Fault Diagnosis in a Grid-Connected Photovoltaic Systems Based on Hierarchical Clustering

Amal Hichri, Mansour Hajji, Majdi Mansouri, Hazem Nounou, Abdelmalek Kouadri and Kais Bouzrara

September 2, 2021

Fault diagnosis in a grid-connected photovoltaic systems based on Hierarchical Clustering*

Amal Hichri, Mansour Hajji
Higher Institute of Applied Sciences and Technology of Kasserine,
Kasserine, Tunisia,
(amelhichri@outlook.fr, hajji_mansour@yahoo.fr)

Majdi Mansouri, Hazem Nounou
Texas A&M University at Qatar
Doha, Qatar
[majdi.mansouri,hazem.nounou]@qatar.tamu.edu

Abdelmalek Kouadri
University M'Hamed Bougara of Boumerdes
Boumerdes, Algieria
(ab_kouadri@hotmail.com)

Kais Bouzrara
National Engineering School of Monastir,
Monastir, Tunisia
(bouzrara.kais@gmail.com)

***Abstract*—This paper proposes an effective fault detection and diagnosis (FDD) of Grid-Connected Photovoltaic (GCPV) systems. The developed approach combines the advantages of both Principal Component Analysis (PCA) model and Hierarchical Clustering (HC) scheme. The PCA model is applied to extract and select the most informative features from GCPV system data. While, the HC metric is used to classify the GCPV faults and distinguish between the operating healthy and faulty modes. The proposed FDD approach, the so-called PCA-based HC is experimentally tested and validated using GCPV system data. Different case studies are investigated in this paper in order to illustrate the efficiency and the robustness of the proposed framework. A comparison with well-known techniques is also presented. The obtained results confirm the high accuracy of the developed technique.**

Keywords—Grid-connected PV systems; principal component analysis (PCA); fault diagnosis; fault classification; Hierarchical Clustering (HC); feature extraction and selection (FES).

I. Introduction

As with any energy system, photovoltaic (PV) systems are subject to failures during operation due to aging effects and external/environmental conditions [1]. The PV systems usually control in harsh outdoor conditions which make them suffering from different faults in the different PV components (PV modules, cabling, converters, inverters …). To solve these problems, it is essential to implement an efficient and comprehensive fault detection and diagnosis (FDD) technique [2], [3].

Model based, image-based and data-driven are the principal approaches used in PV fault diagnosis [4]. Image-based approaches need specific conditions to be performed such appropriate and expensive equipment [5]. Model-based approaches use an analytical model of the PV [6]. The model-based fault diagnosis is based on computing the error between the measured and estimated variables. The main advantage of these techniques is that they have a low hardware requirement and are related to a varied range of PV systems. However, these approaches depend on the adequacy of the mathematical model to describe well the behaviours of the PV system for which additional sensing devices are needed [7]. In fact, the majority of the model-based approaches proposed for PV system fault diagnosis are applied to small scale PV systems [6].

Recently, various studies have been concentrating on the FDD in PV systems using computational intelligence and machine learning techniques. These tools are data-driven approaches and are based on historical data collected during operation of the PV system [8], [9].

Particularly, Principal Component Analysis (PCA) is a well-known multivariate statistical method [10], [11], [12], [13]. PCA is a dimensionality reduction technique able to capture the most dominant variances in the data and it describes the principal component subspace and the residual subspace by means of a linear transformation [14].

In this paper, therefore, we propose an FDD approach that merges the benefits of PCA model and Hierarchical Clustering (HC) scheme. The developed FDD approach is addressed so that the PCA is applied for feature and extraction purposes and HC metric is used for fault classification. Indeed, the HC metric is one of the most applied classifiers in FDD of industrial processes. But its use suffers from certain limitations and disadvantages when using a single variable at each node without considering the correlations between the variables. In addition, to perform FDD, the classical HC only uses raw information from process data by the direct use of measured variables at nodes. The direct raw signals could yield to poor diagnosis results due to noise and redundancies on data.

The developed PCA-based HC method aims first to reduce the amount of the training data, extract and select the most relevant features using PCA based dimensionality reduction scheme. Then, the selected features are fed to the HC classifier for fault diagnosis purposes. The validation is done using a Grid Connected Photovoltaic (GCPV) system data.

The remaining sections of this paper are organized as follows: section 2 is devoted to feature extraction and selection. In section 3, Fault classification approach is presented. Section 4 shows the application results and discussions. The last section is devoted to some conclusions and findings.

II. Feature Extraction and Selection

A. PCA-based feature extraction

We consider the data matrix $X \in \Re^{n \times m}$, collected from a process operates under normal conditions with $n$samples of $m$variables, these data can be stored a zero mean and unit variance matrix $X_s = [X_{s1}^T X_{s2}^T \cdots X_{sm}^T]$. The linear transformation that projects the data onto the subspaces is gives by the following equation

$$T = X_s P \tag{1}$$

Where T is the score matrix. P is an m by m orthonormal matrix consists of the covariance matrix of X_s eigenvectors. The covariance matrix of X_s using the Singular Value Decomposition (SVD) on Φ yields to

$$\Phi = P \Lambda P^T \tag{2}$$

Where $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$ is a diagonal matrix contains the eigenvalues sorted in a decreasing order.

Equation (1) permits to rewrite X_s as

$$X_s = T P^T \tag{3}$$

Now the number of principal components is selected noted l, X_s becomes as

$$X_s = \hat{X}_s + E \tag{4}$$

Such that

$$\hat{X}_s = \hat{T} \hat{P}^T \tag{5}$$

$$E = \tilde{T} \tilde{P}^T \tag{6}$$

Where

$$\hat{P} \in R^{m \times l}, \tilde{P} \in R^{m \times (m^{-l)}}, \hat{T} \in R^{n \times l} and \widetilde{T} \in R^{n \times (m-l)}$$

B. PCA-based feature selection

In the literature, numerous methods exist for selecting the number of principal components l. These methods mainly include the cumulative percentage of variances, cross validation, scree plot and parallel analysis [14], [15]. In this work, the cumulative percentage of variance (CPV) criterion [16] has been used. The CPV is a measure of the percent variance defined by the first principal components l.

$$CPV(l) = \frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \times 100 \tag{7}$$

C. Statistical characteristics extraction

To obtain a good performance of ML classification based approaches, it is significance to extract statistical behaviour via PCA-based model by exhaustively enumerating some possible values. This type of features is simple to compute and occasionally efficient and principled manner. The features considered as appropriate for faults diagnosis in GCPV include the following T^2 statistic measures the variations in principal subspace [17], [18]. It is given by

$$T^2 = x^T \hat{P} \hat{\Lambda}^{-1} \hat{P}^T x \tag{8}$$

Its control limit noted by T_α decides whether the process is healthy and faulty, it is given by

$$T_\alpha = \frac{(n^2 - 1) a}{n(n-a)} F_\alpha(a, n-a) \tag{9}$$

Where n is the number of observations and a is the number PCs. $F_\alpha(a, n-a)$ is an F -distribution of a, $n-a$ degree of freedom evaluated at given confidence level (1-α).

The SPE statistic, also known as Q statistic measures the projection of the sampled data vector onto the residual subspace [19], [20]. It is defined as follows

$$Q = \left\| (I - \hat{P}\hat{P}^T) x \right\|^2 = x^T (I - \hat{P}\hat{P}^T)^2 \tag{10}$$

The control limit of SPE noted by Q_α. The process is under healthy state if inequality given by

$$SPE \leq Q_\alpha \tag{11}$$

And

$$Q_\alpha = \theta_1 \left[c_\alpha \frac{h_0 \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 \theta_0 (\theta_0 - 1)}{\theta_1^2} \right]^{\frac{1}{\theta_0}} \tag{12}$$

Where

$$\theta_i = \sum_{j=a+1}^{m} \lambda_j^i, \; i = 1, 2, 3 \tag{13}$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \tag{14}$$

C_α is the normal deviate corresponding to $(1-\alpha)$ percentile.

The combined index is a combination between $T^2$and SPE [21], [22].

$$\varphi = \frac{SPE(x)}{Q_\alpha} + \frac{T^2}{T_\alpha} \tag{15}$$

III. Fault Classification using Hierarchical Clustering

A. Hierarchical Clustering Algorithm

Hierarchical clustering is viewed as the most significant unsupervised learning algorithm, which is utilized to collect the unlabeled database in clusters. These clusters are shown dendrograms, which are in fact tree representation of points relying on similarity or dissimilarity metrics.

This method contains two approaches: Agglomerative and Divisive. The Agglomerative takes bottom up approach: in which the algorithm begins with taking all data points as single clusters and merging them until one cluster is obtained. While the Divisive takes top down approach: at first, all the data points are presented in one cluster then the algorithm splits it until each data points illustrated as single cluster [23], [24].

This paper deals with the Agglomerative algorithm, which is divided in 3 steps [25]:

- Average-linkage cluster: the distance from one cluster to the other must be the average distance.
- Complete-linkage cluster: also called maximum method, the distance from one cluster to the other must be greatest.
- Single-linkage cluster: also known as minimum method, the distance from one cluster to the other must be shortest.

The fault diagnosis approach is addressed such that the Single-linkage cluster is used for fault classification purposes [26], the following steps are illustrated bellow:

1. Allocate a cluster to each point, such that N clusters for N points.
2. Seek and merge the pair of clusters which are closest to each other.
3. Measure the distances among the new and each of the ancient clusters.
   a) Commence with the disjoint clustering having level, l(0) = 0 and sequence number n=0.
   b) In the present clustering, we have to find the lowest dissimilar pair of clusters say pair (a), (b), according to d[(a),(b)] = min d[(u),(v)] where the minimum is over all pairs of clusters in the current clustering.
   c) Increase the sequence number: n = n+1 and merge clusters (a) and (b) into a single cluster to form the next clustering n. Set the level of this clustering to l(n) = d[(a),(b)].
   d) The further step is to upgrade the proximity matrix, M, by removing the rows and columns corresponding to clusters (a), (b) and inserting a row and column correspond to the recently formed cluster. The proximity among the new cluster, indicated (a,b) and old cluster (k) is identified in this way: d[(k), (a,b)] = min d[(k),(a)], d[(k),(b)].
4. Stop the process when all the data points are in one cluster else, go to step 2.

B. Hierarchical Clustering based on Euclidean distance

This paper deals with the most used metric which is the Euclidean distance. This later calculates the root of square difference between co-ordinates of data points [27].

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2} \tag{16}$$

C. Proposed methodology

The proposed methodology involves three major stages including feature extraction, feature selection and classification. Once the measurements are available representing healthy and different possible faulty scenarios in the process, a PCA model is built under normal operating conditions.

These data are projected onto a subspace of positive right directions by keeping the most captured features information. The structure of the obtained PCA model is represented by the directions of the subspace projector where its dimension is less than that of the original data. The PCA used to extract the important features from the GCPV system then Hierarchical Clustering is applied to these features to implement the fault classification based on Euclidean distance. The main steps in the proposed PCA-based Hierarchical Clustering are summarized in Figure 1 as well as in Algorithm 1.

The steps of the developed approach are summarized in the block diagram in Figure 2. At first, we calculate the covariance matrix for each data then the average of covariance matrix aims to reduce the computing time. Therefore, we determine the multivariate statistical charts for each data.

Algorithm 1 PCA-based Hierarchical Clustering

```
Input: $N \times m$ data matrix $X$
Training phase
1. Standardize the training data set,
2. Determine the covariance matrix for each data set,
3. Determine the average of covariance matrix,
4. Determine the SVD decomposition,
5. Determine the multivariate statistical charts for each data,
6. Compute the minimum distance between each data cluster using the Euclidean distance,
7. Determine the monitoring statistics,
Testing phase
1. Standardize the testing data set using the mean and the variance computed from the healthy training phase,
2. Determine the multivariate statistical charts ($Q$, $T^2$ and the combined $\varphi$ statistics),
3. Compute the minimum distance between each data cluster using the Euclidean distance,
4. Determine the monitoring statistics of the testing data.
```

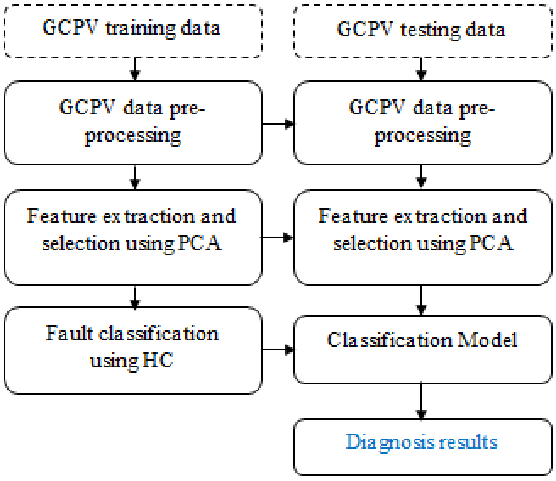

Fig. 1. Illustration of PCA based Hierarchical Clustering for GCPV fault diagnosis.

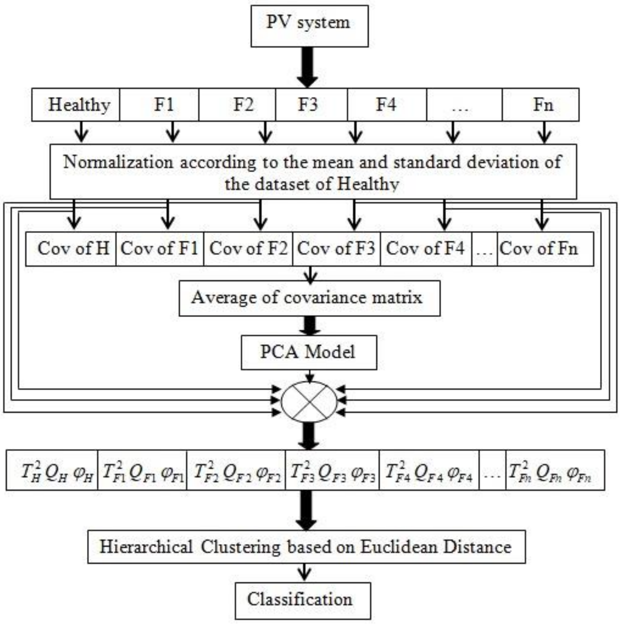
PV system
Healthy
F1
F2
F3
F4
...
Fn
Normalization according to the mean and standard deviation of the dataset of Healthy
Cov of H
Cov of F1
Cov of F2
Cov of F3
Cov of F4
...
Cov of Fn
Average of covariance matrix
PCA Model
$T^2_H\ Q_H\ \varphi_H$
$T^2_{F1}\ Q_{F1}\ \varphi_{F1}$
$T^2_{F2}\ Q_{F2}\ \varphi_{F2}$
$T^2_{F3}\ Q_{F3}\ \varphi_{F3}$
$T^2_{F4}\ Q_{F4}\ \varphi_{F4}$
...
$T^2_{Fn}\ Q_{Fn}\ \varphi_{Fn}$
Hierarchical Clustering based on Euclidean Distance
Classification

Fig. 2. Detail illustration of PCA-based Hierarchical Clustering for fault diagnosis.

IV. Results and Discussion

A. Grid connected PV implementation and data collection

In this paper, data sets have been collected from an emulator of a GCPV system. PV array emulator and grid emulator are used instead of real PV array and grid, in order to able to inject different faults in these two main parts of the GCPV system. Also, they give the flexibility to manage the desired system parameters or the outdoor conditions by setting the humidity, temperature and irradiation. A block diagram of a basic structure of a grid connected PV system is represented in Figure 3.

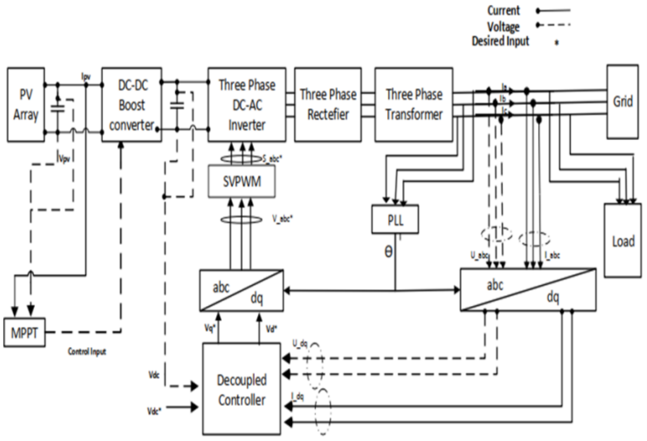

Fig. 3. Block diagram of a basic grid connected PV system.

The implemented system is a programmable DC Power Supply (Solar Array Emulation) Chroma 62150H-1000S, which has been used to emulate a real world PV panel due to its high performance efficiency and its capacity to emulate real PV panel faults. For grid emulation, the programmable AC source Chroma 61511 is used and set to the three-phase mode to match a real grid system network.

A DC-DC converter was implemented using a chopper. Its main task is to maintain the output current of the PV panel emulator to a specific level as much as possible. This DC-DC converter has one operation depending on the operating mode: In MPPT mode, it boosts and amplifies the electrical power to reach the maximum power point of the system. After the DC-DC converter a DC-AC Inverter is implemented using six IGBTs to provide a three-phase signal that will be transmitted to the grid emulator system after rectification and transformation.

The control of this system has been implemented using the DSpace 1104 environment with Matlab. There are two main controllers that have been used and implemented. The first controller is the MPPT controller, used to trigger the DC-DC converter by sending control signals based on the PV output current and voltage. The algorithm used for the MPPT is the PSO algorithm. The second control is used to maintain the synchronization of the three phase DC-AC inverter output current and voltage, with the grid parameters in terms of phase, frequency and magnitude. The phase synchronization is performed using Phase Locked Loop (PLL) algorithm. The frequency and magnitude synchronization are performed using the Voltage Oriented Control (VOC) algorithm combined with Space Vector Pulse Width Modulation (SVPWM). VOC and SVPMW are also used to control the active and reactive power using the DC-AC inverter.

The data are collected using several sensors at several locations of measurement interest, nine variables measurements are considered. The recording length of the data varies from 5 seconds to 15 seconds, depending on the nature of the fault and the damage that can cause when proceeding with the injection of some faults in the system. These measured variables are described in detailed in the table I.

TABLE I. DESCRIPTION OF THE MEASURED GCPV SYSYTEM VARIABLES

Measures	Symbol	Index in data matrix	Variable description
Grid three phase current	I_a I_b I_c	Variable 1 Variable 2 Variable 3	The output three phase current of the transformer after the DC-AC inverter
PV current	I_{pv}	Variable 4	The output current of the PV panel emulator
Grid three phase voltage	V_a V_b V_c	Variable 5 Variable 6 Variable 7	The output three phase voltage of the transformer after the DC-AC inverter
Output voltage	V_{out}	Variable 8	The output voltage of the DC-DC converter
PV voltage	V_{pv}	Variable 9	The output voltage of the PV panel emulator

In this work, the considered faults are experimental faults carried out on the GCPV system simulator. The experiment was conducted by making changes and injecting faults in different components and at different locations to ensure a global analysis and complete study Shown in table II. Six sets (one healthy mode and five faulty modes) of experimental tests have been conducted to assess the fault diagnosis approach; they have been realized under different operating conditions.

TABLE II. DESCRIPTION AND CHARACTERISTIC OF THE DIFFERENT LABLED FAULTS INJECTED

Fault label	Fault Side	Fault type	Fault description
Fault 1	AC Side	Three phase inverter fault	Damage of one IGBT at a time among the total of 6 IGBTs inside the three-phase inverter
Fault 3		Grid external connection fault	Critical external fault at the grid output level. This can be caused by loss or poor grid connection, sudden grid disconnection. The system will switch to a load for protection reasons
Fault 2	DC Side	PV sensor fault	Damage, malfunction or poor connection of the current sensor at the PV output
Fault 4		PV array level fault	Permanent partial shading of 10% to 20%
Fault 5			Critical external fault due to loss of connection / sudden disconnection / open circuit

In order to carry out the different experiments for fault classification purposes, experimental data variables are collected. The GCPV healthy stat us is assigned to class C_0 and the other five operating modes are assigned to classes C_i $(i = 1,\ldots,5)$, respectively, as reported in table III.

To get a good performance of diagnosis-based approaches, it is important to extract the best statistical characteristics from the used data set. Therefore, the objective of the proposed technique is to only keep the most important characteristics to save time in the fault classification procedure. In the current study, Hotelling's T^2 statistic, squared prediction error SPE (Q) statistic, combined index φ are used to select the final efficient features. Examples of these features are illustrated in Figures 4, 5 and 6.

TABLE III. CONSTRUCTIONL OF DATABASE FAULT DIAGNOSIS SYSTEM

Class	State	Training Data	Testing Data
C_0	Healthy	3000	2999
C_1	F1	3000	2999
C_2	F2	3000	2999
C3	F3	3000	2999
C4	F4	3000	2999
C5	F5	3000	2999

B. *Fault classification results*

Data set of GCPV system under healthy operating conditions is normalized to zero mean and unit variance and then used to build a PCA model. A key issue to identify a PCA model is to select the adequate number principal components. The number of retained principal components l has a significant impact on each step of the process modelling and monitoring scheme. The criterion used to select this number is the cumulative percentage of variances (CPV) with 90% as an explained variance threshold. The retained number of PCs using the PCA is equal to 3.

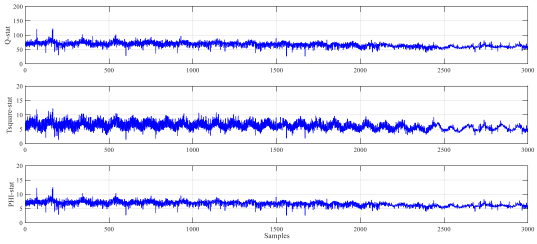

Fig. 4. Q, T^2 and φ statistics for the healthy operating mode.

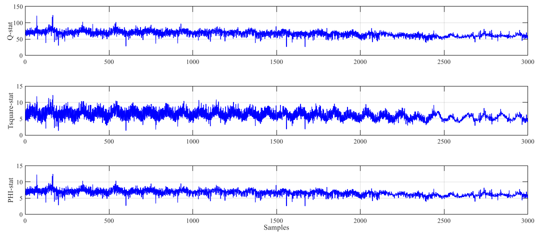

Fig. 5. Q, T^2 and φ statistics under the faulty operating mode 2

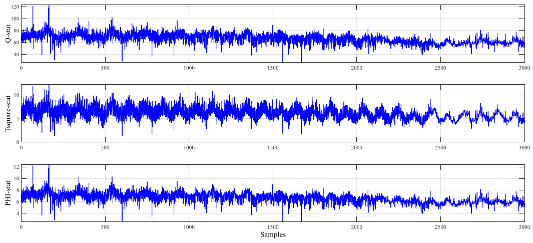

Fig. 6. Q, T^2 and φ statistics under the faulty operating mode 4.

After extracting and selecting the most informative features from the data set, Hierarchical Clustering is applied to these features for fault classification purposes; the main aim is to measure the distance between the features of the data obtained under normal and faulty states.

The diagram, illustrated by Figure 7, is shown the corresponding dendrograms to the Q , T^2 and the combined index φ. Begin with the left side of the figure, corresponding to the Q, in which the class C_0 (1) and C_3 (4) combine together and form a cluster, Then a dendrogram is created, the hight is decided according to the Euclidean distance among the classes. In the following step, C_4 (5) and C_5 (6) form a cluster by creating the corresponding dendrogram. Moreover new dendrograms are created that combine the similar classes C_0 (1), C_3 (4) and C_1 (2) in one dendrogram. Then two clusters get merged into one. In the end, the final dendrogram is created that combines all the classes together.

In the middle of the same figure, corresponding to the T^2. First of all, C_1 and C_5 form a cluster then a dendrogram is created, followed by C_3 which is merged into the same cluster, similarly for C_0. Then C_2 and C_4 combine into one cluster, forming a new dendrogram. Finally, all classes get merged into one.

While in the right side of the same figure, the classes C_0-C_3-C_1 and C_4-C_5 are all under one dendrogram. We finish when cluster is left and finally bring everything together.

Figures 8 and 9 show the results of the application of PCA based-Hierarchical clustering for testing data. We can show from the results that technique is able to distinguish between classes. In such a way that this step deals with adding a new cluster 7 on the right which is consists of selecting the correct class, which represents the correctly classified and high accuracy observations for healthy condition (C_0) and faulty conditions (C_1 to C_5) for Q, T^2 and φ.

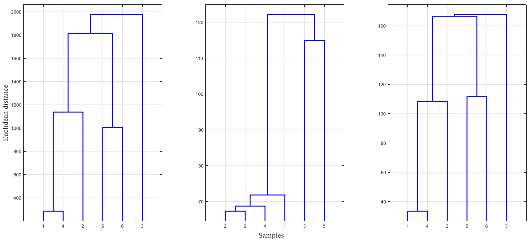

Fig. 7. Dendrogram for healthy mode Training (Q, T^2 and φ).

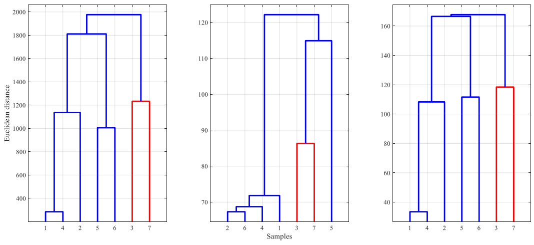

Fig. 8. Dendrogram for class 2 (Q, T^2 and φ).

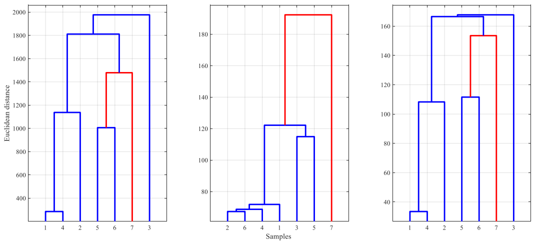

Fig. 9. Dendrogram for class 4 (Q, T^2 and φ).

In this study, Hierarchical Clustering is tested, its performance in term of accuracy via Q, T^2 and φ statistical features. Table IV and VI represent the observations correctly classified for the healthy condition (C_0) and faulty conditions (C_1 to C_5).

Clearly, the PCA based Hierarchical Clustering achieves the best overall performance with an accuracy with 100% for Q and φ. In addition, the developed approach have a high performance in term of accuracy, which is mainly due to the similarity between all extracted features that cannot be distinguished.

Therefore, the proposed technique is considered as a good alternative for faults classification due to its high accuracy and reliability for Q and φ. However, Table V represents the incorrectly classified observations for fault 1 and fault 4 (C_1 and C_4) with an accuracy 0% for T^2.

TABLE IV. MATCHING MATRIX OF THE PROPOSED METHOD FOR Q

Accuracy						
C_0	100%	0%	0%	0%	0%	0%
C_1	0%	100%	0%	0%	0%	0%
C_2	0%	0%	100%	0%	0%	0%
C_3	0%	0%	0%	100%	0%	0%
C_4	0%	0%	0%	0%	100%	0%
C_5	0%	0%	0%	0%	0%	100%

TABLE V. MATCHING MATRIX OF THE PROPOSED METHOD FOR T^2

Accuracy						
C_0	100%	0%	0%	0%	0%	0%
C_1	0%	0%	0%	0%	0%	0%
C_2	0%	0%	100%	0%	0%	0%
C_3	0%	0%	0%	100%	0%	0%
C_4	0%	0%	0%	0%	0%	0%
C_5	0%	0%	0%	0%	0%	100%

TABLE VI. MATCHING MATRIX OF THE PROPOSED METHOD FOR φ

Accuracy						
C_0	100%	0%	0%	0%	0%	0%
C_1	0%	100%	0%	0%	0%	0%
C_2	0%	0%	100%	0%	0%	0%
C_3	0%	0%	0%	100%	0%	0%
C_4	0%	0%	0%	0%	100%	0%
C_5	0%	0%	0%	0%	0%	100%

V. CONCLUSION

In this paper, the problem of faults detection and diagnosis for Grid-Connected PV (GCPV) system was considered. The developed technique was based on the Principal Component Analysis (PCA) and Hierarchical Clustering (HC) classifier. It was addressed so that the PCA technique was applied for features extraction and selection purposes and the HC was used for faults classification. The proposed approach was developed to diagnose the GCPV systems under normal and faulty conditions. Different scenarios were investigated in order to show the robustness and the efficiency of the developed approach. The technique was tested and examined using simulated GCPV data representing different operating conditions. The developed approach showed good diagnosis and higher classification accuracy for Q and φ under different operating modes.

REFERENCES

[1] N. Katayama, S. Osawa, S. Matsumoto, T. Nakano, and M. Sugiyama, “Degradation and fault diagnosis of photovoltaic cells using impedance spectroscopy,” Solar Energy Materials and Solar Cells, vol. 194, pp. 130-136, 2019.

[2] A. Rocky, M. Burhanzoi, O. Kenta, T. Ikegami, and S. Kawai, “Photovoltaic Module Fault Detection Using Integrated Magnetic Sensors,” IEEE Journal of Photovoltaics, vol. 9, no 6, pp. 1783-1789, 2019.

[3] B. K. Karmakar, and A. K. Pradhan, “Detection and Classification of Faults in Solar PV Array Using Thevenin Equivalent Resistance,” IEEE Journal of Photovoltaics, vol. 10, no 2, pp. 644-654, 2020.

[4] M. Tadj, K. Benmouiza, A. Cheknane, and S. Silvestre, “Improving the performance of PV systems by faults detection using GISTEL approach,” Energy conversion and management, vol. 80, pp. 298-304, 2014.

[5] J. A. Tsanakas, D. Chrysostomou, P. N. Botsaris, and A. Gasteratos, “Fault diagnosis of photovoltaic modules through image processing and Canny edge detection on field thermographic measurements,” International Journal of Sustainable Energy, vol. 34, no 6, pp. 351-372, 2015.

[6] A. Chouder, and S. Silvestre, “Automatic supervision and fault detection of PV systems based on power losses analysis,” Energy conversion and Management, vol. 51, no 10, pp. 1929-1937, 2010.

[7] M. Mattei, G. Notton, C. Cristofari, M. Muselli, and P. Poggi, “Calculation of the polycrystalline PV module temperature using a simple method of energy balance,” Renewable energy, vol. 31, no 4, pp. 553-567, 2006.

[8] S. R. Madeti, and S. N. Singh, “Online fault detection and the economic analysis of grid-connected photovoltaic systems,” Energy, vol. 134, pp. 121-135, 2017.

[9] M. Mansouri, A. Al-Khazraji, M. Hajji, M. F. Harkat, H. Nounou, and M. Nounou, “Wavelet optimized EWMA for fault detection and application to photovoltaic systems,” Solar Energy, vol. 167, pp. 125-136, 2018.

[10] Z. Ge, and J. Chen, J. “Plant-wide industrial process monitoring: A distributed modeling framework,” IEEE Transactions on Industrial Informatics, vol. 12, no 1, pp. 310-321, 2015.

[11] X. Liu, K. Li, M. McAfee, and G. W. Irwin, “Improved nonlinear PCA for process monitoring using support vector data description,” Journal of Process Control, vol. 21, no 9, pp. 1306-1317, 2011.

[12] E. L. Russell, L. H. Chiang, and R. D. Braatz, “Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis,” Chemometrics and intelligent laboratory systems, vol. 51, no 1, pp. 81-93, 2000.

[13] J. Zhu, Z. Ge, and Z. Song, “Distributed parallel PCA for modeling and monitoring of large-scale plant-wide processes with big data,” IEEE Transactions on Industrial Informatics, vol. 13, no 4, pp. 1877-1885, 2017.

[14] L. H. Chiang, E. L. Russell, and R. D. Braatz, Fault detection and diagnosis in industrial systems. Springer Science and Business Media, 2000.

[15] M. Z. Sheriff, M. Mansouri, M. N. Karim, H. Nounou, and M. Nounou, “Fault detection using multiscale PCA-based moving window GLRT,” Journal of Process Control, vol. 54, pp. 47-64, 2017.

[16] A. Maulud, D. Wang, and J. A. Romagnoli, “A multi-scale orthogonal nonlinear strategy for multi-variate statistical process monitoring,” Journal of process Control, vol. 16, no 7, pp. 671-683, 2006.

[17] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” Journal of educational psychology, vol. 24, no 6, pp. 417, 1933.

[18] H. Hotelling, “Multivariate quality control. Techniques of statistical analysis,” McGraw-Hill, New York, 1947.

[19] J. E. Jackson, and G. S. Mudholkar, “Control procedures for residuals associated with principal component analysis,” Technometrics, vol. 21, no 3, pp. 341-349, 1979.

[20] F. Xiao, S. Wang, X. Xu, and G. Ge, “An isolation enhanced PCA method with expert-based multivariate decoupling for sensor FDD in air-conditioning systems,” Applied Thermal Engineering, vol. 29, no 4, pp. 712-722, 2009.

[21] A. Raich, and A. Cinar, “Statistical process monitoring and disturbance diagnosis in multivariable continuous processes,” AIChE Journal, vol. 42, no 4, pp. 995-1009, 1996.

[22] H. H. Yue, and S. J. Qin, “Reconstruction-based fault identification using a combined index,” Industrial and engineering chemistry research, vol. 40, no 20, pp. 4403-4414, 2001.

[23] E. Alpaydin, Introduction to machine learning. MIT press, 2020.

[24] E. Masciari, G. M. Mazzeo, and C. Zaniolo, “A new, fast and accurate algorithm for hierarchical clustering on Euclidean distances,” In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg. pp. 111-122, 2013.

[25] S. C. Punitha, P. R. J. Thangaiah, and M. Punithavalli, “Performance analysis of clustering using partitioning and hierarchical clustering techniques,” International Journal of Database Theory and Application, vol. 7, no 6, pp. 233-240, 2014.

[26] I. Davidson, and S. S. Ravi, “Using instance-level constraints in agglomerative hierarchical clustering: theoretical and empirical results,” Data mining and knowledge discovery, vol. 18, no 2, pp. 257-282, 2009.

[27] M. A. M. Khan, Fast Distance Metric Based Data-mining Techniques Using P-trees: K-nearest-neighbor Classification and K-clustering. Diss. North Dakota State University, 2001.