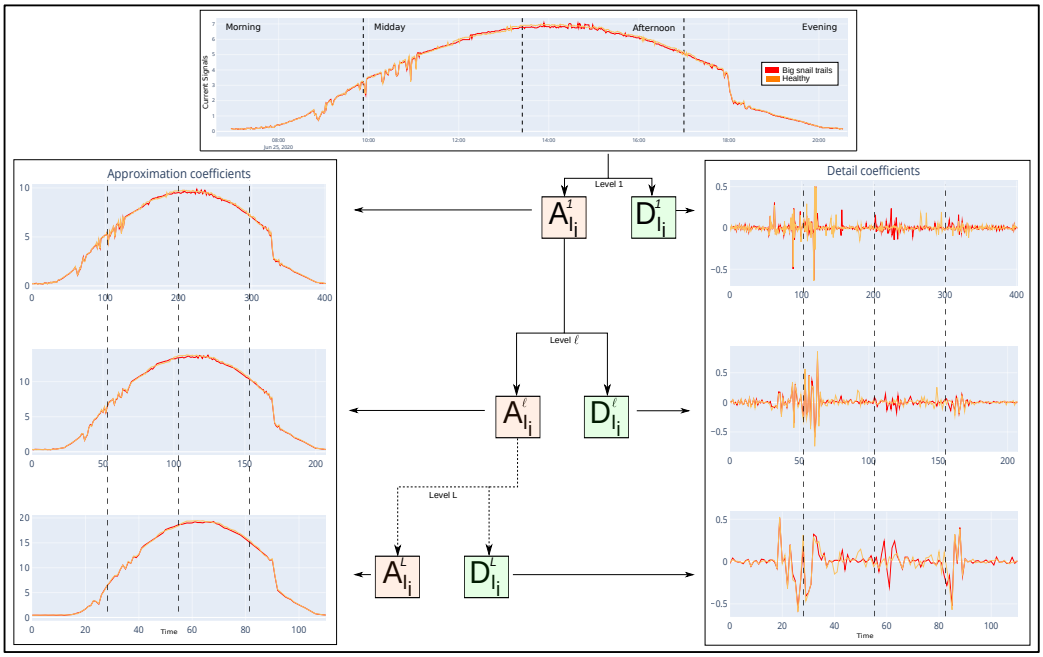


Figure 5: Decomposition into 3 levels of the current signal for a panel with big snail trails (red) and a healthy panel (yellow). The approximation and detail coefficients resulting from the decomposition are presented on the left and right of the figure, respectively.

1. To have a sufficient number of vanishing moments to represent the salient features of the anomalies.
2. To provide sharp cutoff frequencies to reduce the amount of leakage energy into the adjacent resolution levels.
3. The wavelet basis should be orthonormal.

Taking into account these considerations and the fact that most of the work in fault detection in PV systems use wavelets of the Daubechies family (db), the entire family is tested and the Daubechies38 (db38) *mother wavelet* is selected due to its computational speed and good decomposition result.

Multi-resolution signal decomposition can be performed at different levels of decomposition. The result of the decomposition into 3 levels for a current signal I_i is shown in Figure 5. At each level ℓ of decomposition of a current signal I_i, two signals can be created as the result of the wavelet transform. The first signal corresponds to the approximation coefficients ($A^{\ell}_{I_i}$). This signal receives this name due to the fact that it is an approximation of the "low frequency" components of I_i. The second signal corresponds to the detail

coefficients ($D^{\ell}_{I_i}$). This signal represents the "high frequency" corrections of the signal I_i.

The original signal I_i can be reconstructed from the detail and approximation coefficients. The reconstructed signal $I_{i,r}$ is the sum of all the detail coefficients prior to the last selected level L, with the detail and approximation coefficients of level L. This description is formally presented in Equation (7) [17, 118].

$$I_{i,r} = A^{L}_{I_i} + \sum_{\ell=1}^{L} D^{\ell}_{I_i}, \tag{7}$$

5.2. Features based on signal characterization

In this subsection, features are first presented for a generic signal. Then the signals that are used for their extraction in our case study are made explicit.

For a given generic signal X represented by a time series $X_{\{1:n_X\}}$, a number of features can be extracted. Note that the selected features retain only some characteristics of the signal, which has an impact on the possible discrimination of different signals. The n_F selected features have been chosen to capture several characteristics of a signal. These selected features have been considered as they are also used in previous works aimed at fault diagnosis in PV systems [2, 14, 15, 64, 65, 119–121] and works aimed at fault diagnosis in vibration signals [66, 67, 79, 122]. Given a time series $X_{\{1:n_X\}}$ of mean μ, these features are :

- *Skewness (F1)*: skewness represents the asymmetry of the data with respect to the mean and is calculated by Equation (8) [76].

$$F1 = \frac{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t - \mu)^3}{\left(\sqrt{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t - \mu)^2}\right)^3}, \tag{8}$$

- *Kurtosis (F2)*: kurtosis measures the peak of the probability distribution of the data. It also allows knowing how prone to outliers is a distribution. Kurtosis is defined according to Equation (9) [76].

$$F2 = \frac{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t - \mu)^4}{\left(\sqrt{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t - \mu)^2}\right)^4}, \tag{9}$$

- *Variance (F3)*: the variance represents the variability of a series of data with respect to its mean.

$$F3 = \frac{1}{n_X} \sum_{t=0}^{n_X} (X_t - \mu)^2, \tag{10}$$

- $P-P_{value}$ *(F4)*: the peak-to-peak distance (p-p) is the distance between the peak with the highest amplitude and the valley with the lowest amplitude.

$$F4 = max(X_t) - min(X_t), \tag{11}$$

- *Energy (F5)*: explain the energy contained in the signal, it is conserved regardless of whether it is in frequency or in time [2].

$$F5 = \sum_{t=0}^{n_X} X_t^2, \tag{12}$$

The characterization of the operational condition of a PV panel PV_i is performed with the set of $L+1$ time series $\{A_{I_i}^L, D_{I_i}^\ell, \ell = 1, \ldots, L\}$, obtained from the L levels multi-resolution decomposition of the corresponding current signal I_i. These time series are segmented in four time slices corresponding to morning, midday, afternoon, and evening. Sliced signals are indexed accordingly by $* \in \{\ morning, midday, afternoon, evening\ \}$ and we obtain the set $S \in \{A_{I_{i,*}}^L, D_{I_{i,*}}^\ell\}$, $i = 1, \ldots, n_B$, $\ell = 1, \ldots, L$, where n_B is the number of PV panels in cluster B.

The selected features are then determined for each time series in S, forming a feature vector composed by the feature subvector $FA_{I_{i,*}}^L$ for the approximation coefficients and the features subvectors $FD_{I_{i,*}}^\ell$ for detail coefficients, each sub vector being of dimension n_F. The characterization of every time slice can be summarized in a matrix of dimensions $n_B \times ((L+1) \times n_F)$:

$$\mathbb{F}_* = \begin{pmatrix} FA_{I_{1,*}}^L & FD_{I_{1,*}}^1 & \ldots & FD_{I_{1,*}}^\ell & \ldots & FD_{I_{1,*}}^L \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ FA_{I_{n_B,*}}^L & FD_{I_{n_B,*}}^1 & \ldots & FD_{I_{n_B,*}}^\ell & \ldots & FD_{I_{n_B,*}}^L \end{pmatrix}, \tag{13}$$

$$* \in \{\text{morning, midday, afternoon, evening}\}$$

Each row $\mathbb{F}_*(i,.)$ of the matrix $\mathbb{F}_*$ provides the signature of the health state of the PV panel PV_i.

It is important to mention that some of the features in each high dimensionality signature may provide redundant information, which may reduce the performance of data-based diagnosis algorithms [70][71]. Therefore, it is necessary to select the outstanding features by means of an algorithm that identifies a subset of features that preserve the fine details related to a faulty state as represented in the high dimensional space. As mentioned in [76], a high dimensionality data set could be reduced by brute-force with an exhaustive search enumerating and testing all the feature subsets. However, it is more efficient to use feature selection and feature reduction algorithms. For this reason, a two-stage cascading dimensionality selection and reduction method is proposed below.

6. Feature selection

For a given generic matrix of features $\mathbb{F}$ of dimensions $(n_B \times \eta_b)$, whose η_b columns represent the features that characterize the health status of n_B individuals, a set of $\eta_c^{\oplus}$ features, where $\eta_c^{\oplus} \subseteq \eta_b$ features that preserve relevant details for class discrimination can be selected. The selection of the $\eta_c^{\oplus}$ features is first based on correlation, then on variance analysis. In a first selection step, highly correlated features are discarded. Then the remaining weakly correlated features are given as input to the variance based feature selection algorithm. Without lack of generality, feature selection is presented for the matrix of features $\mathbb{F}_{morning}$ obtained with 3 levels of decomposition. It can be easily extrapolated to L levels of decomposition in any of the 4 time slices of interest.

6.1. Correlation based feature selection

Correlation based feature selection allows to choose a subset of η_c uncorrelated relevant features with high predictive value to create solid learning models for the n_B individuals in matrix $\mathbb{F}$. In the literature, it has been previously mentioned that a feature is redundant if one or more other features are highly correlated with it. The use of the *Pearson's correlation matrix* for these analyzes has been proposed in social science works [123–125]. Correlation based feature selection uses the *Pearson's correlation matrix* to

determine the degree of correlation between the initial features η_b. The level of correlation between two features ranges between -1 and 1, with 1 being the highest positive correlation and -1 the highest negative correlation. 0 indicates no correlation at all. The more a feature is correlated to another, the less information it brings while it can introduce noise. Thus, it is recommended to eliminate it [76]. A correlation threshold $\tau_{\mathbb{F}}$ is defined to remove the correlated features that are out of the range $[-\tau_{\mathbb{F}}, \tau_{\mathbb{F}}]$ and form a set of uncorrelated features of cardinal η_c that will be used for class discrimination and that reduce the matrix $\mathbb{F}$ into $\mathbf{F}$.

The selection of the uncorrelated features corresponding to the columns of the matrix $\mathbb{F}_*$ that contain the relevant details of the health states of each PV panel PV_i is performed with a correlation threshold $\tau_{\mathbb{F}_*} = 0.9$. As an example, the correlation based feature selection on the matrix $\mathbb{F}_{morning}$ is presented in Figure 6. Figure 6a provides the correlation matrix crossing the η_b initial features before feature selection. Figure 6b provides the correlation matrix crossing the η_c weakly correlated features after eliminating the strongly correlated features. With this feature selection, the number of features is decreased from 20 to 14 uncorrelated features for the matrix $\mathbb{F}_{morning}$. In other words, the feature dimension is reduced by 40%. Correlation based feature selection is carried out for each matrix $\mathbb{F}_*$, obtaining the matrices $\mathbf{F}_*$, where $* \in \{morning, midday, afternoon, evening\}$.

6.2. Variance based feature selection

Now, it is not because features are not strongly correlated that they have a strong discriminating power for a classification problem. For this reason, a feature selection based on variance is also applied. For this purpose, *parallel coordinates* is used. This technique, based on the variability of the features [75], is widely used in multivariate data analysis [74]. In the parallel coordinates, there are as many normalized axes as features.

For a given matrix of features $\mathbf{F}$ of dimensions $(n_B \times \eta_c)$, whose η_c columns represent the uncorrelated features that characterize the health status of n_B individuals, there are $\eta_c^{\oplus}$ features, $\eta_c^{\oplus} \subseteq \eta_c$, that preserve relevant details and present significant variance between the n_B individuals. To select the $\eta_c^{\oplus}$ features, the variance of the η_c features is compared between the rows $\mathbf{F}(\mathbf{i}, .)$, $i = 1, \ldots, n_B$ of the matrix $\mathbf{F}$. Those features that do not show a significant variation are not selected to form the final set of $\eta_c^{\oplus}$ features, reducing the matrix $\mathbf{F}$ into the matrix F of dimension $(n_B \times \eta_c^{\oplus})$.

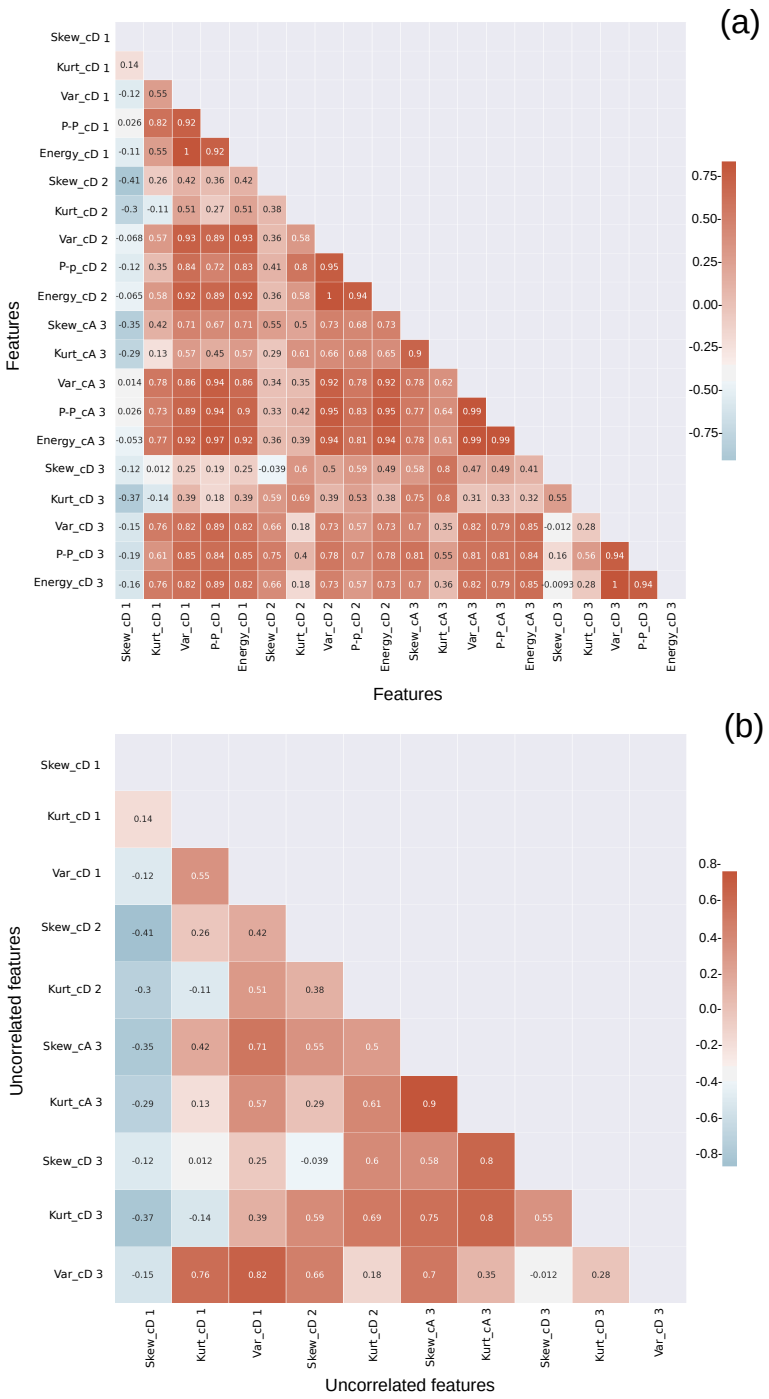


Figure 6: Correlation matrices: (**a**) Pearson correlation matrix of the η_b initial features of $\mathbb{F}_{morning}$, i.e., before correlation based feature selection; (**b**) Pearson correlation matrix of the η_c uncorrelated features of $\mathbf{F_{morning}}$, i.e., after correlation based feature selection.

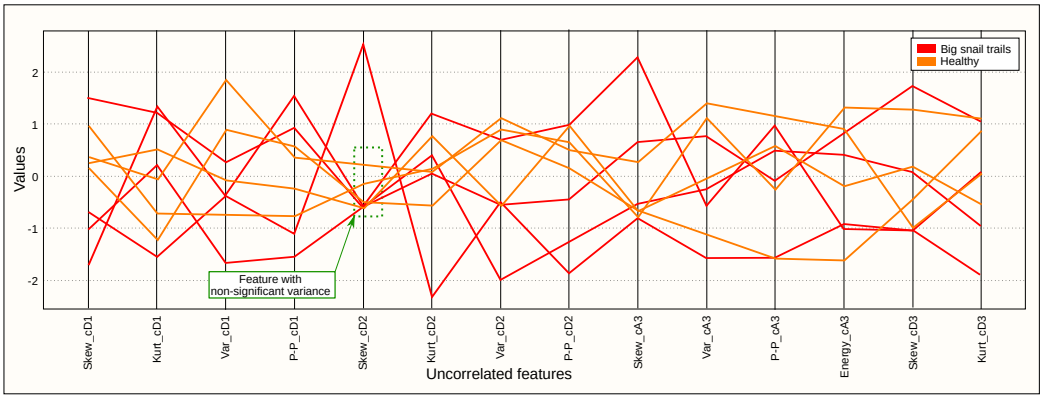


Figure 7: Parallel coordinates plot on the matrix $\mathbf{F_{morning}}$. The normalized values of the uncorrelated features η_c are plotted on the vertical axis. The horizontal axis represents the uncorrelated features η_c.

To illustrate the variance based feature selection, Figure 7 shows the plot of parallel coordinates on the matrix $\mathbf{F_{morning}}$ resulting from the correlation based feature selection. The horizontal axis of Figure 7 represents the η_c weakly correlated features and the vertical axis represents their normalized values. As shown in Figure 7, the feature $Skew\_CD2$ (corresponding to skewness extracted on the detail coefficients at level 2) does not provide significant variance to distinguish between the different operating states (healthy and big snail trails). Therefore, this feature is not selected to form the final set of features $\eta_c^{\oplus}$. By discarding this feature, the matrix $\mathbf{F_{morning}}$ that had 14 features is reduced to matrix $F_{morning}$ with 13 features, maintaining the relevant information to identify the different operating conditions of the PV panels.

Variance based feature selection is applied to all matrices $\mathbf{F}_*$, where $* \in \{\ morning,\ midday,\ afternoon,\ evening\ \}$,leading to four reduced feature matrices F_* of 13, 12, 11 and 16 dimensions respectively.

7. Health status prediction

In the case study, applying feature selection to $\mathbb{F}_*$, where $* \in \{\ morning,\ midday,\ afternoon,\ evening\ \}$, leads to four reduced feature matrices F_*. The selected features aim to solve four classification problems of the health status of the PV panels. Each of these problems can be formulated as a prediction problem based on a regression model or a classification problem

where the response variable is the label, the predictors being the features obtained in section 6.

The PLS algorithm provides very interesting results over other conventional methods when the objective is class prediction [95]. In addition, PLS defines latent components that can be subsequently used as predictors in a classification problem, providing an alternative method to prediction or a validation method of the results of the PLS based prediction. In this sense, the PLS algorithm can be seen as a dimension reduction method that is coupled with a regression model. It performs dimensionality reduction and classification based on regression simultaneously [126].

7.1. PLS Regression model

The PLS algorithm is based on the iterative nonlinear partial least squares algorithm (NIPALS) [127, 128] adapted to reduce the dimensionality in ill-conditioned over-determined regression problems [95]. Assume the matrix of predictors to be given by a centralized and normalized matrix F of dimension $(n_B \times \eta_c^{\oplus})$, and the matrix of targets or response variables be given by a matrix Y of dimension $(n_B \times q)$. The PLS algorithm is based on the decomposition of Y and F into latent components T such that:

$$Y = TQ^T + U, \tag{14}$$

$$F = TP^T + E, \tag{15}$$

where, P and Q are matrices of coefficients, of dimensions $(\eta_c^{\oplus} \times \eta_{PLS})$ and $(q \times \eta_{PLS})$ respectively, that show how the latent components are related to F and Y. E and U are matrices of random errors of dimensions $(n_B \times \eta_c^{\oplus})$ and $(n_B \times q)$ respectively. Finally, T is a $(n_B \times \eta_{PLS})$ matrix giving the uncorrelated latent or PLS components of n_B observations. T can be seen as a linear transformation of F given by Equation (16).

$$T = FK, \tag{16}$$

where K is a $(\eta_c^{\oplus} \times \eta_{PLS})$ matrix of weights. The columns of T and K are denoted as $T(., h) = (t_{1,h}, \ldots, t_{n_B,h})^T$ and $K(., h) = (k_{1,h}, \ldots, k_{\eta_c^{\oplus},h})^T$, $h = 1, \ldots, \eta_{PLS}$. The rows of F are denoted as $F(j, .) = (f_{j,1}, \ldots, f_{j,\eta_c^{\oplus}})$,

$j = 1, \ldots, n_B$. Based on Equation (16), each term $t_{j,h}$ of $T(., h)$ is calculated according to:

$$t_{j,h} = (f_{j,1}, \ldots, f_{j,\eta_c^{\oplus}}) * (k_{1,h}, \ldots, k_{\eta_c^{\oplus},h})^T = \sum_{i=1}^{\eta_c^{\oplus}} f_{j,i} k_{i,h}, \tag{17}$$

where each element $k_{i,h}$, $i = 1, \ldots, \eta_c^{\oplus}$, corresponds to the normalized covariance of the response variable with each predictor given by:

$$k_{i,h} = \frac{Cov(f_{j,i}, y_j)}{\sqrt{\sum_{i=1}^{\eta_c^{\oplus}} Cov^2(f_{j,i}, y_j)}}, \tag{18}$$

Once T is constructed, the matrix Q^T is obtained as the least squares solution of the equation (14). Then, the regression model is defined according to:

$$Y = FB + U, \tag{19}$$

Where, B is a $(n_B \times q)$ matrix of regression coefficients defined according to:

$$B = KQ^T, \tag{20}$$

7.2. Prediction of the health status

In the case study presented in this article, the response variables Y are categorical. In other words, each response variable y_i, $i = 1, \ldots, n_B$, of the matrix Y takes only one of the possible n_B unordered values. For example, in our case, each categorical variable y_i takes the value of $y_i = 2$ (big snail trails), $y_i = 3$ (healthy) or $y_i = 0$ otherwise.

In the proposed approach, we first use the non-linear PLS algorithm as a dimensionality reducer. In [129], the PLS and other dimensionality reduction algorithms are analyzed. Particularly in categorical scenarios, dimensionality reduction using PLS shows results similar to PCA [126] with high prediction accuracy [130, 131]. The set of components that are obtained as a result of dimensionality reduction using PLS is called the set of PLS latent components. These PLS latent components are used for the prediction of the health status based on the regression model of Equation (19). The PLS is fitted with 60% of the data and tested with the remaining 40% of the data.

In order to evaluate the accuracy of the regression model, the complementary metrics Root Mean Squared Error ($RMSE$) and R-Squared or Coefficient of determination metrics (R^2) are used. The $RMSE$ measures the standard deviation between the predicted values and the actual values of the observation [132]. A number close to zero implies a high precision of the model. The $RMSE$ for n_B samples is defined as:

$$RMSE = \sqrt{\frac{1}{n_B}\sum_{i=1}^{n_B}(y_i - \hat{y_i})^2}, \tag{21}$$

where y_i are observed values and $\hat{y_i}$ are the fitted values of the response variable Y for the ith case. The $RMSE$ does not provide information about the explained component of the regression fit [133]. Because of this, the metric R^2 is used in a complementary way. R^2 measures the percentage of variation in the response variable Y explained by the predictors F [133]. The value of R^2 ranges from 0 to 1, where 1 corresponds to the best prediction and 0 corresponds to a poor prediction. The R^2 metric for n_B samples is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_B}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n_B}(y_i - \bar{y})^2}, \tag{22}$$

where $\bar{y} = \sum_{i=1}^{n_B} y_i$ represents the mean value of the response variable Y. Similarly, the confusion matrix is used as a tool for evaluating the performance of the PLS algorithm. The confusion matrix represents a count of the number of accurately classified negative and positive samples represented as True Negative (TN) and True Positive (TN) respectively. Also, it represents the number of real negative samples classified as positive stands for False Positive (FP) and the number of real positive samples classified as negative stands for False Negative (FN) [134].

The results of the prediction of health status for all matrices F_*, where $* \in \{\ morning, midday, afternoon, evening\ \}$, are reported in Figure 8.

As can be seen in Figure 8, the PLS algorithm is able to correctly predict 7 of the 8 PV panels of cluster B in the 4 time slices. In the $Midday$ time slice, it is possible to observe how the PLS algorithm classifies a Big Snail Trail panel as a new different class (label 0). Furthermore, the performance of the prediction of the PLS method on the time slices $Midday$ and $Afternoon$ is related to the similarity of the current signals between the PV panels PV_i,

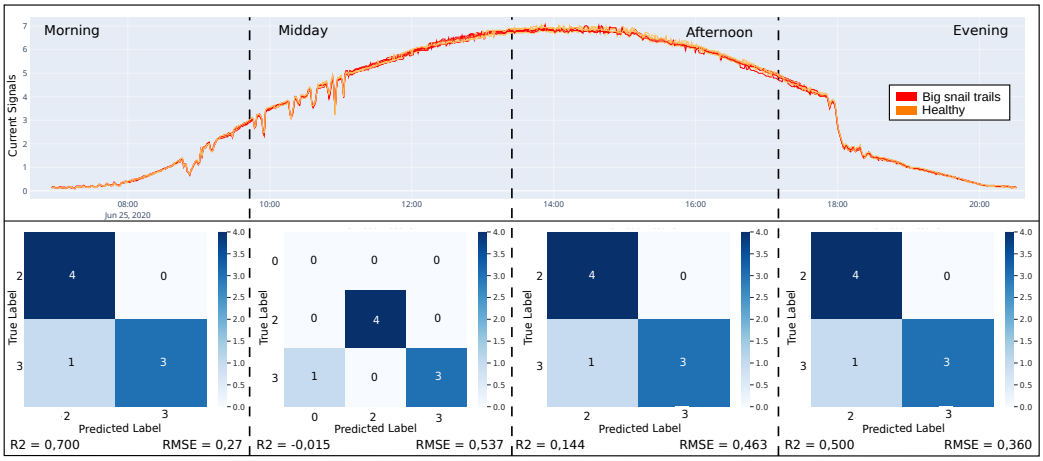


Figure 8: Prediction of the health status of the PV panels of cluster B with PLS. Prediction accuracy with R^2 and $RMSE$ metrics for the four time slices *morning*, *midday*, $afternoon$, and *evening* with PLS.

$i = 1, \ldots, n$ when solar irradiation is at its highest value. In the same Figure 8, analyzing the value of R^2 and $RMSE$, in the $Morning$ and $Evening$ slices, it is possible to observe that the model can explain the 70% and 50%, respectively, of what is happening in the actual data. While in the $Midday$ and $afternoon$ slices, the model reaches a maximum of 14% of the data. The performance of the PLS algorithm is strongly affected by the number of individuals who are used to fit the model.

7.3. PLS-LDA classification method

Alternatively, a health status classification method that uses the PLS latent components (given by T) as input of a classical classification method is proposed. The use of PLS as a dimension reducer for classification problems is studied in [95–97]. This method allows to classify the health status and to validate the health status results generated with the PLS prediction of section 7.2.

The classification algorithm has been selected to be Linear Discriminant Analysis (LDA) due to the interesting results reported when it is used with the PLS dimensionality reduction [97, 135]. In addition, this algorithm has already been used in fault detection in PV systems [8]. The LDA algorithm projects the original data matrix T (predictors) from a high-dimensional space into a new low-dimensional space that makes within-class scatter as small as possible and between-class scatter as large as possible.

Given a number of classes G, the LDA determines the center class φ_{C_g}, $g = 1, \ldots, G$, for each class C_g according to:

$$\varphi_{C_g} = \frac{1}{n_e} \sum_{i=1}^{n_e} e_i, \tag{23}$$

where n_e is the number of elements e_i in class C_g. Then, the LDA algorithm computes the within-class S_W and the between-class S_B scatters. The S_W is calculated according to:

$$S_W = \sum_{g=1}^{G} S_{C_g}, \tag{24}$$

where S_{C_g} is defined as:

$$S_{C_g} = \sum_{i=1}^{n_e} (e_i - \varphi_{C_g})(e_i - \varphi_{C_g})^T, \tag{25}$$

The between-class scatter S_B is calculated according to the expression:

$$S_B = \sum_{i=1}^{G} (\varphi_g - \varphi)(\varphi_g - \varphi)^T, \tag{26}$$

where, φ is the mean value of all data in matrix T. Finally, the LDA finds a linear projection v that discriminates as much as possible the set of classes of the data. This projection is obtained by maximizing the expression:

$$J(v) = \frac{v^T S_w v}{v^T S_B v}, \tag{27}$$

The discriminant axes of v have as eigenvalues $\lambda_1, \ldots, \lambda_{\eta_{PLS}}$ and correspond to the decomposition of the matrix $S_w S_B^{-1}$. This decomposition into eigenvalues defines the projection space of the original data of the matrix T. To evaluate the degree of correct predictions (ability to identify positive and negative samples) the confusion matrix and the F-Value are used. The F-Value metric does not take into account the true negatives (TN), for this reason, in cases of unbalanced classes it improves the perception of the performance of the algorithm [136]. The F-Value ranges from 0 to 1, where 1

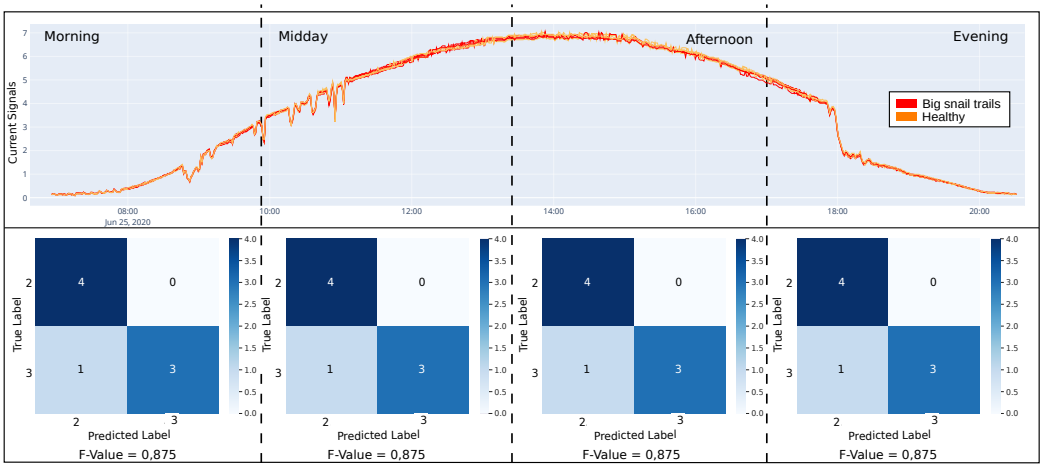


Figure 9: Prediction of the health status of the PV panels of cluster B with $PLS-LDA$. Prediction accuracy with $F-Value$ metric for the four time slices for the time slices *morning*, *midday*, $afternoon$, and *evening* with $PLS-LDA$.

indicates the best performance and 0 the worst. The F-value is defined as:

$$F-Value = 2 * \frac{precision * recall}{precision + recall} \tag{28}$$

where the precision, $precision = TP/(TP+FP)$, allows to measure the cost of false positives. The recall, $recall = TP/(TP+FN)$, allows estimating the number of individuals correctly classified as true positives compared to the total number of elements belonging to the class. The LDA algorithm is trained and tested with the same PV panels that fit the model of section 7.2. The total number of components generated in dimensionality reduction using PLS are used. The classification results using the PLS-LDA method, together with the F-Value and the confusion matrix for each time slice are presented in Figure 9.

As shown in Figure 9, with the exception of the $Midday$ time slice, in the other time slices the PLS-LDA method classifies the PV panels in the same classes as using PLS algorithm. In the $Midday$ time slice the different class (label 0) generated by the PLS algorithm is removed. As a summary, Table 2 presents the final prediction accuracy for the time slices *morning*, *midday*, $afternoon$, and *evening* of the PLS-LDA and PLS methods.

As seen in Table 2, the PLS-LDA method classifies the four time slices with an F-value of 0.875 (high precision) compared to the prediction accuracy

Time slice	PLS		PLS-LDA
	R^2	RMSE	F-Value
Morning	0,700	0,274	0,875
Midday	-0,015	0,537	0,875
Afternoon	0,144	0,463	0,875
Evening	0,500	0,360	0,875

Table 2: Prediction accuracy for the four time slices with the PLS and $PLS - LDA$ methods.

presented by the PLS method that does not give homogeneous results for all the time slices (see Midday line in Table 2. Let us recall that in this article it is considered that if it is possible to discriminate healthy PV panels from another set of PV panels in at least one time slice, then it is possible to establish which PV panels are faulty with the available data.

8. Discussion and Conclusions

The approach presented in this article responds to current energy concerns regarding the guarantee of continuous energy production in photovoltaic systems. These systems distribute approximately 2% of the energy consumed in the world [137] and present annual losses of around 18.9% of power due to the presence of faults [138].

This work proposes and develops a health state prediction dedicated to photovoltaic systems. The method is based on a set of features all extracted from the MPP current signal. This approach was tested with a string of 12 photovoltaic panels and validated for efficiency by separating three different health scenarios: healthy, big snail trail, and broken glass.

To summarize, the approach uses, in a first stage, a simple hierarchical clustering based on Dynamic Time Warping, to group the PV panels into two clusters A and B, where cluster A contains the severely affected PV panels and group B contains the others. At this early stage, the method clearly discriminates between healthy and broken glass types, which points at priority predictive maintenance actions and reduces overall costs consequently. In a second stage, the use of a set of in-depth time-frequency features allows for a more precise approach to detect tiny faults and shows its ability to discriminate weakly affected panels from healthy panels. The second stage was validated by advantageously identifying photovoltaic panels with big snail

trail faults despite the difficulty of discriminating them from healthy panels. This represents a clear contribution with respect to previous works such as [139] that fails to detect faults whose behaviors is highly similar to that of healthy panels. It is also important to highlight that our method has the clear advantage to require very simple monitoring. Indeed, only the MPP current is required. Nowadays, this type of detection can only be achieved by regularly visiting the PV plant, which is extremely more expensive.

A further advantage is that the approach proposed in this paper only requires a reduced number of individuals of each class, which reduces the cost of data acquisition and storage. Another interesting point is that faults that occur under low irradiation (*Morning* and *Evening*) are generally the most difficult to diagnose, however, the proposed method presents the best performance in these situations.

Another contribution is to base the diagnosis process on four time slices of the day. The detection of a fault in a time slice may grow into a serious fault later or vanish simply inducing a slight loss of performance. The method hence provides information about specific time points of the day that should be monitored. Therefore, this diagnosis by temporary slices allows analyzing the impact and evolution of faults over time. Let us note that different time slices could be used to increase resolution in diagnosing faults such as arc faults [25], partial shadowing [18], LL-faults [65] that occur with low levels of irradiation.

Referring to time aspects, it should also be noted that multi-resolution signal decomposition is extremely efficient at detecting the exact time a signal changes as well as the type and extent of the change [140]. This provides an advantage over the Fourier transform because if the fault manifests faster than the sampling window of the Fourier analysis, like it is the case of arc faults, it is very likely that they go completely undetected.

The various contributions highlighted above make the proposed method an effective method for monitoring PV systems and likely to significantly reduce maintenance costs.

Interestingly, the method that is proposed is based on generic algorithms that could be applied to PV array faults that are not considered in this article, and also to other applications of the energy sector. This is considered in our future work. It is also envisaged to make the measurements of the electrical quantities, including the current, at a higher frequency than that used in the tests of this article in order to check whether the diagnosis is thereby improved.

References

[1] O. Onar, M. Uzunoglu, M. Alam, Modeling, control and simulation of an autonomous wind turbine/photovoltaic/fuel cell/ultra-capacitor hybrid power system, Journal of Power Sources 185 (2008) 1273–1283.

[2] P. K. Ray, A. Mohanty, B. K. Panigrahi, P. K. Rout, Modified wavelet transform based fault analysis in a solar photovoltaic system, Optik 168 (2018) 754–763.

[3] E. Romero-Cadaval, B. Francois, M. Malinowski, Q.-C. Zhong, Grid-connected photovoltaic plants: An alternative energy source, replacing conventional sources, IEEE Industrial Electronics Magazine 9 (2015) 18–32.

[4] A. Shahsavari, M. Akbari, Potential of solar energy in developing countries for reducing energy-related emissions, Renewable and Sustainable Energy Reviews 90 (2018) 275–291.

[5] M. Seyedmahmoudian, B. Horan, T. K. Soon, R. Rahmani, A. M. Than Oo, S. Mekhilef, A. Stojcevski, State of the art artificial intelligence-based mppt techniques for mitigating partial shading effects on pv systems – a review, Renewable and Sustainable Energy Reviews 64 (2016) 435–455.

[6] B. Parida, S. Iniyan, R. Goic, A review of solar photovoltaic technologies, Renewable and Sustainable Energy Reviews 15 (2011) 1625–1636.

[7] S. Upadhyay, M. Sharma, A review on configurations, control and sizing methodologies of hybrid energy systems, Renewable and Sustainable Energy Reviews 38 (2014) 47–63.

[8] S. Fadhel, A. Migan, C. Delpha, D. Diallo, I. Bahri, M. Trabelsi, M. Mimouni, Data-driven approach for isolated pv shading fault diagnosis based on experimental i-v curves analysis, 2018, pp. 927–932.

[9] C. Ferrara, D. Philipp, Why do pv modules fail?, Energy Procedia 15 (2012) 379–387.

[10] A. Dhoke, R. Sharma, T. K. Saha, Pv module degradation analysis and impact on settings of overcurrent protection devices, Solar Energy 160 (2018) 360–367.

[11] M. Köntges, G. Oreski, U. J. Magnus Herz, P. Hacke, K.-A. Weiss, Assessment of Photovoltaic Module Failures in the Field, Technical Report, International energy agency photovoltaic power systems programme, 2017.

[12] E. Jamshidpour, P. Poure, E. Gholipour, S. Saadate, Single-switch dc–dc converter with fault-tolerant capability under open- and short-circuit switch failures, IEEE Transactions on Power Electronics 30 (2015) 2703–2712.

[13] P. Jain, J.-X. Xu, S. K. Panda, J. Poon, C. Spanos, S. R. Sanders, Fault diagnosis via pv panel-integrated power electronics, in: IEEE 17th Workshop on Control and Modeling for Power Electronics (COMPEL), 2016, pp. 1–6.

[14] S. Ahmad, N. Hasan, V. S. Bharath Kurukuru, M. Ali Khan, A. Haque, Fault classification for single phase photovoltaic systems using machine learning techniques, in: 8th IEEE India International Conference on Power Electronics (IICPE), 2018, pp. 1–6.

[15] A. Haque, K. V. S. Bharath, M. A. Khan, I. Khan, Z. A. Jaffery, Fault diagnosis of photovoltaic modules, Energy Science & Engineering 7 (2019) 622–644.

[16] V. S. B. Kurukuru, A. Haque, M. A. Khan, A. K. Tripathy, Fault classification for photovoltaic modules using thermography and machine learning techniques, in: International Conference on Computer and Information Sciences (ICCIS), 2019, pp. 1–6.

[17] Z. Wang, S. McConnell, R. S. Balog, J. Johnson, Arc fault signal detection - fourier transformation vs. wavelet decomposition techniques using synthesized data, in: IEEE 40th Photovoltaic Specialist Conference (PVSC), 2014, pp. 3239–3244.

[18] B. P. Kumar, G. S. Ilango, M. J. B. Reddy, N. Chilakapati, Online fault detection and diagnosis in photovoltaic systems using wavelet packets, IEEE Journal of Photovoltaics 8 (2018) 257–265.

[19] W. Chine, A. Mellit, A. M. Pavan, S. Kalogirou, Fault detection method for grid-connected photovoltaic plants, Renewable Energy 66 (2014) 99–110.

[20] P. Sobański, P. T. Orłowska-Kowalska, Application of open-circuit igbt faults diagnostic method in dtc-svm induction motor drive, Automatika 57 (2016) 387–395.

[21] F. Meinguet, P. Sandulescu, B. Aslan, L. Lu, N.-K. Nguyen, X. Kestelyn, E. Semail, A signal-based technique for fault detection and isolation of inverter faults in multi-phase drives, in: IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), 2012, pp. 1–6.

[22] S. Nie, Y. Chen, X. Pei, H. Wang, Y. Kang, Fault diagnosis of a single-phase inverter using the magnetic field waveform near the output inductor, in: Twenty-Sixth Annual IEEE Applied Power Electronics Conference and Exposition (APEC), 2011, pp. 1648–1655.

[23] V. S. B. Kurukuru, A. Haque, M. A. Khan, Fault detection in single-phase inverters using wavelet transform-based feature extraction and classification techniques, in: Applications of Computing, Automation and Wireless Systems in Electrical Engineering (MARC), Springer Singapore, Singapore, 2019, pp. 649–661.

[24] Z. Yi, A. H. Etemadi, Line-to-line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine, IEEE Transactions on Industrial Electronics 64 (2017) 8546–8556.

[25] Z. Wang, R. S. Balog, Arc fault and flash detection in dc photovoltaic arrays using wavelets, in: IEEE 39th Photovoltaic Specialists Conference (PVSC), 2013, pp. 1619–1624.

[26] Y. Zhao, J.-F. de Palma, J. Mosesian, R. Lyons, B. Lehman, Line–line fault analysis and protection challenges in solar photovoltaic arrays, IEEE Transactions on Industrial Electronics 60 (2013) 3784–3795.

[27] R. Hariharan, M. Chakkarapani, G. S. Ilango, Challenges in the detection of line-line faults in pv arrays due to partial shading, in: International Conference on Energy Efficient Technologies for Sustainability (ICEETS), 2016, pp. 23–27.

[28] C. Strobl, P. Meckler, Arc faults in photovoltaic systems, in: Proceedings of the 56th IEEE Holm Conference on Electrical Contacts (HOLM), 2010, pp. 1–7.

[29] Z. Wang, S. McConnell, R. Balog, J. Johnson, Arc fault signal detection - fourier transformation vs. wavelet decomposition techniques using synthesized data, 2014, pp. 1–6.

[30] M. Rabla, E. Tisserand, P. Schweitzer, J. Lezama, Arc fault analysis and localisation by cross-correlation in 270 v dc, in: IEEE 59th Holm Conference on Electrical Contacts (HOLM), 2013, pp. 1–6.

[31] M. Dhimish, V. Holmes, B. Mehrdadi, M. Dales, Simultaneous fault detection algorithm for grid-connected photovoltaic plants, IET Renewable Power Generation 11 (2017).

[32] F. Harrou, Y. Sun, B. Taghezouit, A. Saidi, M.-E. Hamlati, Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches, Renewable Energy 116 (2018) 22–37.

[33] J. A. Tsanakas, L. Ha, C. Buerhop, Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges, Renewable and Sustainable Energy Reviews 62 (2016) 695–709.

[34] E. Kaplani, Detection of Degradation Effects in Field-Aged c-Si Solar Cells through IR Thermography and Digital Image Processing, International Journal of Photoenergy (2012) 396792.

[35] A. Livera, M. Theristis, G. Makrides, G. E. Georghiou, Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems, Renewable Energy 133 (2019) 126–143.

[36] Y. Zhao, L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian, R. Lyons, Decision tree-based fault detection and classification in solar photovoltaic arrays, in: Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), 2012, pp. 93–99.

[37] M. N. Akram, S. Lotfifard, Modeling and health monitoring of dc side of photovoltaic array, IEEE Transactions on Sustainable Energy 6 (2015) 1245–1253.

[38] D.-M. Tsai, G.-N. Li, W.-C. Li, W.-Y. Chiu, Defect detection in multi-crystal solar cells using clustering with uniformity measures, Advanced Engineering Informatics 29 (2015) 419–430.

[39] E. H. Sepúlveda Oviedo, L. Travé-Massuyès, A. Subias, C. Alonso, M. Pavlov, Hierarchical clustering and dynamic time warping for fault detection in photovoltaic systems, in: X Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización (XCIMM), Bogotá, Colombia, 2021, pp. 1–2.

[40] Y. Chouay, M. Ouassaid, A Multi-stage SVM Based Diagnosis Technique for Photovoltaic PV Systems, in: Advances in Robotics, Automation and Data Analytics, volume 1350, Springer International Publishing, Cham, 2021, pp. 183–193.

[41] Z. Chen, L. Wu, S. Cheng, P. Lin, Y. Wu, W. Lin, Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics, Applied Energy 204 (2017) 912–931.

[42] W. Chine, A. Mellit, V. Lughi, A. Malek, G. Sulligoi, A. Massi Pavan, A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks, Renewable Energy 90 (2016) 501–512.

[43] H. Ding, K. Ding, J. Zhang, Y. Wang, L. Gao, Y. Li, F. Chen, Z. Shao, W. Lai, Local outlier factor-based fault detection and evaluation of photovoltaic system, Solar Energy 164 (2018) 139–148.

[44] W. He, D. Yin, K. Zhang, X. Zhang, J. Zheng, Fault Detection and Diagnosis Method of Distributed Photovoltaic Array Based on Fine-Tuning Naive Bayesian Model, Energies 14 (2021) 1–17.

[45] Z. Zhang, M. Ma, H. Wang, H. Wang, W. Ma, X. Zhang, A fault diagnosis method for photovoltaic module current mismatch based on numerical analysis and statistics, Solar Energy 225 (2021) 221–236.

[46] E. Garoudja, F. Harrou, Y. Sun, K. Kara, A. Chouder, S. Silvestre, Statistical fault detection in photovoltaic systems, Solar Energy 150 (2017) 485–499.

[47] S. R. Madeti, S. N. Singh, Modeling of PV system based on experimental data for fault detection using kNN method, Solar Energy 173 (2018) 139–151.

[48] A. Hazra, S. Das, M. Basu, An efficient fault diagnosis method for PV systems following string current, Journal of Cleaner Production 154 (2017) 220–232.

[49] Q. Zhou, P. Yan, Y. Xin, Research on a knowledge modelling methodology for fault diagnosis of machine tools based on formal semantics, Advanced Engineering Informatics 32 (2017) 92–112.

[50] J. Lines, A. Bagnall, Time series classification with ensembles of elastic distance measures, Data Min Knowl Disc 29 (2015) 565–592.

[51] X. Wang, F. Yu, W. Pedrycz, L. Yu, Clustering of interval-valued time series of unequal length based on improved dynamic time warping, Expert Systems with Applications 125 (2019) 293–304.

[52] Y.-S. Jeong, M. K. Jeong, O. A. Omitaomu, Weighted dynamic time warping for time series classification, Pattern Recognition 44 (2011) 2231–2240.

[53] M. Łuczak, Hierarchical clustering of time series data with parametric derivative dynamic time warping, Expert Systems with Applications 62 (2016) 116–130.

[54] M. Sammour, Z. A. Othman, A. M. M. Rus, R. Mohamed, Modified dynamic time warping for hierarchical clustering, International Journal on Advanced Science, Engineering and Information Technology 9 (2019) 1481–1487.

[55] A.-Z. Fatama, A. Haque, M. A. Khan, A multi feature based islanding classification technique for distributed generation systems, in: International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 160–166.

[56] W. Ji, J. Zhang, Phase error evaluation technique based on fourier transform for refractive index detection limit of microfluidic differential refractometer, Optik 127 (2016) 7973–7977.

[57] N. Haje Obeid, A. Battiston, T. Boileau, B. Nahid-Mobarakeh, Early intermittent interturn fault detection and localization for a permanent magnet synchronous motor of electrical vehicles using wavelet transform, IEEE Transactions on Transportation Electrification 3 (2017) 694–702.

[58] D. Bayram, S. Şeker, Redundancy-based predictive fault detection on electric motors by stationary wavelet transform, IEEE Transactions on Industry Applications 53 (2017) 2997–3004.

[59] F. B. Costa, B. A. Souza, N. S. D. Brito, J. A. C. B. Silva, W. C. Santos, Real-time detection of transients induced by high-impedance faults based on the boundary wavelet transform, IEEE Transactions on Industry Applications 51 (2015) 5312–5323.

[60] N. Sangeetha, X. Anita, Entropy based texture watermarking using discrete wavelet transform, Optik 160 (2018) 380–388.

[61] Z. Yi, A. H. Etemadi, Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems, IEEE Transactions on Smart Grid 8 (2017) 1274–1283.

[62] M. Ahmadipour, H. Hizam, M. L. Othman, M. A. M. Radzi, A. S. Murthy, Islanding detection technique using slantlet transform and ridgelet probabilistic neural network in grid-connected photovoltaic system, Applied Energy 231 (2018) 645–659.

[63] M. Ahmadipour, H. Hizam, M. Lutfi Othman, M. Amran Mohd Radzi, An anti-islanding protection technique using a wavelet packet transform and a probabilistic neural network, Energies 11 (2018).

[64] V. S. B. Kurukuru, F. Blaabjerg, M. A. Khan, A. Haque, A novel fault classification approach for photovoltaic systems, Energies 13 (2020).

[65] K. Dadhich, V. S. B. Kurukuru, M. A. Khan, A. Haque, Fault identification algorithm for grid connected photovoltaic systems using machine

learning techniques, in: International Conference on Power Electronics, Control and Automation (ICPECA), 2019, pp. 1–6.

[66] A. Sharma, M. Amarnath, P. Kankar, Feature extraction and fault severity classification in ball bearings, Journal of Vibration and Control 22 (2016) 176–192.

[67] K. Arunkumar, D. T. Manjunath, A brief review/survey of vibration signal analysis in time domain, SSRG International Journal of Electronics and Communication Engineering 3 (2019) 12–55.

[68] K. H. Hui, C. S. Ooi, M. H. Lim, M. S. Leong, S. M. Al-Obaidi, An improved wrapper-based feature selection method for machinery fault diagnosis, PLOS ONE 12 (2017) 1–10.

[69] A. Nanopoulos, R. Alcock, Y. Manolopoulos, Feature-Based Classification of Time-Series Data, Nova Science Publishers, Inc., USA, 2001, p. 49–61.

[70] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, A. Linney, Classification of audio signals using statistical features on time and wavelet transform domains, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 6, 1998, pp. 3621–3624 vol.6.

[71] F. Chen, B. Tang, T. Song, L. Li, Multi-fault diagnosis study on roller bearing based on multi-kernel support vector machine with chaotic particle swarm optimization, Measurement 47 (2014) 576–590.

[72] I. Jolliffe, Principal Component Analysis, Springer, New York, 2002.

[73] A. Arnaout, B. Alsallakh, R. Fruhwirth, G. Thonhauser, B. Esmael, M. Prohaska, Diagnosing drilling problems using visual analytics of sensors measurements, in: IEEE International Instrumentation and Measurement Technology Conference Proceedings (I2MTC), 2012, pp. 1750–1753.

[74] J. Johansson, C. Forsell, Evaluation of parallel coordinates: Overview, categorization and guidelines for future research, IEEE Transactions on Visualization and Computer Graphics 22 (2016) 579–588.

[75] C. A. Steed, G. Shipman, P. Thornton, D. Ricciuto, D. Erickson, M. Branstetter, Practical application of parallel coordinates for climate model analysis, Procedia Computer Science 9 (2012) 877–886.

[76] B. Esmael, A. Arnaout, R. Fruhwirth, G. Thonhauser, A statistical feature-based approach for operations recognition in drilling time series, International Journal of Computer Information Systems and Industrial Management Applications 4 (2012) 100–108.

[77] X. Wang, Y. Zheng, Z. Zhao, J. Wang, Bearing fault diagnosis based on statistical locally linear embedding, Sensors 15 (2015) 16225–16247.

[78] M. E. Tipping, C. M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society 61 (1999) 611–622.

[79] Z. Xia, S. Xia, L. Wan, S. Cai, Spectral regression based fault feature extraction for bearing accelerometer sensor signals, Sensors 12 (2012) 13694–13719.

[80] B. Basnet, H. Chun, J. Bang, An intelligent fault detection model for fault detection in photovoltaic systems, Journal of Sensors (2020) 1–11.

[81] X. Zhao, J. Guo, F. Nie, L. Chen, Z. Li, H. Zhang, Joint principal component and discriminant analysis for dimensionality reduction, IEEE Transactions on Neural Networks and Learning Systems 31 (2020) 433–444.

[82] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–23.

[83] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[84] O. Alter, P. Brown, B. D., Singular value decomposition for genome-wide expression data processing and modeling, Proc Natl Acad Sci U S A 97 (2000).

[85] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.

[86] B. Schölkopf, A. J. Smola, K.-R. Müller, Kernel Principal Component Analysis, MIT Press, Cambridge, MA, USA, 1999, p. 327–352.

[87] D. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking, Int J Comput Vis 77 (2008) 125–141.

[88] M. Usman, S. Ahmed, J. Ferzund, A. Mehmood, A. Rehman, Using pca and factor analysis for dimensionality reduction of bio-informatics data, International Journal of Advanced Computer Science and Applications 8 (2017).

[89] D. L. Donoho, C. Grimes, Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences 100 (2003) 5591–5596.

[90] A. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: NIPS, 2001.

[91] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, SIAM Journal on Scientific Computing 26 (2004) 313–338.

[92] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (2000) 411–430.

[93] C. Huang, J. Du, B. Nie, R. Yu, W. Xiong, Q. Zeng, Feature selection method based on partial least squares and analysis of traditional chinese medicine data, Computational and Mathematical Methods in Medicine (2019) 1–12.

[94] R. Jenatton, G. Obozinski, F. Bach, Structured sparse principal component analysis, in: Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), Chia Laguna Resort, Sardinia, Italy, 2010, pp. 366–373.

[95] Y. Liu, W. Rayens, Pls and dimension reduction for classification, Computational Statistics 22 (2007) 189–208.

[96] D. V. Nguyen, D. M. Rocke, Tumor classification by partial least squares using microarray gene expression data , Bioinformatics 18 (2002) 39–50.

[97] A. Boulesteix, Pls dimension reduction for classification with microarray data, Statistical Applications in Genetics and Molecular Biology 3 (2004) 1 – 30.

[98] F. Nielsen, In Introduction to hpc with mpi for data science, Springer International Publishing, 2016, p. 195–211.

[99] T. M. Cesar, S. P. Pimentel, E. G. Marra, B. P. Alvarenga, Wavelet transform analysis for grid-connected photovoltaic systems, in: 6th International Conference on Clean Electrical Power (ICCEP), 2017, pp. 1–6.

[100] H. Li, Time works well: Dynamic time warping based on time weighting for time series data mining, Information Sciences 547 (2021) 592–608.

[101] B. Jun, Fault detection using dynamic time warping (dtw) algorithm and discriminant analysis for swine wastewater treatment, Journal of Hazardous Materials 185 (2011) 262–268.

[102] Y. Tanaka, M. Takahashi, Dynamic time warping-based cluster analysis and support vector machine-based prediction of solar irradiance at multi-points in a wide area, International Symposium on Stochastic Systems Theory and its Applications (ISCIE) (2016) 210–215.

[103] H. Badr, B. Zaitchik, A. Dezfuli, A tool for hierarchical climate regionalization, Earth Sci Inform 8 (2016) 949–958.

[104] S. Aminikhanghahi, D. Cook, A survey of methods for time series change point detection, Knowl Inf Syst 51 (2017) 339–367.

[105] S. Rani, G. Sikka, Article: Recent techniques of clustering of time series data: A survey, International Journal of Computer Applications 52 (2012) 1–9.

[106] M. Saleh, Z. Othman, M. S. Saleh, Characteristics of agent-based hierarchical diff-edf schedulability over heterogeneous real-time packet networks, European journal of scientific research 27 (2009) 431–453.

[107] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Transactions on Information Theory 36 (1990) 961–1005.

[108] K. Kashyap, U. Shenoy, Classification of power system faults using wavelet transforms and probabilistic neural networks, in: International Symposium on Circuits and Systems (ISCAS), 2003, pp. 1–4.

[109] A. Etemadi, M. Sanaye-Pasand, High-impedance fault detection using multi-resolution signal decomposition and adaptive neural fuzzy inference system, IET Generation, Transmission and Distribution 2 (2008) 110 – 118.

[110] S. Mallat, A Wavelet Tour of Signal Processing 3rd ed., Academic Press, Cambridge, MA, USA, 2008, p. 1 – 745.

[111] A. Belaout, F. Krim, A. Mellit, B. Talbi, A. Arabi, Multiclass adaptive neuro-fuzzy classifier and feature selection techniques for photovoltaic array fault detection and classification, Renewable Energy 127 (2018) 548–558.

[112] A. G. Shaik, R. R. V. Pulipaka, A new wavelet based fault detection, classification and location in transmission lines, International Journal of Electrical Power & Energy Systems 64 (2015) 35–40.

[113] P. K. Ray, B. K. Panigrahi, P. K. Rout, A. Mohanty, H. Dubey, Detection of faults in power system using wavelet transform and independent component analysis, in: First International Conference on Advancement of Computer Communication & Electrical Technology, 2016, pp. 1–5.

[114] W. Zhao, Y. Song, Y. Min, Wavelet analysis based scheme for fault detection and classification in underground power cable systems, Electric Power Systems Research 53 (2000) 23–30.

[115] C. Pang, M. Kezunovic, Fast distance relay scheme for detecting symmetrical fault during power swing, IEEE Transactions on Power Delivery 25 (2010) 2205–2212.

[116] K. L. V. Iyer, X. Lu, Y. Usama, V. Ramakrishnan, N. C. Kar, A twofold daubechies-wavelet-based module for fault detection and voltage regulation in seigs for distributed wind power generation, IEEE Transactions on Industrial Electronics 60 (2013) 1638–1651.

[117] B. N. Singh, A. K. Tiwari, Optimal selection of wavelet basis function applied to ecg signal denoising, Digital Signal Processing 16 (2006) 275–287.

[118] A. Jensen, A. Cour-Harbo, Ripples in Mathematics:the Discrete Wavelet Transform, Springer ed, 2001, pp. 1–246.

[119] S. Vergura, G. Acciani, V. Amoruso, G. E. Patrono, F. Vacca, Descriptive and inferential statistics for supervising and monitoring the operation of pv plants, IEEE Transactions on Industrial Electronics 56 (2009) 4456–4464.

[120] N. Ismail, F. Nordin, A. Alkahtani, S. ZAM., Detection of the source of the incipient faults produced by single phase inverter using feed-forward back-propagation neural network., Indian Journal of Science and Technology 9 (2016) 1–9.

[121] F. Wang, Y. Yu, Z. Zhang, J. Li, Z. Zhen, K. Li, Wavelet decomposition and convolutional lstm networks based improved deep learning model for solar irradiance forecasting, Applied Sciences 8 (2018).

[122] D. Goyal, A. Choudhary, B. Pabla, S. Dhami, Support vector machines based non-contact fault diagnosis system for bearings, J Intell Manuf 31 (2020) 1275–1289.

[123] R. B. Zajonc, A note on group judgements and group size, Human Relations 15 (1962) 177–180.

[124] V. Buško, Psychological Testing Theory, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 1138–1139.

[125] R. M. Hogarth, Methods for Aggregating Opinions, Springer Netherlands, Dordrecht, 1977, pp. 231–255.

[126] A.-L. Boulesteix, K. Strimmer, Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, Briefings in Bioinformatics 8 (2006) 32–44.

[127] H. Wold, Estimation of principal components and related models by iterative least squares, 1966.

[128] H. Wold, Soft modelling: The basic design and some extensions, 1982.

[129] J. Dai, L. H. Lieu, D. M. Rocke, Dimension reduction for classification with gene expression microarray data, Statistical Applications in Genetics and Molecular Biology 5 (2006).

[130] M. Z. Man, G. Dyson, K. Johnson, B. Liao, Evaluating methods for classifying expression data, Journal of Biopharmaceutical Statistics 14 (2004) 1065–1084.

[131] X. Huang, W. Pan, S. Grindle, X. Han, Y. Chen, S. Park, L. Miller, J. Hall, A comparative study of discriminating human heart failure etiology using gene expression profiles, BMC Bioinformatics 6 (2005) 1065–1084.

[132] H. Pham, A new criterion for model selection, Mathematics 7 (2019).

[133] E. Ostertagová, Modelling using polynomial regression, Procedia Engineering 48 (2012) 500–506.

[134] A. Kulkarni, D. Chong, F. A. Batarseh, 5 - foundations of data imbalance and solutions for a data democracy, in: Data Democracy, Academic Press, 2020, pp. 83–106.

[135] L. Tang, S. Peng, Y. Bi, P. Shan, X. Hu, A new method combining lda and pls for dimension reduction, PLOS ONE 9 (2014) 1–10.

[136] C. Ferri, J. Hernández-Orallo, R. Modroiu, An experimental comparison of performance measures for classification, Pattern Recognition Letters 30 (2009) 27–38.

[137] D. S. Pillai, F. Blaabjerg, N. Rajasekar, A Comparative Evaluation of Advanced Fault Detection Approaches for PV Systems, IEEE Journal of Photovoltaics 9 (2019) 513–527.

[138] S. Firth, K. Lomas, S. J. Rees, A simple model of pv system performance and its use in fault detection, Sol Energy 84 (2010) 624–635.

[139] E. Garoudja, A. Chouder, K. Kara, S. Silvestre, An enhanced machine learning based approach for failures detection and diagnosis of PV systems, Energy Conversion and Management 151 (2017) 496–513.

[140] M. Misiti, Y. Misiti, G. Oppenheim, J. Poggi, Wavelet toolbox - user's guide, 2013.