



HAL
open science

Fault diagnosis of grid-connected photovoltaic systems based on unsupervised ensemble clustering and multi layer perceptron model

Mohsen Zargarani, Anne Migan-Dubois, Demba Diallo, Claude Delpha, Sara Zermani, Chabakata Mahamat, Laurent Linguet

► To cite this version:

Mohsen Zargarani, Anne Migan-Dubois, Demba Diallo, Claude Delpha, Sara Zermani, et al.. Fault diagnosis of grid-connected photovoltaic systems based on unsupervised ensemble clustering and multi layer perceptron model. 40th European Photovoltaic Solar Energy Conference and Exhibition, EU PVSEC, Sep 2023, Lisbon (Portugal), Portugal. 10.4229/EUPVSEC2023/4CV.1.27 . hal-04807800

HAL Id: hal-04807800

<https://hal.science/hal-04807800v1>

Submitted on 27 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FAULT DIAGNOSIS OF GRID-CONNECTED PHOTOVOLTAIC SYSTEMS BASED ON UNSUPERVISED ENSEMBLE CLUSTERING AND MULTI LAYER PERCEPTRON MODEL

Mohsen Zargarani^{3,1,*}, Anne Migan-Dubois¹, Demba Diallo¹, Claude Delpha², Sara Zermani³, Chabakata Mahamat³ and Laurent Linguet³

Université Paris-Saclay, CentraleSupélec, CNRS, Sorbonne Université, GeePs, Gif Sur Yvette, 91192, France

2. Université Paris-Saclay, CNRS, CentraleSupélec, L2S, Gif Sur Yvette, 91192, France

3. University of French Guiana, 97323 Cayenne cedex, French Guiana,

*Corresponding Author: mohsen.zargarani@centralesupelec.fr

ABSTRACT: Fault diagnosis in PV systems is studied broadly from various aspects. There are different techniques for FFD in PV systems. Among them unsupervised methods are appreciated for not needing labeled datasets or huge amounts of historical data. Clustering the data to study the similarities and differences is one of reliable unsupervised methods. To obtain proper accuracy in this paper, we introduce new integrated techniques for fault diagnosis and prediction. By generating different PV fault scenarios, different clustering techniques yield labels separately and all these clustering labels are gathered in an ensemble area to obtain best labels by voting. Voting majority empowers our ensemble clustering techniques. By interpreting the centroids of consensus clusters based on electrical characteristics of faults, we can generate fault labels. The final classification is outlined by neural networks (MLP). The results and experiments in a real test PV system show higher accuracy than other state of art unsupervised techniques.

Keywords: Photovoltaic (PV), fault diagnosis, consensus clustering, multilayer perceptron (MLP).

1 INTRODUCTION

Fault diagnosis in PV systems is studied broadly from various aspects.. Among them unsupervised methods are appreciated since they dispense with labeled datasets or huge amounts of historical data. Honglu Zho et al. proposed a fault diagnosis approach which was designed based on fuzzy clustering and probabilistic neural networks [2]. Shengyang Liu et al introduced an integrated method based on Gaussian kernel fuzzy C means clustering and back propagation neural networks [8]. Abbas M et al. presented an ANFIS system trained by means of subtractive clustering and grid partition technique [9]. In general, integration of clustering and neural networks played a novel role in data analysis and in different areas of science. Tanmoy Chakraborty applied both classification and clustering together and evaluated all by an optimization function [7]. It must be clarified that most of above mentioned methods applied clustering as conventional sole clustering inside their analysis and did not study the ensemble clustering with their integrated diagnosis methods. In fact, to the best of our knowledge, the proposed method in this paper is the first time that ensemble clustering and neural networks integrate in an overall diagnosis architecture. To better understand the state of art in the ensemble clustering part, the vast majority of research and methods in ensemble clustering are verified. Strehl and Ghosh developed the problem of ensemble clustering by utilizing three new techniques for solving cluster ensemble target problem. They introduced an objective function based on the ANMI and also algorithms based on meta clustering and the hyper graph partitioning [4]. Khedairia and Khadir used an iterative voting process presented in a multiple clustering combination method [6]. Compared to our proposed method, the ensemble technique in their method is somehow divergent since they used only jaccard index while in our method, all Jaccard, overlapping, and similarity coincidence indices are calculated in an iterative algorithm to distinguish the sub-clusters and similarity sets in a smarter routine. Ayad and Kamel introduced a general formulation for the voting problem as a multi-response regression

problem. Using probabilistic mapping in their voting process makes their method dissimilar to the other methods and different from the method in this paper[5].

In this paper, we introduce new integrated techniques for ensemble clustering and fault diagnosis to obtain superior accuracy. Generating different clustering yields inputs to the voting process. The algorithm interprets the centroids of final consensus clusters. The MLP accomplishes classification and fault prediction. The results show competitive accuracy than other state of art unsupervised routines.

Table I: Nomenclature

Abbreviation	description
MLP	Multi-Layer Perceptron
ANMI	Average Normalized Mutual Information
ANFIS	Adaptive Neuro-Fuzzy Inference System
D&EC	Diagnosis based on Ensemble Clustering
D&SC	Diagnosis based on Spectral Clustering
D&AC	Diagnosis based on Agglomerative Clustering
D&KC	Diagnosis based on K-means Clustering
RF	Random Forest
DT	Decision Tree
GNB	Gaussian Naïve Bayes
LR	Logistic Regression

2 STUDY AREA AND DATA

In this paper, we propose an approach for fault diagnosis in a context of use in French Guyana, where the climate is tropical and even “in some regions, the mean daily potential exceeds 5000 Wh/m²” [1]. To study the effect of internal and external (environmental) factors on PV faults, maintenance and performance, a model based on a real PV plant in French Guiana is simulated and designed. Studies on real plant and empirical results from different methods encouraged us to select unsupervised and clustering techniques. In addition, for global challenging fault scenarios, we designed different fault

category tests in a PV test module installed in SIRTAs atmospheric observatory center located in Paris.



Figure 1: PV test module, SIRTAs, Paris

3 METHODOLOGY

Overall unsupervised diagnosis comprises steps that are listed in the following algorithm table.

Algorithm I: Overall unsupervised diagnosis

	<i>description</i>
Input:	<i>data points, features extracted from the I-V curves (Isc, Voc, Imp, Vmp)</i>
Output:	<i>Predicted Fault from classification</i>
Steps:	
I.	<i>Feature Correction and Normalization) by using Equations (1) and (2)</i>
II.	<i>Various clustering generations</i>
III.	<i>Consensus clustering based on new coincidence voting majority</i>
IV.	<i>Centroid interpretation</i>
V.	<i>Fault declaration & Relabeling (Fault label generation)</i>
VI.	<i>Fault Classification : Neural Networks Classification</i>

3.1 Normalization

I-V curve features are corrected according to the standard IEC 60891 at STC condition of environment for temperature and irradiance [3]. All corrected values are normalized based on related nominal values. The collected I-V curves are corrected based on the following procedure [3]:

$$I_2 = I_1(1 + \alpha_{rel}(T_{m2} - T_{m1})\frac{G_2}{G_1}) \quad (1)$$

$$V_2 = V_1 + V_{oc1}(\beta_{rel}(T_{m2} - T_{m1}) + a \cdot \ln(\frac{G_2}{G_1})) - R_s(I_2 - I_1) - k \cdot I_2(T_{m2} - T_{m1}) \quad (2)$$

Where V_1 and V_2 , I_1 and I_2 represent voltage and current before correction and after correction. The same are for T_{m1} and T_{m2} , G_1 and G_2 which mean temperature and irradiance before and after correction. β_{rel} and α_{rel} are relative temperature coefficients for V_{oc} and I_{sc} [3].

3.2 Base clustering generation

Among numerous clustering methods, three dissimilar

techniques are preferred. K-means, spectral clustering and agglomerative clustering are selected because of their better performance and variety in their techniques and calculation philosophy.

3.3 Consensus clustering and Coincidence voting

Given two clusters from different partitions, their similarity measure as Jaccard index can be represented as:

$$Jx_{n,m}^{i,j} = \frac{|z_n^i \cap z_m^j|}{|z_n^i \cup z_m^j|} \quad (1)$$

Where z_n^i in eq. (1) represents the cardinality of elements with cluster label n in partition i and z_m^j means cardinality of elements with cluster label m in partition j. Obviously, $Jx_{n,m}^{i,j}$ is between zero and one. Here, $i = 1$ and $j = 1, \dots, P$. Values of n and m vary between one and K. Another similarity measure is interiority or overlapping index. The interiority index is measured by eq. (2) [11].

$$Ix_{n,m}^{i,j} = \frac{|z_n^i \cap z_m^j|}{\min\{|z_n^i|, |z_m^j|\}} \quad (2)$$

Although the Jaccard index seems to be a compelling tool in quantifying similarity and intersection, it is limited in considering the interiority between two clusters of separate partitions. For instance, two sample cluster sets may have the same Jaccard index but different interiority. The influence of interiority in I-V curve features can be better figured out when two clusters have the maximum interiority. In this case, it means one cluster from partition i is inside another cluster that belongs to partition j. In this instance, from a fault diagnosis point of view, they naturally exist in the same fault category. But in intersection clusters, they may either belong to the same fault category or different fault categories. The intersection clusters are more likely to be misleading in PV systems since some different faults have partly common behaviors. For example, open circuit fault and aging fault have both similar declines in voltage that may outline intersection areas of different faults while they obviously belong to different categories. For instance, two clusters of different data partitions that have similar Jaccard index but different interiority index are considered. In this case, higher interiority index value is preferred because it can demonstrate the relevant similarity of I-V curve more precisely. The coincidence similarity index covers both intersection and interiority concerns, which represents both by multiplying the two values as eq. (3) [10].

$$Cx_{n,m}^{i,j} = Jx_{n,m}^{i,j} \cdot Ix_{n,m}^{i,j} \quad (3)$$

After running diverse clustering algorithms, all cluster labels are listed separately in the new matrix. From this cluster matrix, three different performance indices make the proper stand for confusion matrix. The confusion matrices are of Jaccard, Interiority and Coincidence index

$$A^{i,j} = \begin{bmatrix} Cx_{1,1}^{i,j} & \dots & Cx_{1,k}^{i,j} \\ \vdots & \ddots & \vdots \\ Cx_{k,j}^{i,j} & \dots & Cx_{k,k}^{i,j} \end{bmatrix} \quad (4)$$

$$A = [A^{1,2} \quad \dots \quad A^{1,j} \quad \dots \quad A^{1,P}] \quad (5)$$

To secure final consensus clustering, first, the algorithm groups similar labeled clusters of the first clustering method Z^1 in rows and the clusters of other methods as Z^j in columns. Then, it calculates the cardinality of intersection for common members of grouped clusters to make intersection sets as S . In the next step, it calculates Jaccard index J_x , Interiorty Index I_x and Coincidence similarity index C_x for each member of intersect clusters. It will be followed inside intersection sets S until to come by similarity matrix A . The process attempts to realize the maximum value of coincidence similarity index, C_x in each row of similarity matrix A , then, it generates new labels for voting. By voting process, a voted list and a non-voted list will be provided separately. Non-voted list builds the new dataset and this loop frequently iterates until all remaining rows of non-voted clusters compare their intersect Coincidence similarity index and become relabeled. This method outperforms all previous methods that rely only on a single Jaccard index. However, coincidence index can reach a more sensitive and more accurate comparison base than Jaccard index or interiorty index while logically covering both of them with product operation.

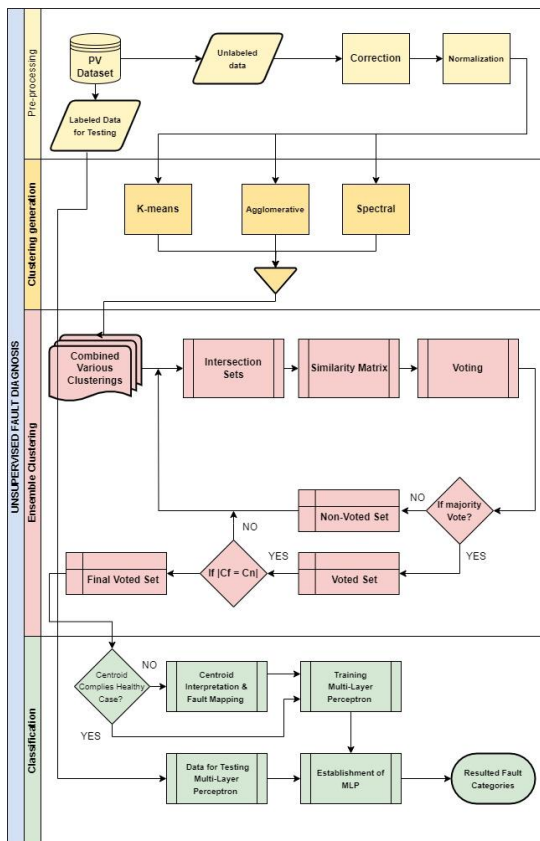


Figure 3: Overall unsupervised fault diagnosis flowchart

3.4 Mapping

After labeling the elements by ensemble clustering, the process interprets each cluster centroid for electrical characteristics to verify they belong to which fault category.

3.5 Classification

Since the mapping of the dataset is done, the process generates a labeled dataset. Proposed MLP undertakes classification and fault prediction. Based on the empirical

results, for the activation functions, RELU is used for input and hidden layers, Softmax for output layer and Adam is selected as optimizer.

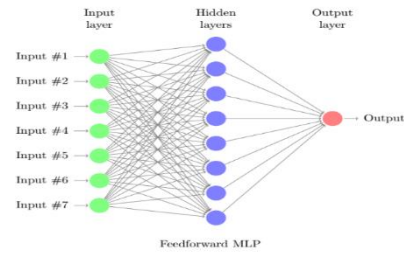


Figure 2: Feedforward MLP

4. EXPERIMENTAL SETUP FOR TEST

An experimental setup is built to verify the efficiencies of the introduced diagnosis technique. All test scenarios of PV panels are carried out for different faults and numerous weather conditions. The test platform comprises 8 PV modules which are arranged four by four in parallel. Each four panels are connected in series.

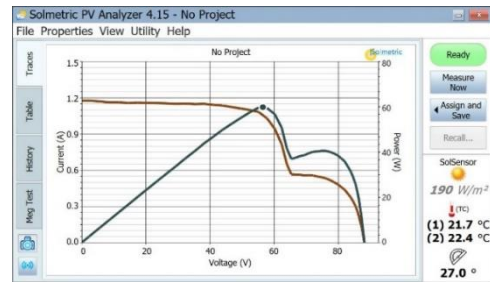


Figure 4: Partial shading in PV test module.

5. RESULT ANALYSIS

This study mainly intended to develop a novel fault diagnosis for PV systems using “ensemble-clustering then labeling algorithm” which is integrated with MLP for classification. The proposed algorithm accepts four main I-V features (I_{sc} , V_{oc} , I_{mp} and V_{mp}) plus T (temperature) and G (irradiance), as inputs.

Table II: Accuracy comparison of proposed unsupervised ensemble clustering with other unsupervised clustering methods

	D&EC	D&SC	D&AC	D&KC
Open circuit	90.83	77.59	79.0	78.5
Average aging	89.31	68.12	69.71	64.08
Partial shading	92.22	75.03	76.50	66.67

Table III: Accuracy comparison of D&EC and supervised methods

	D&EC	RF	DT	GNB	LR
Open circuit	90.83	97.83	97.29	95.08	90.22
Average aging	89.31	96.42	97.11	91.67	90.45
Partial shading	92.22	100	100	97.97	93.91

To illustrate our main contribution, compared to the state of art methods [2],[4],[5],[6] and [8], we must point out that, on the one hand, we integrated neural networks and ensemble clustering together, in a novel architecture, and on the other hand, we designed smarter philosophy for similarity analysis embedded inside iterative voting process. The performance of overall unsupervised diagnosis can be observed in tables. II and III. The accuracy of ensemble clustering diagnosis outperforms other clustering based diagnosis methods. Furthermore, the accuracy of the proposed unsupervised method is competitive to conventional supervised methods.

6. CONCLUSION

Integrated ensemble-clustering and neural networks method shapes the simultaneous advantages of both neural networks and novel ensemble-clustering notion, that improves the capability in carrying out diagnosis of the PV plant without historical knowledge about their past faults.

REFERENCES

- [1] Linguet, Laurent & Atif, Jamal. A Bayesian approach for solar resource potential assessment using satellite images. 2014, IOP Conference Series, DOI: 10.1088/1755-1315/17/1/012171.
- [2] Honglu Zhu et al. 2018. Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model. doi.org/10.1016/j.solener.2018.10.054
- [3] Baojie Li, Anne Migan-Dubois, Claude Delpha, Demba Diallo. Analysis of the Performance of the I-V Curve Correction Methods in the Presence of Defects. 37th European Photovoltaic Solar Energy Conference and Exhibition (EU PVSEC 2020), Sep 2020, Lisbon, Portugal. fhal-02903352
- [4] Strehl, A., Ghosh, J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3, 583–617.
- [5] Ayad, H.G., Kamel, M.S., 2010. On voting-based consensus of cluster ensembles. *Pattern Recogn.* 43, 1943–1953
- [6] Khedairia Soufiane et al. 2022. A multiple clustering combination approach based on iterative voting Process. *Computer and Information Sciences* 34 (2022) 1370–1380.
- [7] T Chakraborty, F Pierazzi, VS Subrahmanian. 2017. Ensemble clustering and classification for predicting android malware families - *IEEE Transactions on Dependable and Secure.*
- [8] Shengyang Liu et al. 2019. Photovoltaic array fault diagnosis based on Gaussian Kernel Fuzzy C-Means clustering algorithm. *Sensors* 2019, 19, 1520; doi:10.3390/s19071520
- [9] Muhammad Abbas and Duanjin Zhang. 2021. A smart fault detection approach for PV modules using adaptive Neuro-Fuzzy inference framework. *Energy reports* 7 (2021) 2962-2975
- [10] Luciano da Fontoura Costa. 2022. A brief Guide to the coincidence similarity and its applications. DOI: 10.13140/RG.2.2.16322.25285/7
- [11] M.K.Vijaymeena and K.Kavitha. 2016. A survey on similarity measures in text mining. *Machine learning and applications. MLAIJ Vol.3, No.1, March2016*