PHOTOVOLTAIC SYSTEM FAULT DETECTION AND CLASSIFICATION BASED ON K-NEAREST NEIGHBOR

PHÁT HIỆN VÀ PHÂN LOẠI LỖI TRÊN HỆ THỐNG PV DỰA TRÊN GIẢI THUẬT K HÀNG XÓM GẦN NHẤT

Bui Quang Minh[1,*], Nguyen Dang Duong[1], Dao Quang Tung[1], Nguyen Duc Tuyen[1]

DOI: https://doi.org/10.57001/huih5804.2023.037

ABSTRACT

Faults in photovoltaic systems (PV) have critical damage to the operation and be the main reason for power loss. To avoid these consequences, the detection and analysis of PV systems faults must be enhanced effectively and promptly. In this paper, a fault detection and classification model is developed and improved. Faults such as open-circuit (OC), line-to-line (L-L), and partial shading of the systems are detected and classified in K-nearest neighbor (kNN). Based on the simulation performed in MATLAB/Simulink, the dataset of the three errors is collected and split at the rate of 75% for the training and 25% for the testing. The dataset is built with noise conditions as in practice, and the collecting progress is automated. The absolute error is confined to the range of 0.5% and 5%. It is observed that the result from the proposed model gives high accuracy of 99.8%.

Keywords: *Photovoltaic, fault detection and classification, k-nearest neighbor.*

TÓM TẮT

Các lỗi trong hệ thống điện năng lượng mặt trời (PV) ảnh hưởng rất tiêu cực lên năng suất và vận hành. Để hạn chế hậu quả này, việc phát hiện và phân tích những lỗi trong hệ thống PV cần được phát triển nhanh chóng và hiệu quả để kịp thời sửa chữa. Trong bài báo này, một mô hình phát hiện và phân loại lỗi được xây dựng và phát triển. Những lỗi như hở mạch, ngắn mạch và bóng che một phần ở trong hệ thống sẽ được phát hiện và phân loại bằng giải thuật K hàng xóm gần nhất (kNN). Dựa vào mô phỏng MATLAB/Simulink, dữ liệu của ba loại lỗi được thu thập và phân loại với tỷ lệ 75% cho tập huấn luyện và 25% cho tập kiểm tra. Tập dữ liệu được xây dựng với điều kiện nhiễu như trong thực tế và quy trình thu thập được tự động hóa. Sai số tuyệt đối được giới hạn trong phạm vi 0,5% và 5%. Theo quan sát, kết quả từ mô hình đề xuất cho độ chính xác cao đến 99,84%.

Từ khóa: *Hệ thống PV, phát hiện và phân loại lỗi, giải thuật K hàng xóm gần nhất.*

[1]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology

[*]Email: minh.bq210585@sis.hust.edu.vn

Received: 22/10/2022

Revised: 03/02/2023

Accepted: 15/3/2023

1. INTRODUCTION

Renewable energy is gradually increasing with rapid scaling due to global concern about climate change and greenhouse emissions. Among all the renewable energy sources, photovoltaic (PV) power generation makes up the majority because of its safety, unlimited source distribution, and high energy quality. According to the 21st-century Renewable Energy Policy Network report [1], the global installed capacity of PV power generation in 2020 was 760GW in total and 133GW newly installed in 2021, accounting for 51.8% of 2021 renewable electricity expansion, followed by wind power [2]. With the innovation in PV technologies, solar power has drawn global interest and has become one of the most sustainable energy sources in the future. However, PV systems rely on environmental conditions and often operate under several harsh weather statuses, which not only directly affects systems performance but also leads to various faults. Therefore, to improve the performance of the PV system and minimize power loss, faults appearing in the system need to be clear promptly, and faults detection and classification become a necessity.

Solar cell I-V characteristics curves are basically a representation of the operation of a solar cell or PV module summarising the relationship between the current and voltage at the certain conditions of irradiance and temperature. Specifically, the different faulty modes affect the I-V characteristics of the PV string in different ways, leaving distinct signatures during its operation. In [3], mismatch and shading faults, connectivity faults, and short circuit faults are investigated, the author uses the CIE algorithm to classify different behaviours of a PV system based on the differences in the I-V characteristic of each condition. Applying machine learning has also been developed to enhance the result of PV fault detection and classification. The author of [4] used Support Vector Machines (SVM) with Principal Component Analysis (PCA) for feature extraction to diagnosis four types of faults (short circuit fault, inverse bypass diode fault, shunted bypass diode fault, and shadowing effect). The paper [5] being about Artificial Neural Network (ANN) was studied with the goal of monitoring the working condition of three photovoltaic modules. A radial basis network was trained to classify faults with realistic levels of noise. The impressive result of this model is over 97% in classification. Ref-[6] is

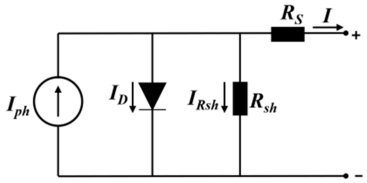
based on k-nearest neighbours (kNN). In this paper, experimental data from the manufacturer's datasheet is reported under standard test conditions (STC) and normal operating cell temperature (NOCT). The observed result is accurate, which is 98.70% in detecting and classifying open circuit faults, line-to-line faults, partial shading with and with-out bypass diode faults and partial shading with inverted bypass diode faults in real time.

Three faults, commonly occuring in constructing the system and operating under practice weather conditions, which are analysed in this paper including: open-circuit (OC) faults, line-to-line (L-L) faults and partial shading conditions. The authors also propose an optimized kNN model to diagonose the faults. The remaining parts of this paper are organized as follows: Section 2 shows the simulation setup, data collection, and proposed detect-and-classify model. Section 3 reports the result and decomposes the efficiency of the method. The final section presents conclusions and future work of the paper.

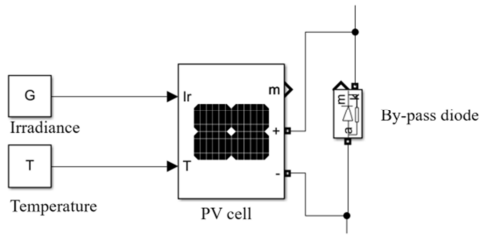
2. METHODOLOGY

2.1. Simulation setup

2.1.1. The photovoltaic array



(a) Single diode model



(b) Simulink model of PV panel

Figure 1. Simplified equivalent circuit of PV panel

The PV module is widely constructed of a single diode model (SDM) [7]. As can be seen in Fig. 1a, the SDM consists of a current source, a diode, a series and a parallel resistance. The current-voltage I-V characteristic of the PV module can be given by Kirchhoff's current law [7].

$$I = I_{ph} - I_0 \cdot \left\{ \exp\left(\frac{V + IR_s}{n_s V_t} \right) - 1 \right\} - \frac{V + IR_s}{R_{sh}} \quad (1)$$

where V and I are the ouput voltage and current, I_{ph} and I_o are the photo-generated current and the dark saturation current, V_t is the junction thermal voltage, R_s and R_{sh} are the series and parallel resistances, and n_s is the number of cells in the module connected in series. Fig. 1b shows its simulation on MATLAB used in this paper. The irradiance and temperature can be adjusted from MATLAB workspace using the G and T constant blocks, respectively. The case study is conducted in mono-crystalline PV module name WE5M-18 which is supplied by World Energy. The datasheet of the PV module is given in Table 1.

As seen in Fig. 2, the entire PV array is constructed by three strings, each consisting of three series connected PV modules. The currrent and voltage of the system are measured by sensors and then transmitted to the power box to monitor generated power. The load box presents a common electric load. The current (A), voltage (V) and power (W) values are collected and sent to MATLAB workspace for the data collection step.

Table 1. Data sheet of WE5M-18 PV module

Parameter	Value
Peak power (Pmpp) (W)	5
Production Tolerance (%)	-33
Maximum Power Current (I_{mp}) (A)	0.55
Maximum Power Voltage (V_{mp}) (A)	9
Short Circuit Current (I_{sc}) (A)	0.64
Open Circuit Voltage (V_{oc}) (V)	10.53
Temperature coefficient of V_{oc} (%/°C)	-0.361
Temperature coefficient of I_{sc} (%/°C)	0.102

A program is built with MATLAB script for running the simulation automatically. With each environmental conditions and faulty, the system will run once to get and save the data to long-term storage. The irradiance varies from 600W/m^2 to 1000W/m^2 with a fixed step of 20W/m^2 while the temperature performs from 23°C to 50°C with a fixed step size 2°C to match the noise weather condition as practical PV systems in operation. The system is set up under the normal working conditions, and three faulty ones containing partial shading, open-circuit and line-to-line. The MATLAB-based program runs to generate the desired information as of the four variables: irradiance, temperature, voltage and current. The data after each iteration is collected and then automatically stored in a CSV file into four fields corresponding to four collected variables. The I-V profile of each condition is collected with the 0.001-second resolution by adjusting the load profile.

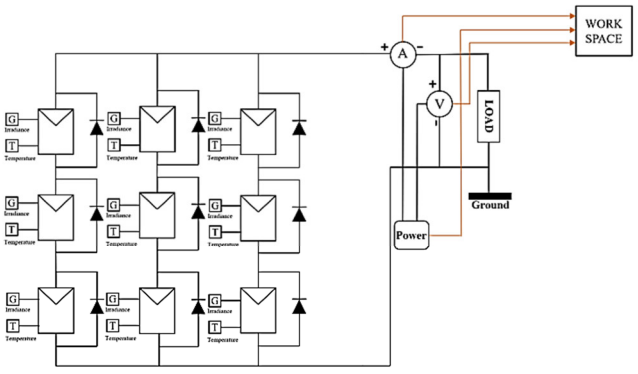


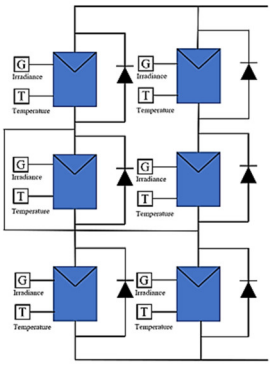
Figure 2. MATLAB/Simulink simulation model

2.1.2. Type of PV array faults

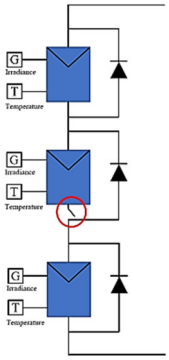
An accidental connection between two PV array nodes leads to the L-L fault. These issue can occur because of cable insulation failure, short-circuit problems between the current-carrying conductors, mechanical damage, water intrusion, or corrosion of the DC junction box. This problem causes the voltage across the affected string to drop, which causes the remaining strings to draw a substantial amount of back-feeding current. The back-feeding current may be stopped, and the faulty strings can be isolated using typical over-current protection devices. In this paper, the L-L fault is reproduced by connecting 2 points in the middle of 2 parallel strings together. The cases study in this type are generated variously, such as connecting between line 1 and line 2 as in Fig. 3a, between line 1 and line 3 or line 2 and line 3.

The open circuit defect can occur for several causes, including damaged cells, physically worn-out cable couplings, loose connections, and aging power cables close to the terminal lead. The bypass diode keeps the circuit continuous during this malfunction and permits current to flow into the inverter. However, this defect causes the voltage of the faulty string to decrease while the current stays constant, therefore significantly reducing the output power. As illustrated in Fig. 3b, faults are simulated as losing connection between the negative pole of PV panel and the system string. The error is reproduced with one opened-circuit to six opened-circuit panels.

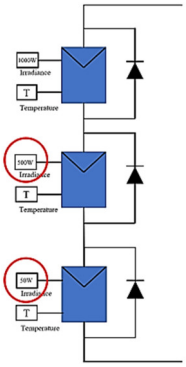
The partial shading problem can be brought on by uneven solar radiation on modules as a result of moving clouds, nearby trees and buildings, and module mismatches. If any of the PV cells or modules in a PV string are shaded, those cells or modules become reverse-biased and function as a resistive load to dissipate electricity. This raises the temperature of a PV module or cell, which causes fire dangers and ultimately results in the irreversible loss of the entire PV system. The losses brought on by partial shading can be decreased by the use of bypass diodes. Similar to the case of an open circuit fault, the partial shading conditions are divided from 1 shaded panel to 6 shaded panels, corresponding to 10% to 70% shaded rate of the PV array. Fig. 3c shows an instance of a partially shaded PV array when two modules receive less irradiance than the other which is unshaded.



(a) Line-to-line



(b) Open-circuit



(c) Partial shading

Figure 3. Type of PV array faults

2.2. Application of k-nearest neighbor on PV faults classification

KNN is an example-based approach [8] used to calculate how many test samples are closest to k. Based on the labels of these samples, the test label is chosen. The program calculates the Euclidean distance between each sample in a dataset and updates the data as follows:

$$(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

KNN algorithm generally requires an integer k which is a set of labeled samples (training data), and a metric to measure distances. Due to its simplicity, the implementation is straightforward. In Fig. 4, the kNN method labeled the data as type 2 data for a "star" data when k = 4 and labeled the data as type 1 as k = 10. In kNN, the number of considered nearest neighbors k is the most crucial hyperparameter. If k is too small, the model will be under-fitting, whereas being over-fitting and require high computational time if k is too large. Therefore, a hyper-parameter-based optimization has been applied to maximize the efficiency of the model. The model can determine the most appropriate algorithm itself by setting the "algorithm" object to "auto" in a Python module for machine learning called Scikit-learn [9]. Fig. 5 shows the flowchart of detection and classification based on the generated dataset.

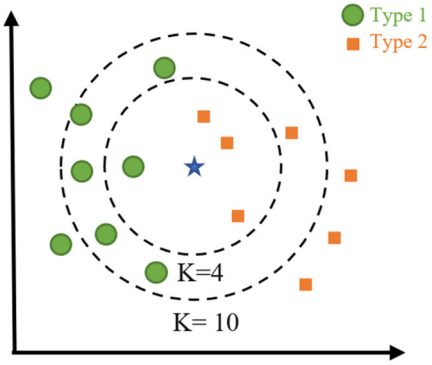


Figure 4. Data labelling by the kNN

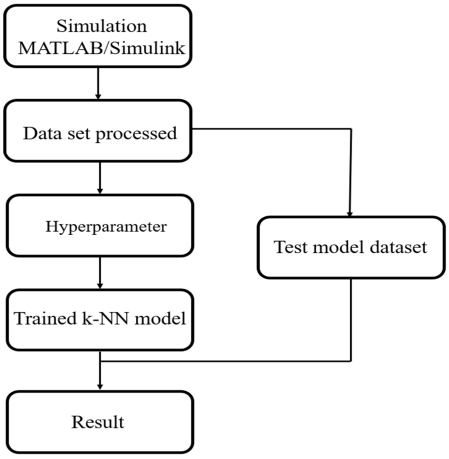


Figure 5. Main process

During operation, weather conditions such as radiation and temperature are key factors in shaping system outputs. The fault-type parts are also affected during different weather. Two different error types, which operate under different irradiance and temperature, can produce two I-V identical curves, leading to confusing error classification results based on the model's curve. To overcome the consequences, the model analyses the combination between 2 indicators of the characteristic curve and two variables representing environmental conditions, such as irradiance and temperature, to increase the model's accuracy and place the problem-solving problem close to the practice.

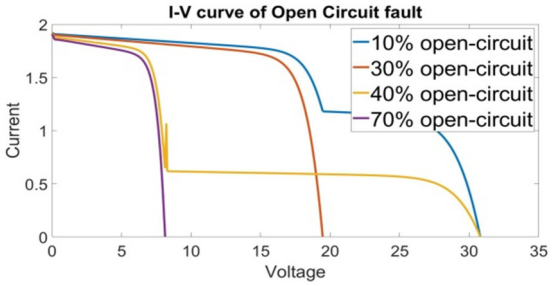
3. RESULT AND DISCUSSION

In the simulation, the system operating under each state produces different I-V curves. Fig. 6 sample representation of the three error states reproduced in this article. The cases are all carried out under the STC (G = 1000W/m^2 and T = 25°C). The difference in characteristics can be seen in each case of different conditions. Under different environmental conditions (different radiation and temperatures), the I-V curve also represents the state of the entire system. Based on this difference, a dataset consisting of a sequence of data of I-V lines is collected, as mentioned in the previous section. Table 2 shows the amount of data collected after the simulation. With the case of partial shading, there are more cases than other states, so the dataset of this case has higher qualitative.

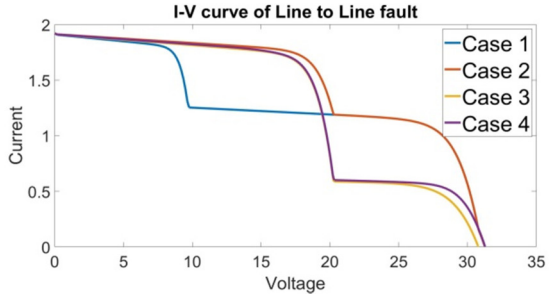
Table 2. Dataset statistics

Condition	Number of samples
Normal	156,315
Partial shading	357,164
Open circuit	187,149
Line-to-line	224,504
Total	**925,132**

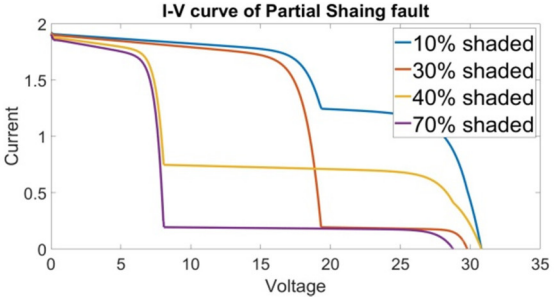
After the required dataset is completely collected, is divided into two sets: 693848 samples (75%) are used to train the model and the rest (25%) for testing. During the training process, the model will run through hyperparameters to maximize the output of the problem. The model is built with the help of GridSearchCV library constructed by Python programming language. The optimization process is set up and results in k = 2. Therefore, the k-parameter is used in the kNN model 2. After the model was trained, the rest of the dataset will be passed through the model and the predicted ressult is produced. A result with high accuracy was obtained, with 99.84%. With this result, error detection and error classification can be considered very accurate. Fig. 7 shows the confusion matrix of the model. The confussion chart also indicates that the open-circuit and line-to-line cases are a bit difficult to classify, but not too significant compared to the sum of the adequately classified data. As can be observed, the model has the highest accuracy rate for classifying the line-to-line condition (99.92%), followed by the partial shading case (99.89%). The open-circuit, with a classification accuracy of 99.57%, is the worst scenario. The I-V curve in this state under the same environmental conditions is similar to the remainder of the mistakes, which is why. The model is still quite powerful even with an open circuit error misclassification rate of 0.43%.



(a) Open circuit



(b) Line-to-line



(c) Partial shading

Figure 6. The I-V curves from the considered faulty conditions

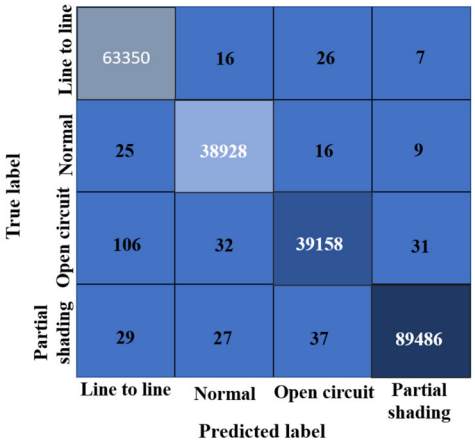


Figure 7. The confussion chart of the result

4. CONCLUSION

This paper presents a PV fault detection and classification strategy conducted by the k-Nearest Neighbor model with a hyperparameter-based optimizing process. The dataset used for training and validating is achieved by automatic simulation. The proposed approach shows its effectiveness in diagnosing multiple faults available in PV systems. The model performs effectively with an accuracy of 99.84% with a testing dataset. In future work, the authors will carry out a 3×3 experimental PV array with the types of faults reproduced as in the simulation to validate whether the module works well in practice. Moreover, the dataset will be improved by an extended range of temperature and irradiance, adding the parameters of different PV panels to help the model operate successfully with various types of PV systems.

REFERENCES

[1]. *Renewable energy policy network for the 21st century.* https://www.ren21.net/wp-content/uploads/2019/05/GSR2021_Full_Report.pdf.

[2]. *Renewable Energy and Jobs Annual Review 2022 of IRENA.* https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2022/Sep/IRENA_Renewable_energy_and_jobs_2022.pdf?fbclid=IwAR09zS78Q4_pIg73JUUUYv0dB_E1Syc39foyFcvxBXurVJFg6m7OmAvR0c.

[3]. B. Zbib, H. Al Sheikh, 2020. *Fault Detection and Diagnosis of Photovoltaic Systems through I-V Curve Analysis*. 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1-6, doi: 10.1109/ICECCE49384.2020.9179390.

[4]. R. K. Mandal, N. Anand, N. Sahu, P. Kale, 2020. *PV System Fault Classification using SVM Accelerated by Dimension Reduction using PCA*. 2020 IEEE 9th Power India International Conference (PIICON), pp. 1-6, doi: 10.1109/PIICON49524.2020.9112896.

[5]. Muhammed Hussain, Mahmoud Dhimish, Sofya Titarenko, Peter Mather, 2020. *Artificial neural network based photovoltaic fault detection algorithm integrating two bi-directional input parameters.* Renewable Energy, Volume 155, p. 1272-1292, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2020.04.023.

[6]. Hosseinzadeh J., Masoodzadeh F., Roshandel E., 2019. *Fault detection and classification in smart grids using augmented K-NN algorithm*. SN Appl. Sci. 1, 1627. https://doi.org/10.1007/s42452-019-1672-0.

[7]. A. Chatterjee, A. Keyhani, D. Kapoor, 2011. *Identification of Photovoltaic Source Models*. in IEEE Transactions on Energy Conversion, vol. 26, no. 3, pp. 883-889, doi: 10.1109/TEC.2011.2159268.

[8]. Guo G., Wang H., Bell D., Bi Y., Greer K., 2003. *KNN Model-Based Approach in Classification*. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62.

[9]. Ghawi Raji, Pfeffer Jürgen, 2019. *Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity*. Open Computer Science, vol. 9, no. 1, pp. 160-180. https://doi.org/10.1515/comp-2019-0011.

THÔNG TIN TÁC GIẢ

Bùi Quang Minh, Nguyễn Đăng Dương, Đào Quang Tùng, Nguyễn Đức Tuyên

Trường Điện - Điện tử, Đại học Bách khoa Hà Nội