Applied Chemical Engineering (2025) Volume 8 Issue 2 ISSN: 2578-2010 (O)
doi: 10.59429/ace.v8i2.5659

ORIGINAL RESEARCH ARTICLE

Smart fault isolation and diagnosis in PV setups utilizing random forest classifiers

Cherifa KARA MOSTEFA KHELIL[1,2*], Ihssen HAMZAOUI[1,3], Fatma Zohra BAOUCHE[1,4], Mohamed Nadjib BENALLAL[1], Badia AMROUCHE[2,3], Kamel KARA[2]

[1] *Electrical Engineering Department, Djillali Bounaama-Khemis Miliana University, Thniet El Had Street, Khemis Miliana, Ain Defla, 44001, Algeria*

[2] *Electronics Department, SET Laboratory, Saad Dahleb Blida 1University, BP 270 Blida, 09000, Algeria*

[3] *Acoustics and civil engineering laboratory, Djillali Bounaama-Khemis Miliana University, Thniet El Had Street, Khemis Miliana, Ain Defla, 44001, Algeria*

[4] *LESI laboratory, Djillali Bounaama-Khemis Miliana University, Thniet El Had Street, Khemis Miliana, Ain Defla, 44001, Algeria*

***Corresponding author:** Cherifa KARA MOSTEFA KHELIL, k.karamostapha@univ-dbkm.dz; karamostefa_cherifa@univ-blida.dz

ABSTRACT

In the modern era, there has been a growing focus among researchers on the transition from fossil fuels to renewable energy sources, particularly photovoltaic (PV) energy, which is gaining popularity worldwide. As the development and installation of PV systems accelerate globally, it is essential to address the various faults and failures these systems may encounter. Consequently, fault diagnosis and evaluation have emerged as critical areas of study aimed at enhancing performance, improving system efficiency, and reducing maintenance costs and repair times. This paper proposes the use of a Random Forest classifier (RF) for diagnosing short circuit and open circuit faults in PV systems. The classifier is trained using machine learning algorithms to accurately identify different fault types based on real measured data from an experimental PV setup. This data encompasses weather conditions such as cell temperature and solar irradiation, as well as system parameters like current and voltage at the maximum power point, alongside performance metrics. The Random Forest classifier serves as a proactive tool for maintenance and fault diagnosis in PV systems, contributing to better overall performance and reliability. Testing on real-world data from a PV system demonstrates that this approach achieves remarkable accuracy in fault diagnosis, with a precision of 100% for current classification and around 97% for voltage classification, all within a few seconds for each parameter.

Keywords: PV setup; faults; isolation; diagnosis; machine learning; random forest classifier

ARTICLE INFO

Received: 19 April 2025
Accepted: 11 June 2025
Available online: 24 June 2025

1. Introduction

The increasing focus on renewable energies, particularly photovoltaic energy, necessitates that researchers and industry stakeholders prioritize maintaining the quality and longevity of these installations to ensure their performance, stability, efficiency, and reliability. Consequently, advanced fault diagnosis has become an indispensable trend in photovoltaic (PV) systems. In the realm of PV system diagnosis, numerous machine learning algorithms have been employed, including various types of artificial neural networks (ANN)[1-8], fuzzy logic (FL)[9], k-nearest neighbors (kNN)[10-11], and particle swarm optimization (PSO)[12]. Additionally, decision trees and

random forest classification algorithms are widely utilized for fault diagnosis in PV systems[13-16].

Random Forest (RF) classification algorithms are particularly effective in identifying and diagnosing faults such as short circuits, open circuits, shading, or module degradation. These algorithms can analyze data from real PV power plants, including current, voltage, solar irradiation, and temperature readings, to detect faults affecting system performance. Their ability to provide recommendations for troubleshooting and maintenance, combined with ease of interpretation and the capacity to handle both numerical and categorical data, makes them a valuable tool for fault diagnosis in PV systems.

In our earlier research[1], we created a new smart fault detection system (called Intelligent Fault Diagnosis or IFD) designed specifically for solar power systems connected to the electrical grid. This system proved highly effective at spotting problems, achieving impressive accuracy.

This paper utilizes the Random Forest classification algorithm for diagnosing open circuit faults, short circuit faults, and normal condition cases in PV systems. The discussion will cover the benefits and challenges of using this algorithm, as well as potential improvements and future research directions in this area.

The paper is structured as follows: Section 2 outlines the methodology, Section 3 details the experimental design and data, Section 4 presents the results and discussion, and Section 5 concludes with future perspectives on this approach.

2. Methodology

In this paper, the fault diagnosis approach is structured into three primary phases and illustrated as follow in the **Figure 1**.

2.1. Data collection phase

This initial phase involves gathering essential data from the photovoltaic (PV) array. The input data include PV temperature and solar irradiance, while the output data comprise voltage and current at the maximum power point (MPP), collected from the Maximum Power Point Tracking (MPPT) system of the PV setup. These data are crucial for initiating the subsequent phase, where machine learning (ML) algorithms are applied.

2.2. Isolation phase

This phase focuses on isolating faults using two Random Forest classifiers. The purpose is to classify cases of normal operating conditions and three distinct fault scenarios. The first Random Forest (RF) classifier is dedicated to current classification, requiring solar irradiance and current at the MPP as input data. Conversely, the second RF classifier is used for voltage classification, utilizing temperature and voltage at the MPP as inputs. Both ML algorithms have been pre-trained to classify current and voltage at the MPP separately, enabling them to categorize fault information based on detected residual data.

2.3. Identification phase

In this final phase, the outputs from both Random Forest classifiers are analyzed to locate and diagnose the specific fault within the PV array. This comprehensive analysis ensures accurate fault detection and diagnosis, facilitating targeted maintenance and repair efforts.

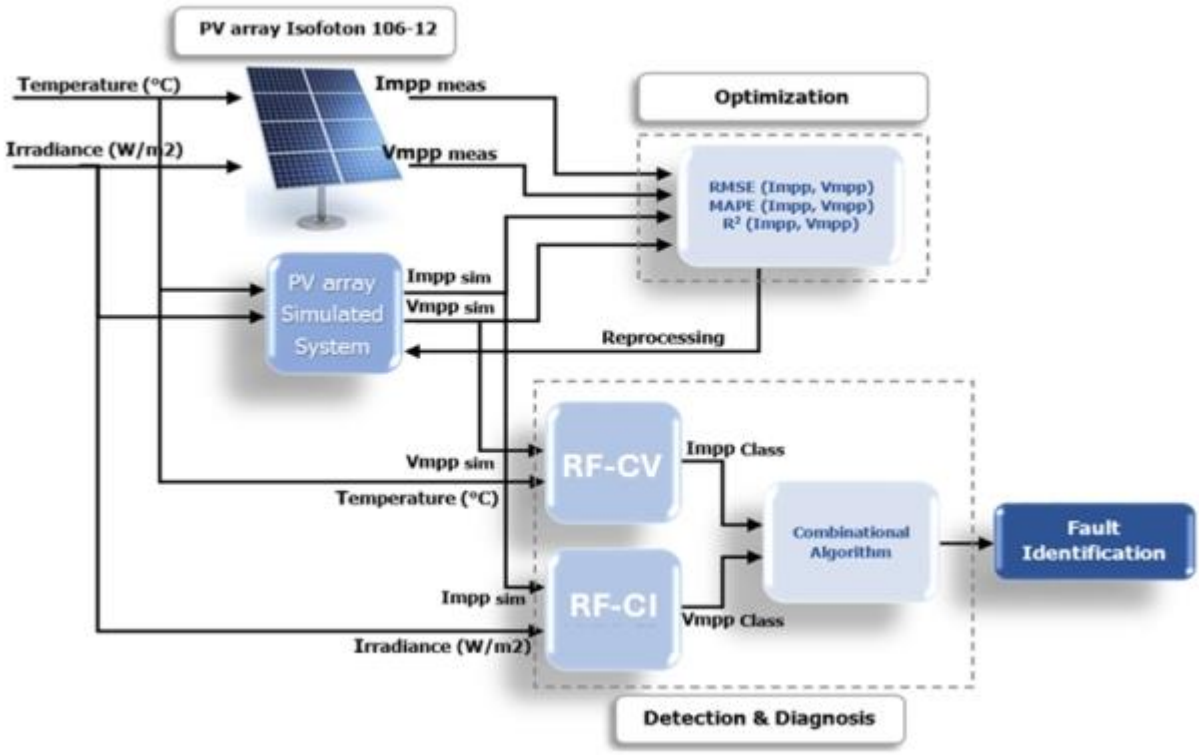

Figure 1. An overview of how the Machine Learning fault diagnosis methodology works[1].

3. Experimental plan and data

The actual approach has been focused on an experimental PV plan located in the capital Algiers, Algeria. This small grid connected PV plan is placed on the roof. As illustrated in **Figure 2**, this PV array contains 90 monocrystalline modules. The experimental PV plan contains three sub arrays and has been built to meet the needs of different research subjects. The employed data used in this study are issued from a sub array which comes from a global PV array that includes 30 PV modules linked within two parallel strings that each one string is composed by 15 modules interconnected in series[1,2]. **Figure 3** illustrates the one diode model of PV cell used in this study.

Figure 2. Roof small grid connected PV plant in Algiers, Algeria.[3]

The modeling of the GCPV to DC system requires a parameter model, defined by the Newton-Raphson Eq.1[3]:

$$I = Iph - I_0\left(exp\frac{q(V + R_sI)}{nkT_c} - 1\right) - \frac{V + IR_s}{R_{sh}} \quad (1)$$

Where:

Iph: photo generated current at STC,

Rs: cell series resistance,

Rp: cell parallel resistance,

Tc: Temperature of the cell,

K: constant of Boltzmann (1.38 x 10^{-23} J/°K),

q: charge of the electron (1.6 x 10^{-19} C),

I0: saturation current at STC.

N: the diode ideality factor.

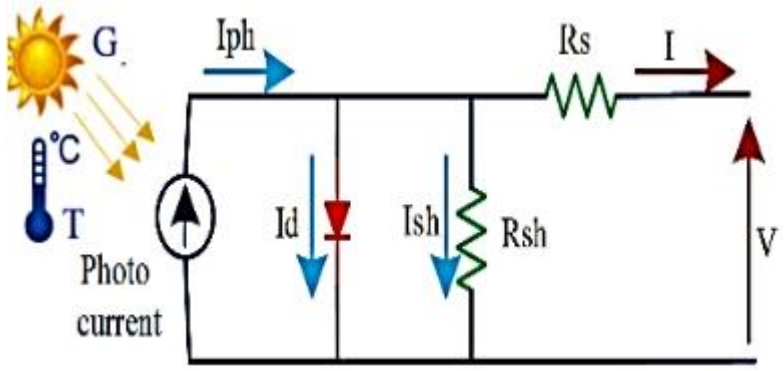

Figure 3. One diode model of the PV module.

The characteristics of the PV module as well as the electrical properties of the PV array used in this work are summarized in **Table 1** and **Table 2** respectively.

Table 1. Electrical properties of the isofoton 106-12 PV module[2].

Solar Panel electrical characteristics	Value
Peak power	106 W
Short circuit current (Isc)	6.54 A
Open circuit voltage (Voc)	21.6 V
Voltage at Maximum Power Point (Vmpp)	17.4 V
Current at Maximum Power Point (Impp)	6.10 A
Number of cells connected in Series	36
Number of cells connected in Parallel	2
Cell Short circuit current	3.27 A
Cell Open circuit Voltage	0.6 V

Table 2. Components and characteristics of PV installation[3,6].

Components	Characteristics
Global PV array	90 PV modules with monocrystalline technology
PV Sub array studied	30 PV modules divided in two strings: 15 x 15
Sunlight apparatus	Thermoelectric Pyranometer
Temperature apparatus	K-type thermocouple Pilot PV cell

Components	Characteristics
Data logger	Agilent 34970A
Inverter	IG30 Fronius

Table 2. (*Continued*)

The collected data obtained from the PV array are integrated in the two random forest classifiers as explained in section 2 (Methodology).

4. Random forest classifier

A Random Forest classifier (RD) is considered as a predictive Machine Learning (ML) model employed for classification and regression. As demonstrated in **Figure 4**, this ML classifier builds multiple decision trees during training and outputs the class that is the mode of the classes (i.e., the most common class) of the individual trees[14,16]. It enhances the performance of decision trees by averaging the results of multiple trees and introducing randomness into the model-building process to ensure robustness and reduce variance. Each decision tree in the Random Forest is a binary tree where each node represents a feature (attribute) and a threshold for splitting the data. The trees are constructed by recursively splitting the data to maximize some criterion, typically Gini impurity or information gain that is a measure employed especially in decision tree algorithms to measure the impurity or purity of a dataset. For classification, each tree in the Random Forest outputs a class label. The final prediction is determined by the majority vote among all the trees.

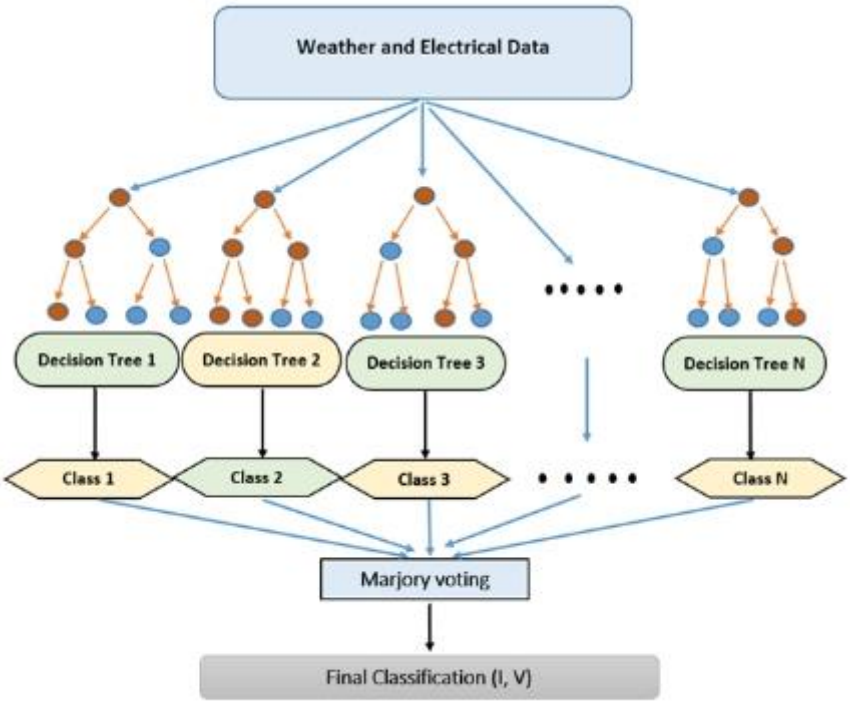

Figure 4. General flowchart of the classification of PV systems based random forest algorithm.

In Eq.2, if T is the number of trees and fi(x) is the prediction of the i-th tree for input x, then the Random Forest prediction $\hat{y}$ is[17,15]:

$$y = mode\{f_1(x), f_2(x), ..., f_T(x)\} \quad (2)$$

where the mode function returns the most frequently occurring class label among the trees.

Random Forests can also provide insights into feature importance, which measures the contribution of each feature to the prediction accuracy[16]. A common method to estimate feature importance is to use the

decrease in node impurity attributed to a feature. In Eq.3, if Impurity(T) is the impurity of the tree before the split, and Impurity(T′) is the impurity after the split, the importance of a feature can be computed as:

$$Feature\ importance = \frac{1}{N}\sum_{j}(Impurity(T) - Impurity\,(T')) \tag{3}$$

where the sum is over all nodes where the feature was used to split, and N is the number of trees or the number of times the feature is used across all trees.

For a node with class distribution p1, p2, …, pk, the Gini impurity IG is calculated as:

$$Gini(S) = 1 - \sum_{i=1}^{k} p_i^2 \tag{4}$$

Where k represents the total number of classes, while pi denotes the proportion of instances belonging to the i-th class. Then, calculating the Gini gain for each attribute in the overall dataset and selecting the attribute to provide the highest value along with creating a node for that attribute is completely necessary. The formula used to calculate the Gini gain for each attribute is as follow[16]:

$$ini_{Gain(S,A)} = Gini(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Gini\,(S_v) \tag{5}$$

Where A represents an attribute, while S refers to the dataset. Sv represents the subset of instances in S where attribute A has a value of v. Repeat steps 2 recursively for each subset of data generated by the division until all instances in a subset are classified into the same class or there are no more attributes left to divide the data.

5. Impact of random forest classifier on the quality of photovoltaic systems

Integrating a Random Forest classifier into the monitoring systems of photovoltaic (PV) installations offers a practical and powerful way to improve how faults are detected and diagnosed. By using machine learning, these systems can analyze large amounts of operational data and quickly identify issues that might otherwise go unnoticed. This approach is especially valuable because it works well with existing monitoring infrastructure, meaning operators do not need to invest in expensive new hardware or overhaul their current setups.

A key advantage of Random Forest classifiers is their ability to handle complex and varied fault conditions, such as line-to-line faults, partial shading, and temperature fluctuations. These algorithms are trained on historical and real-time data, allowing them to distinguish between different types of faults with high accuracy—recent studies have shown detection rates as high as 100% and classification accuracy near 95% when properly optimized.

The process typically involves extracting crucial parameters from the PV system, preprocessing the data to ensure quality, and tuning the classifier to maximize performance.

Moreover, integrating these models enables real-time diagnostics, so operators are alerted to problems as soon as they arise. This rapid response helps prevent minor issues from escalating into major failures, ultimately reducing downtime and maintenance costs. The use of user-friendly interfaces further ensures that the insights provided by the classifier are accessible to operators, regardless of their technical background.

Overall, embedding Random Forest classifiers into PV monitoring frameworks not only streamlines fault detection but also helps maximize energy output and operational efficiency. By leveraging advanced data analysis and existing resources, PV operators can ensure their systems remain reliable, productive, and cost-effective.

6. Effectiveness evaluation of the PV Systems diagnosis based on random forest

In order to evaluate the effectiveness of the Random Forest classifier algorithms, their results are analyzed using the most frequently used diagnosis performance indicators in science and engineering fields[2]:

a. *Accuracy:* implies how nearest is the results to the real value.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (6)$$

b. *Sensivity:* measures in what way the positive samples are correctly classified.

$$Sensivity = \frac{TP}{TP + FN} \times 100 \quad (7)$$

c. *Specificity:* measures in what way the negative samples are correctly classified.

$$Specificity = \frac{TN}{FP+TN} \times 100 \quad (8)$$

Where:

TP: true positive, signifies that the samples contain characteristics of a specific class and indeed they are classified in this class.

TN: true negative, signifies that the samples does not contain characteristics of a specific class and indeed they are not classified in this class.

FP: false positive, signifies that the samples does not contain characteristics of a specific class and they are classified in this class.

FN: false negative, signifies that the samples contain characteristics of a specific class and indeed they are not classified in this class.

Table 3 summarizes the four major categories as result of binary classification containing two rows and two columns into confusion matrix called confusion table in the intension to confirm the performance evaluation related to the classifier. The number of rows and columns depends on the number of classes. The terms true and false refer to whether the prediction corresponds to the external criticism conversely to the terms positive and negative that refer to the prediction of the classifiers.

Table 3. Confusion matrix under intermittent classification troubles[2].

Classification outcome from RFs		Classification outcome from experimental data Real label	
		True Class	False Class
Predicted Label	True Class	TP	FP
	False Class	FN	TN

7. Results and discussions

In this approach as demonstrated in **Figure 5**, global samples used are 16000 samples divided in three phases, 2000 samples are used in training phase for each attribute each attribute (PV temperature, solar irradiance, current and voltage at maximum power point) means (2000 x 4) where 400 samples are consecrated for each case, 1200 samples are employed in validation phase for each attribute (240 samples are utilized for each case) and 160 samples for each treated case that means 800 samples are turned to account in test phase

for each attribute. **Table 4** resumes all classes treated in this approach, either for current or for voltage at Maximum power point. **Figures 6** and **7** illustrate the classification results for the first and the second Random Forest classifiers algorithms respectively. This reveals the ability of the existed Random Forest to bring the incoming samples within their true classes, where all samples for the current classification in **Figure 6** are in their right classes without any misclassification between healthy case and open circuit fault in PV array, while for the voltage classification in **Figure 7** most samples are in their correct classes with few less misclassification data between healthy case and one PV module short circuited and between one PV module short circuited and three PV modules short circuited that is due to the temperature's variations. This excellent classification with a best precision is due to its high accuracy through entirety learning, robustness overfitting as well as is due to its flexibility with data types and missing values.

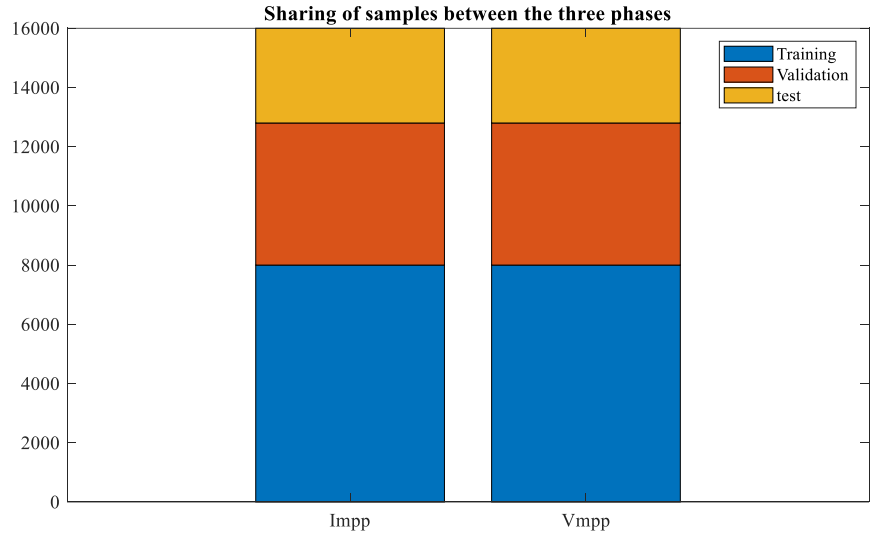

Figure 5. Measured and predicted current classification.

Table 4. Global Classes treated in fault diagnosis of PV systems.

	Classes				
Electrical Parameters	Current		Voltage		
Codes	1	2	3	4	5
Identification	Healthy	Open circuit	Healthy	1 PV module Short-Circuit	3 PV modules Short-Circuit

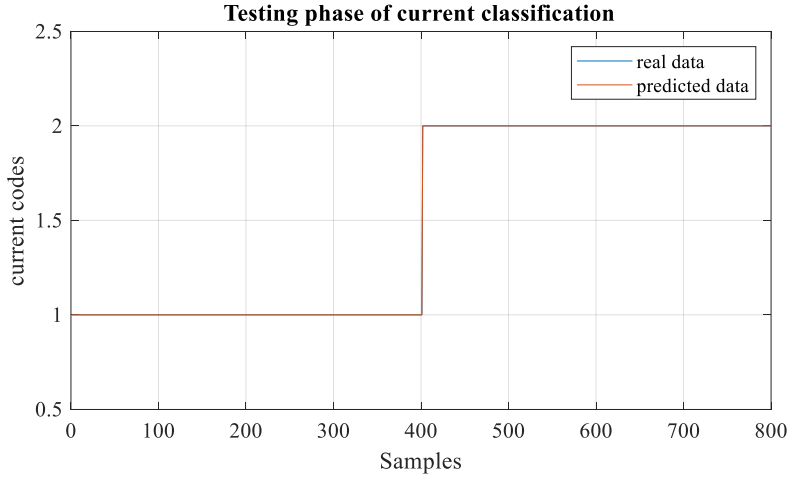

Figure 6. Measured and predicted current classification.

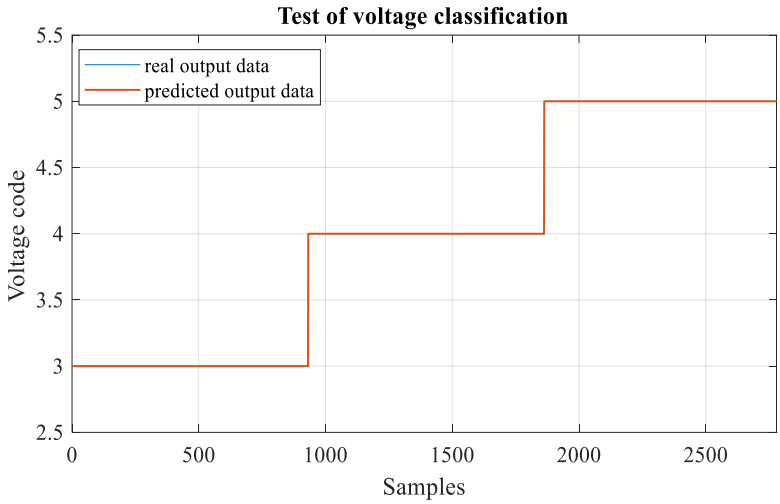

Figure 7. Measured and predicted voltage classification.

High accuracies are represented in **Figures 8** and **9** consequently displaying 100% for both classes of current classification and around 97 % for all classes of voltage classification. Current and voltage classifications represent excellent accuracy for both algorithms from the first stage of diagnosis on which the identification of fault class depends.

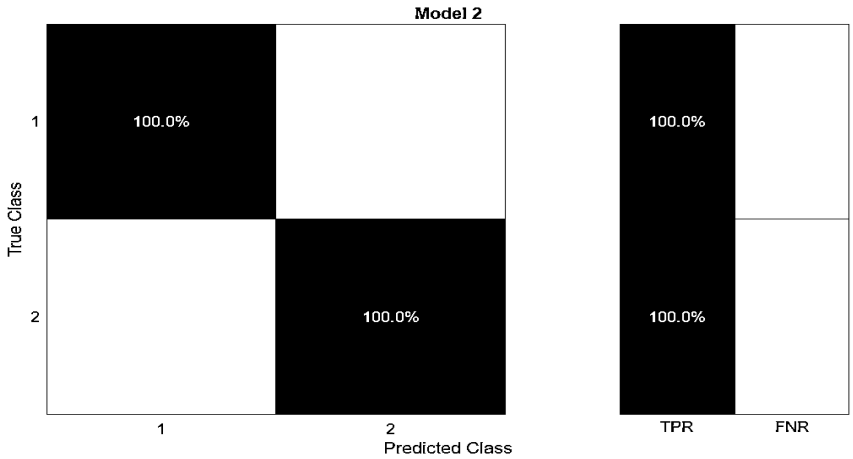

Figure 8. Current multi-classification results of RF-based PV systems.

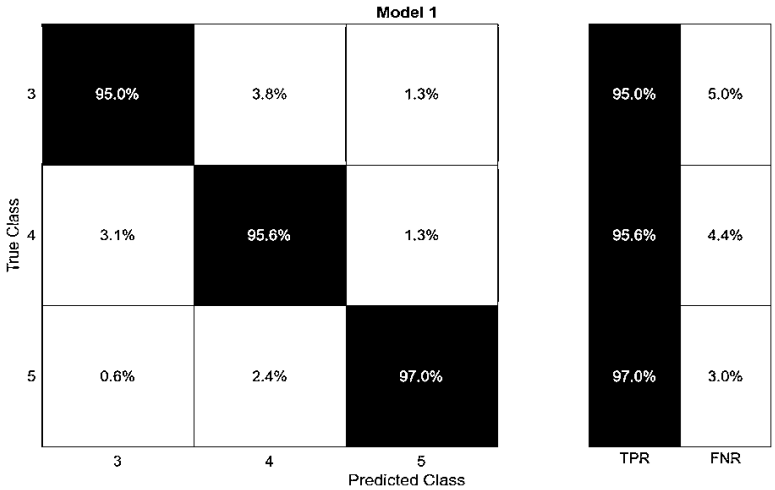

Figure 9. Voltage multi-classification results of RF-based PV systems.

The obtained results about current and voltage classification employing the Random Forest classifier display an impressive percentage accuracy which mean that this machine learning algorithm offers a robust, versatile, and user-friendly approach for diagnosing faults in photovoltaic systems. Its strengths in handling high-dimensional data, evaluating feature importance, and resisting overfitting are contributed significantly to improving the reliability and efficiency of solar photovoltaic energy systems while facilitating timely maintenance interventions.

Table 5. Effectiveness evaluation results of current and voltage in validation and test phases.

		Validation	Test
Precision (%)	Impp	99	97
	Vmpp	89	91
Sensivity (%)	Impp	100	97
	Vmpp	86	92
Specificity (%)	Impp	98	97
	Vmpp	96	97

According to the obtained results illustrated on the **Table 5**, the random forest classifiers algorithms reveal a very good outcome with high accuracy displaying from 89 to 97 % for the three key statistical concepts citing: precision, sensivity and specificity respectively. The response time of this approach is very fast according to the dataset as it requires just 30 seconds for the global diagnosis.

8. Conclusion

To enhance the performance and reliability of photovoltaic (PV) systems, fault detection and diagnosis are essential. Random Forest classifier algorithms provide a robust framework for data analysis, enabling the identification of potential faults within these systems. By leveraging Random Forest models, researchers and engineers can effectively classify and diagnose faults, thereby informing maintenance strategies and optimizing overall system efficiency. This study examines the application of Randon Forest classifier algorithms in PV fault detection and diagnosis, highlighting both their benefits and challenges. Notably, the Random Forest classifier has demonstrated exceptional precision in fault classification, achieving an accuracy of 100% for current classification and approximately 96% for voltage classification, with response times of just a few seconds for each parameter. These results are attributed to the inherent advantages of Random Forest classifiers in PV fault diagnosis, including their ability to handle complex datasets and provide rapid, accurate assessments.

Declaration of competing interest

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

Acknowledgements

The authors acknowledge the financial support data from CDER plant in BP.62 Bouzereah Observatory Road.16340, Algiers (Algeria) as well as equipment support from Laboratory of electrical systems and remote control (SET Laboratory), Electronics department, Blida 1 University, BP270 Blida, Algeria

References

1. Kara Mostefa Khelil, C., Amrouche, B., Benyoucef, A. S., Kara, K and Chouder, A. New Intelligent Fault Diagnosis (IFD) Approach for grid-connected photovoltaic systems. J Energy. 2020;211:118591. DOI: 10.1016/j.energy.2020.118591
2. Kara Mostefa Khelil, C., Amrouche, B., Kara, K and Chouder, A. The impact of the ANN's choice on PV systems diagnosis quality. j.enconman. 2021;240: 114278. DOI: 10.1016/j.enconman.2021.114278
3. Kara Mostefa Khelil, C., Amrouche, B and Kara K. Fault detection and diagnosis of GCPV systems using bayesian neural network. In Journal of Physics: Conference Series. IOP Publishing. 2022 Mar 1;2208(1):012019. DOI: 10.1088/1742-6596/2208/1/012019
4. Li B, Delpha C, Diallo D, Migan-Dubois A. Application of Artificial Neural Networks to photovoltaic fault detection and diagnosis: A review Ren & Sust Ener Rev. 2021;138:110512. DOI: 10.1016/j.rser.2020.110512
5. Islam M, R R Masud, A Md Tofael, I A. K. M. Kamrul and Tlemçani M. Artificial Intelligence in Photovoltaic Fault Identification and Diagnosis: A Systematic Review. Energies.2023;16:7417. DOI: 10.3390/en16217417
6. Amiri A F, Kichou S, Oudira H, Chouder A and S Santiago. Fault Detection and Diagnosis of a Photovoltaic System Based on Deep Learning Using the Combination of a Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU). Sustainability. 2024;16:1012. DOI: 10.3390/su16031012
7. Arévalo P, Cano A, Darío B, Jurado F .Fault analysis in clustered microgrids utilizing SVM-CNN and differential protection. j.asoc.2024;164:112031. DOI: 10.1016/j.asoc.2024.112031
8. Kara Mostefa Khelil, C., Kara, K and Chouder, A. Fault detection of the photovoltaic system by artificial neural networks. CEEE. 2017; 4:60-65.
9. Kara Mostefa Khelil, C., Amrouche, B., Kara, K and Chouder, A. Newfound Intelligent solution for grid connected PV systems diagnosis based on CANFIS algorithm. Tob Regul Sci. 2023; 9(1):3809-3844. DOI: doi.org/10.18001/TRS.9.1.286
10. Madeti, S.R and Singh, S.N. Modeling of PV system based on experimental data for fault detection using kNN method. Sol Energy .2018;173:139–51. DOI: 10.1016/j.solener.2018.07.038
11. Godfrey B Z, Kara Mostefa Khelil C, Godfrey M G, Taane Z. Identification of PV Fault Classes Using Intelligent Method KNN (K-Nearest Neighbours). IJRSI. 2024;10:1108093. DOI: 10.51244/IJRSI.2024.1108093
12. Liao Z, Wang D, Tang L, Ren J, Liu Z. A Heuristic diagnostic method for a PV System: Triple-Layered Particle Swarm Optimization–Back-Propagation Neural Network. Energies. 2017;10: 226. https://doi.org/10.3390/en10020226
13. Liu, Y., Zhang, L., & Zhang, L [Random Forest with Class-Balanced Decision Trees for Multi-class Classification. IEEE Transactions on Neural Networks and Learning Systems. 2019; 30(8):2286-2297.
14. Somesh, L., Chakravarty, A and Maiti, A. Fault Diagnosis in Power Transmission Line using Decision Tree and Random Forest Classifier. 2022 IEEE 6th International Conference on Condition Assessment Techniques in Electrical Systems (CATCON). 17-19 December 2022.
15. Qiang Wang, Thanh-Tung Nguyen, et al An efficient random forests algorithm for high dimensional data classification. Advances in Data Analysis and Classification. 2018; 12: 953–972. DOI: 10.1007/s11634-018-0318-1
16. Amiri A F, Oudira H, Chouder A, Kichou S. Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier. j.enconman. 2024; 301: 18076. DOI: 10.1016/j.enconman.2024.118076
17. Zhi-Hua Zhou.(2025). Ensemble Methods Foundation and algorithms second edition. Taylor and Francis group. ISBN: 978-1-032-96061-6. DOI:10.1201/9781003587774.