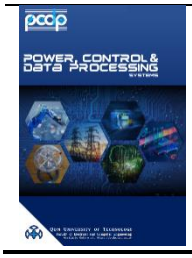
DOI: 10.30511/pcdp.2026.2092605.1089 - ISSN: 3060-8376 - Vol 3(4), 2026



POWER, CONTROL AND DATA PROCESSING SYSTEMS

Available Online at: https://pcdp.qut.ac.ir/

Fault Detection and Classification in PV Arrays with Machine Learning Algorithms

ARTICLE INFO

Article Type
Original Research

Authors
Afshin Najafi[1]
Mohammadhossein Tamimi[1]
Ehsan Shahi[1]

[1] Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran

*** *Correspondence***
Address: Faculty of Electrical Engineering, K. N. Toosi University of Technology, Seyed-Khandan Bridge, Shariati Ave., Tehran, Iran.
Postal Code: 1631714191.
Phone: +9821-88881003
Fax: -
afnajafiafshin@gmail.com
m.tamimi@email.kntu.ac.ir
e.shahi@email.kntu.ac.ir

Article History
Received: June 26, 2026
Accepted: July 21, 2026
ePublished: December 01, 2026

ABSTRACT

Fault detection and classification in photovoltaic arrays are important for improving the reliability, safety, and energy production of photovoltaic systems. This study presents a two-stage evaluation of machine learning-based PV fault diagnosis using both simulation-generated and measured experimental data. First, a 64-module PV array with an 8×8 configuration is modeled in MATLAB/Simulink under ten normal and faulty operating conditions. Support Vector Machine and Naive Bayes classifiers are evaluated before and after Principal Component Analysis preprocessing. PCA-SVM achieves the highest accuracy of 97.93% in the ten-class simulation study, compared with 86.50% and 43.125% for SVM and Naive Bayes, respectively. Second, external validation is performed using twelve measured current-voltage curves obtained from six physical PV panels under clean and partial-shading conditions. The measured curves are processed into 3,600 observations and evaluated using leave-one-panel-out validation. Six classifier families, including SVM, Naive Bayes, Random Forest, XGBoost, LightGBM, and MLP are investigated both with and without PCA. Random Forest achieves the highest experimental-data accuracy of 97.53%. PCA increases SVM accuracy from 68.14% to 79.69% and XGBoost accuracy from 83.33% to 90.03%, while its effect is unfavorable for several other classifiers. The results demonstrate that PCA can improve feature separability, but its effectiveness is classifier- and dataset-dependent. The external validation confirms the usefulness of electrical and environmental features for measured partial-shading diagnosis while also highlighting the need for broader field validation.

Keywords: Photovoltaic array; Fault diagnosis; Fault classification; Principal component analysis; External validation; Support vector machine; Random forest.

1 Introduction

According to the International Renewable Energy Agency, the global installed solar photovoltaic capacity reached approximately 2,391.584 GW by the end of 2025 [1]. The continued expansion of photovoltaic (PV) generation has been supported by reductions in module costs, improvements in manufacturing and maximum power point tracking technologies, and the growing need to reduce dependence on fossil-fuel-based electricity generation. However, the increasing scale of PV installations has made system reliability, fault diagnosis, and preventive maintenance increasingly important. Undetected faults can reduce energy production, accelerate component degradation, increase maintenance costs, and, in severe cases, create electrical and fire hazards.

PV arrays are generally constructed through series and parallel connections of multiple PV modules and are continuously exposed to changing irradiance, temperature, humidity, dust, storms, and partial shading. Consequently, they are vulnerable to electrical, environmental, and physical abnormalities, including short-circuit faults, open-circuit faults, bridge faults, line-to-ground faults, partial shading, hot spots, bypass-diode failures, cable degradation, and increased series resistance [2-4]. A recent review emphasized that reliable PV monitoring should not only detect the presence of an abnormal condition but should also classify its type and, where possible, identify its location. Nevertheless, robustness under real operating conditions remains a major challenge for existing diagnostic techniques [2].

Conventional PV protection devices, including fuses, relays, and circuit breakers, are not always capable of identifying all DC-side faults. The nonlinear current-voltage characteristics of PV modules, low fault currents under weak irradiance, operation of bypass diodes and maximum power point tracking systems, and the dependence of PV output on environmental variables can weaken or mask fault signatures [3], [4]. Visual inspection and thermal imaging can identify physical defects and hot spots, while electrical methods based on current-voltage curves, power losses, and model residuals can provide useful diagnostic information [5]. However, these techniques may require specialized equipment, periodic manual inspection, accurate physical models, or fault-specific thresholds that are difficult to define under continuously changing environmental conditions.

For these reasons, data-driven fault diagnosis has received considerable attention. Machine learning models can learn relationships between electrical and environmental measurements and PV operating conditions without requiring a separate analytical threshold for every fault type. Earlier studies applied k-nearest neighbor algorithms to experimental PV measurements [6], random forests to array voltage and string-current signals [7], and principal component analysis to extract informative characteristics from current-voltage curves [8]. PCA has also been combined with support vector machines to reduce feature redundancy and improve fault separation [9], while SVM-based methods have been used to detect line-to-line faults under variable PV operating conditions [10]. The effect of data contamination has also been investigated through a two-stage framework involving data filtering, decision-tree labeling, and random-forest classification, demonstrating that noisy measurements and incorrect labels can considerably affect PV fault-diagnosis performance [11].

Recent studies have further extended the use of conventional machine learning in PV diagnosis. Nedaei et al. reconsidered classical machine learning and dataset-reduction techniques to develop less complex fault-classification structures [12]. Hassan et al. investigated concurrent PV faults using feature selection and several SVM kernels and compared the resulting models with random forest, logistic regression, and k-nearest neighbor classifiers [13]. Their results demonstrated that feature selection can reduce computational time while maintaining diagnostic performance, although simultaneous faults with similar electrical characteristics remain difficult to distinguish. Pillai et al. evaluated machine learning models using operational measurements from a 33-kW PV installation, demonstrating the importance of validation using field data rather than relying exclusively on simulation-generated samples [14]. These studies indicate that conventional machine learning remains relevant when transparency, computational efficiency, and ease of implementation are important.

Deep learning has also become an important direction in PV fault diagnosis. Ali and Ding proposed a feature-engineering and one-dimensional convolutional neural network framework that was evaluated on four PV array topologies under clean and noisy conditions [15]. Sabati et al. combined several pretrained convolutional neural networks with Bitterling Fish Optimization to select informative image features and reduce feature redundancy [16]. Navid et al. introduced an experimentally validated fully connected neural network using back-surface temperature and electrical measurements for PV fault diagnosis [17]. Seghiour et al. combined convolutional neural networks with Wigner–Ville time-frequency analysis and compared the developed framework with several conventional machine learning models [18]. More recently, Talaat et al. incorporated deep learning and visual explanation mechanisms into PV fault classification, reflecting the increasing attention given to the interpretability of diagnostic decisions [19].

Label availability and model transparency represent two additional challenges in intelligent fault diagnosis. Vahidi et al. developed a modified semi-supervised generative adversarial network for PV arrays that uses both labeled and unlabeled samples. The method achieved high classification performance with only a small proportion of labeled data and was further assessed under Gaussian noise and using a real PV benchmark dataset [20]. In a related industrial fault-diagnosis application, Vahidi et al. combined a convolutional neural network with Gradient-weighted Class Activation Mapping to provide interpretable thermal-image-based diagnosis of induction-motor faults [21]. At the broader power-system level, data-driven surrogate models have also been employed to learn complex optimization responses and reduce the need for repeated model-based calculations [22]. Although the latter two studies do not directly address PV array fault classification, they demonstrate the broader movement toward data-driven, computationally tractable, and explainable decision-making in energy and industrial systems.

Despite these advances, several research gaps remain. First, many existing studies focus on a single classifier family or evaluate dimensionality-reduction techniques only in combination with one specific model. Consequently, it remains unclear whether the benefit of PCA is general or strongly dependent on the learning mechanism

of the classifier. Second, simulation-based assessments remain common in PV fault diagnosis, whereas independent validation using measured data is frequently limited. Third, controlled comparisons among margin-based, probabilistic, ensemble-tree, boosting, and neural-network classifiers before and after identical PCA preprocessing remain relatively scarce. Addressing these issues requires a framework in which the same feature definitions, validation procedure, PCA criterion, and performance metrics are consistently applied to different classifier families.

Accordingly, the revised study does not claim that PCA or the investigated classifiers are newly developed algorithms. Instead, its scientific contribution is a controlled analysis of the interaction between PCA-based feature transformation and classifiers with different learning mechanisms, supported by both a ten-class simulation dataset and an independent measured PV dataset. The main contributions are summarized as follows:

- A structured ten-class simulation dataset is generated using a 64-module PV array with an 8×8configuration under different irradiance, temperature, fault-type, and fault-severity conditions.
- An external experimental validation is conducted using 3,600 observations extracted from twelve measured I–V curves of six physical PV panels under clean and partial-shading conditions.
- The effect of PCA is systematically evaluated across six classifier families, namely SVM, Naive Bayes, Random Forest, XGBoost, LightGBM, and MLP, using identical classifier parameters before and after PCA.
- A leave-one-panel-out validation strategy is adopted for the measured dataset to prevent information leakage between observations originating from the same physical panel.
- A comprehensive evaluation is provided using accuracy, precision, recall, Macro-F1, Weighted-F1, Cohen's kappa, confusion matrices, and panel-level statistical comparisons.
- The simulation results demonstrate that PCA substantially improves both SVM and Naive Bayes. The external measured-data results show that PCA improves SVM and XGBoost, while its effect is classifier-dependent and is not necessarily beneficial for all probabilistic, tree-based, or neural-network classifiers.

The remainder of this paper is organized as follows. Section 2 presents the simulated PV array, the studied fault scenarios, the PCA formulation, the classification models, the measured experimental dataset, and the adopted validation procedures. Section 3 reports the ten-class simulation results, the external measured-data validation, the comparison of classifiers before and after PCA, the statistical analysis, and the limitations of the study. Finally, Section 4 summarizes the main conclusions and presents specific directions for future research.

2 Methodology

The methodology consists of two complementary evaluation stages. In the first stage, a 64-module PV array is modeled in MATLAB/Simulink under ten normal and faulty operating conditions. The generated dataset is used to compare SVM, Naive Bayes, PCA-SVM, and PCA-Naive Bayes. In the second stage, an external dataset containing measured I-V curves from physical PV panels is used to evaluate six classifier families before and after PCA. The same seven electrical and environmental variables are employed in both stages. This two-stage structure enables the original classification framework to be evaluated under a controlled ten-class simulation problem and subsequently assessed using independently measured PV data.

2.1 PV Array Model

Photovoltaics is the process in which a PV cell or module converts solar irradiance into electrical energy. The output voltage and current of a PV module change with irradiance and temperature. Therefore, the voltage-current and power-voltage characteristics are important for analyzing the behavior of PV systems under normal and faulty operating conditions. Figure 1 shows the V-I and P-V characteristics of a PV module under standard test conditions.

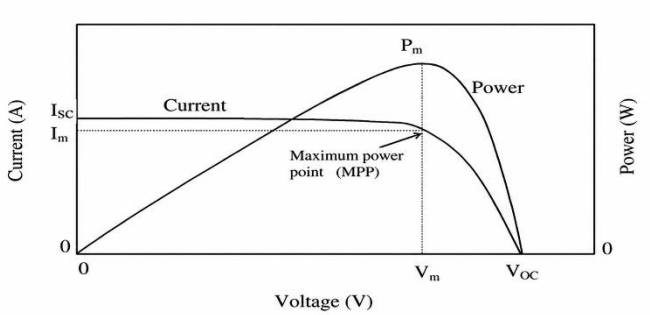


Figure 1: V-I and P-V characteristics of the PV module under standard test conditions.

As shown in Figure 1, the main electrical parameters of a PV module include open-circuit voltage, short-circuit current, maximum power point voltage, maximum power point current, and maximum output power. These variables are important because different PV array faults can change the shape of the V-I and P-V curves and shift the main operating points from their normal values.

A PV power system is usually composed of several parts, including PV arrays, DC-DC converters, DC-AC inverters, controllers, and the grid-side connection. Figure 2 shows the general structure of a PV system. In this structure, PV modules are connected in series and parallel to increase the generated voltage, current, and power. The DC-DC converter and its controller regulate the DC-side operating point, while the inverter converts the DC power into AC power for grid connection.

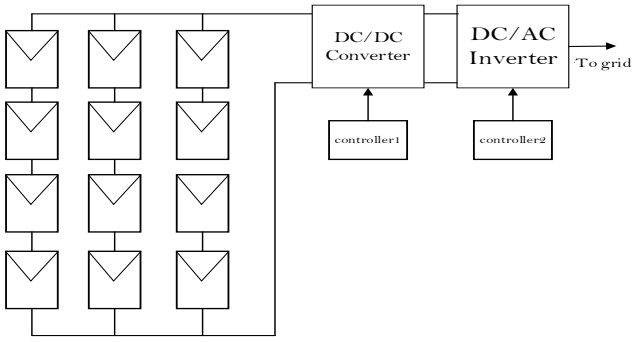


Figure 2: General structure of the photovoltaic system.

Although Figure 2 presents the general PV system structure, the main focus of this paper is on the PV array part, because most DC-side electrical and environmental faults directly affect the output characteristics of the array. Therefore, the simulated PV array and its fault conditions are explained in the next subsection.

2.2 PV Array Model and Fault Scenarios

In this study, a PV array is simulated in MATLAB/Simulink to generate the required dataset. The studied array consists of 64 PV modules arranged in an 8 × 8 structure. This configuration includes several series and parallel connections, which makes it suitable for studying different fault types, such as short-circuit, open-circuit, partial shading, and diode-related faults. Figure 3 shows the simulated PV array structure and the locations of the studied fault scenarios.

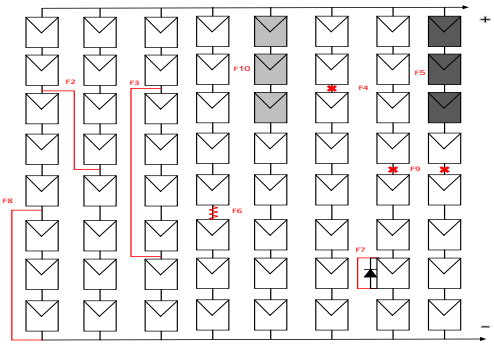


Figure 3: Simulated 8 × 8 PV array structure and the locations of the studied fault scenarios.

As shown in Figure 3, different faults are applied to selected modules, strings, or connections of the PV array. These faults are selected to represent common operating problems in PV arrays. In total, ten operating conditions are considered, including one normal condition and nine fault conditions. The studied operating conditions are summarized in Table 1.

Table 1: Studied operating conditions of the PV array.

Fault type	Fault sign
Normal operation	F1
Bridge fault	F2
Short circuit fault 50%	F3
One string open circuit fault	F4
Partial shading fault - receives 65% of the radiation	F5
Series resistance increase fault	F6
Module diode short circuit fault	F7
Short circuit fault 37.5%	F8
Two strings open circuit fault	F9
Partial shading fault - receives 85% of the radiation	F10

As shown in Table 1, the dataset includes both normal and faulty operating states. Some faults have similar physical origins but different severity levels. For example, two short-circuit faults and two partial shading faults are considered. This makes the classification task more challenging than a simple normal/fault detection problem, because the classifier must distinguish between different fault types and different fault severities. The simulations are performed under different environmental conditions to make the generated dataset more representative. The temperature is varied from 7°C to 55°C, and the irradiance is varied from 100 W/m² to 1200 W/m². For each operating condition, seven input variables are recorded. These variables include five electrical variables, namely voltage, current, power, maximum power point voltage, and maximum power point current, and two environmental variables, namely temperature and irradiance. Therefore, for each operating condition, the generated samples form an input feature matrix with seven features. The samples are labeled according to their corresponding operating condition from F1 to F10. As a result, the PV fault diagnosis task is formulated as a ten-class supervised classification problem.

A total of 8,000 simulation samples were generated, including 800 samples for each of the ten operating conditions. The dataset was divided using a stratified hold-out procedure, with 80% of the samples used for training and the remaining 20% reserved as an independent test set. Consequently, the training and test subsets contained 6,400 and 1,600 samples, respectively. For each operating condition, 640 samples were used for training and 160 samples were used for testing. Feature standardization and PCA were fitted exclusively using the training data and were subsequently applied to the independent test samples.

2.3 Overall Classification Framework

The overall classification framework is illustrated in Figure 4. The study uses two complementary data sources: a ten-class MATLAB/Simulink dataset and an external measured PV dataset. For both datasets, seven input variables are employed, including voltage, current, power, maximum-power-point voltage, maximum-power-point current, temperature, and irradiance.

In the simulation-based evaluation, SVM and Naive Bayes are assessed using both the original and PCA-transformed feature spaces. In the external validation, SVM, Naive Bayes, Random Forest, XGBoost, LightGBM, and MLP are evaluated before and after PCA. Feature standardization and PCA are fitted exclusively using the corresponding training data to prevent information leakage. The models are evaluated using accuracy, precision, recall, F1-score, Macro-F1, Weighted-F1, Cohen's kappa, and confusion matrices.

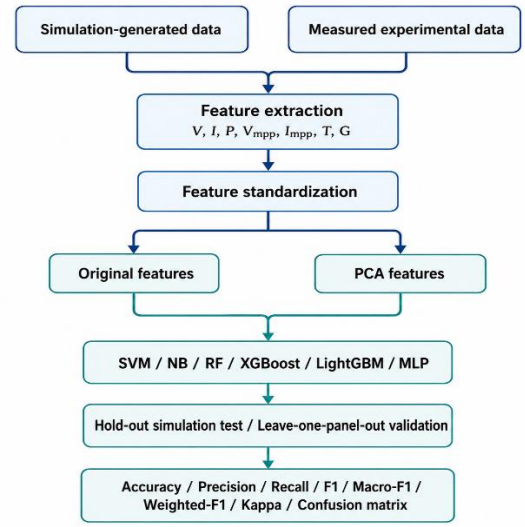


Figure 4: Overall workflow of the simulation-based and external experimental PV fault-classification framework.

2.4 Principal Component Analysis

The original input dataset may include redundant information because several electrical variables of the PV array are correlated with each other. For example, changes in irradiance and temperature can affect voltage, current, and power at the same time. This redundancy can increase the computational burden and may reduce the classification performance. Therefore, PCA is used as a preprocessing step before applying the classification models. PCA transforms the original input variables into a new set of uncorrelated variables called principal components. These components are ordered according to the amount of variance that they explain in the original dataset. In this study, PCA is applied to reduce the input feature space and preserve the most useful information for the classification process.

Let X be the input data matrix, where each row represents one sample and each column represents one input variable. First, the input data are mean-centered as follows:

$$X_c = X - \mu \quad \text{Eq 1}$$

where μ is the mean vector of the input variables, and X_c is the mean-centered data matrix. Then, the covariance matrix of the input data is calculated as:

$$C = 1/(N-1)\, X_c{}^{\mathrm{T}} X_c \quad \text{Eq 2}$$

where N is the number of samples. After calculating the covariance matrix, eigenvalue decomposition is applied. The eigenvectors corresponding to the largest eigenvalues are selected to construct the reduced feature space. The transformed feature matrix is obtained as:

$$Z = X_c W_k \quad \text{Eq 3}$$

where W_k contains the selected eigenvectors, and Z is the reduced feature matrix. In this paper, PCA is used to investigate whether feature reduction can improve the performance of SVM and Naive Bayes classifiers.

Two PCA-selection procedures were applied according to the evaluation stage. For the ten-class simulation dataset, the first three principal components were retained. These components explain 62.1%, 21.3%, and 12.0% of the total variance, respectively, corresponding to a cumulative explained variance of 95.4%. For the external measured dataset, PCA was fitted independently within each leave-one-panel-out training fold. The minimum number of components preserving at least 95% of the cumulative variance was retained. Depending on the held-out panel, four or five components were selected, explaining between 95.17% and 97.55% of the training-data variance. Figure 7 illustrates the cumulative explained-variance profile of the measured dataset.

2.5 Support Vector Machine Classifier

Support Vector Machine is a supervised learning method that has been widely used for classification problems. The main idea of SVM is to find a separating hyperplane that maximizes the margin between different classes. The samples located near the separating hyperplane are called support vectors, and they play an important role in defining the decision boundary. Figure 5 illustrates the general concept of the SVM classifier.

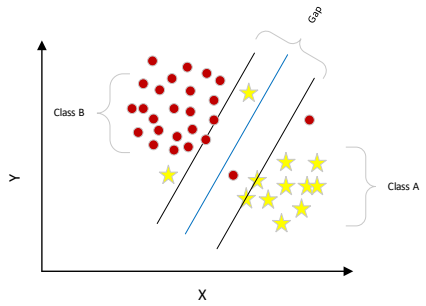


Figure 5: General concept of the Support Vector Machine classifier.

For a binary classification problem, the SVM optimization problem can be written as:

$$min\ 1/2\ ||w||^2 + C \sum_{r=1}^{l} \zeta_r \quad \text{Eq 4}$$

subject to:

$$y_r(w^T \varphi(X_i) + b) \geq 1 - \zeta_r \quad \text{Eq 5}$$

where X_i is the input sample, y_r is the class label, w and b define the separating hyperplane, ζ_r is the slack variable, C is the penalty parameter, and $\varphi(.)$ is the kernel mapping function. If the input data are not linearly separable, the kernel function maps the input data into a higher-dimensional feature space to improve class separation. Since the PV fault diagnosis problem in this paper includes ten operating classes, a multi-class SVM structure is used to classify the samples from F1 to F10. The SVM classifier is trained and tested in two cases. In the first case, the original input features are used directly. In the second case, the PCA-transformed features are used as the input of the SVM classifier.

In this study, a multi-class SVM configuration is implemented using the one-vs-one (OVO) reduction strategy to effectively handle the ten-class PV fault diagnosis task. To map the highly non-linear relationships between the PV array operating points under varying environmental fluctuations, a Radial Basis Function (RBF) kernel is selected. The RBF kernel function is mathematically defined as follows:

$$K(X_i, X_j) = e^{\left(-\gamma ||X_i - X_j||^2\right)} \quad \text{Eq 6}$$

where γ represents the kernel scale parameter. To ensure optimal generalization ability and avoid overfitting, a systematic grid-search optimization technique integrated with 5-fold cross-validation was applied to determine the optimal hyperparameters. Consequently, the penalty coefficient (C) was selected as 10, and the optimized kernel scale parameter (γ) was established as 0.25. The model was trained and validated using the cross-validation dataset before evaluating its final performance on the independent test samples.

The same RBF kernel, C, and γvalues were retained for each SVM/PCA-SVM comparison. Therefore, the performance difference within each pair can be attributed to PCA preprocessing rather than to different classifier settings.

2.6 Naive Bayes Classifier

Naive Bayes is used as a benchmark classifier in this study. This method is based on Bayes' theorem and estimates the probability of each class for a given input sample. The main assumption of Naive Bayes is that the input features are conditionally independent given the class label. Although this assumption is simple, Naive Bayes is a fast and useful baseline classifier for many classification problems.

For an input sample z, the predicted class is obtained as:

$$\hat{y} = arg\,max\, P(c)\ \Pi\ P(zj \mid c) \quad \text{Eq 7}$$

where c is a possible class label, $P(c)$ is the prior probability of class c, and $P(zj \mid c)$ is the conditional probability of the j-th input feature given class c. In this study, Naive Bayes is applied to the same dataset used for SVM. Similar to the SVM case, the classifier is evaluated once using the original input features and once using the PCA-transformed features. Gaussian Naive Bayes was evaluated using both the original and PCA-transformed features. The same probabilistic model configuration was employed before and after PCA, allowing the effect of feature transformation to be assessed independently of classifier hyperparameter changes.

2.7 Classification Procedure and Performance Evaluation

After generating the dataset, the samples are used to train and test the selected machine learning classifiers. The main purpose of this stage is to compare the classification ability of SVM and Naive Bayes and also to examine the effect of PCA preprocessing on their performance. For this purpose, four cases are evaluated. In the first case, the original input features are directly used with the SVM classifier. In the second case, the original input features are used with the Naive Bayes classifier. In the third case, PCA is first applied to the input data, and the reduced feature matrix is then used with the SVM classifier. In the fourth case, the PCA-transformed features are used with the Naive Bayes classifier.

The performance of the classifiers is evaluated using confusion matrices, accuracy, precision, recall, Macro-F1, Weighted-F1, Cohen's kappa, and mean squared error. MSE is retained only as an auxiliary index for the simulation-based comparison. The confusion matrix provides a detailed view of the classification results by comparing the actual and predicted labels for all operating conditions. In this matrix, the diagonal elements show the correctly classified samples, while the off-diagonal elements show the misclassified samples. Therefore, a classifier with larger diagonal values and smaller off-diagonal values has better classification performance. The classification accuracy is calculated as:

$$Accuracy = Nc\ /\ Nt\ \times\ 100\% \quad \text{Eq 8}$$

where Nc is the number of correctly classified samples, and Nt is the total number of tested samples.

Since accuracy alone may not fully represent multiclass classification performance, class-specific precision, recall, and F1-score are also calculated. For class c, these metrics are defined as:

$$Precision_c = \frac{TP_c}{TP_c+FP_c} \quad \text{Eq9}$$

$$Recall_c = \frac{TP_c}{TP_c+FN_c} \quad \text{Eq 10}$$

$$F1_c = \frac{2Precision_c Recall_c}{Precision_c+Recall_c} \quad \text{Eq 11}$$

Macro-F1 is calculated as:

$$Macro-F1 = \frac{1}{k}\sum_{c=1}^{K} F1_c \quad \text{Eq 12}$$

Weighted-F1 is defined as:

$$Weighted-F1 = \sum_{c=1}^{K} \frac{n_c}{N} F1_c \quad \text{Eq 13}$$

where K is the number of classes, $n_c$is the number of test samples belonging to class c, and $N$is the total number of test samples.

Cohen's kappa is additionally employed in the external validation to quantify agreement beyond chance. Mean squared error is retained as an auxiliary performance index and is calculated as follows:

$$MSE\ =\ 1/N\ \Sigma(yi\ -\ \hat{y}i)^2 \quad \text{Eq 14}$$

where yi is the actual numerical class label, $\hat{y}i$ is the predicted numerical class label, and N is the number of tested samples. Mean squared error is retained only as an auxiliary index for consistency with the original simulation-based comparison. Since MSE depends on the numerical coding assigned to the class labels, it is not considered a primary performance metric for multiclass classification. Accuracy, precision, recall, F1-score, Macro-F1, Weighted-F1, Cohen's kappa, and confusion matrices are used as the main classification measures in the revised evaluation.

2.8 External Experimental Dataset and Data Processing

To complement the simulation-based evaluation, an external validation was conducted using the publicly available experimental PV dataset cited in [23] and [24]. The dataset contains measured current-voltage characteristics, solar irradiance, module temperature, and thermographic information obtained from outdoor experiments on SP090P and HYBRYTEC-M5-30/12 photovoltaic modules under clean, dirty, and partial-shading conditions.Since partial shading is directly included among the operating conditions considered in the simulation study, only the clean and partial-shading measurements were selected for external validation. The selected dataset contains twelve measured I-V sweeps obtained from six physical panels, including six clean curves and six partial-shading curves. Table 2 summarizes the composition of the selected experimental data.

Table 2: Composition of the external measured PV dataset used for validation.

Condition	Measured I–V sweeps	Physical panels	Processed observations
Clean	6	6	1,800
Partial shading	6	6	1,800
Total	12	6 unique panels	3,600

Each measured I-V curve contains a dense sequence of voltage and current observations. To obtain a uniform representation and reduce high-frequency measurement fluctuations, the voltage range of each curve was divided into 300 equal-width intervals. The median voltage and current values within each interval were retained. Empty intervals, when present, were completed through one-dimensional interpolation between neighboring current measurements.The corresponding power was calculated as $P = VI$. The maximum-power-point voltage $V_{\mathrm{mpp}}$and maximum-power-point current I_{mpp}were determined from the operating point producing the maximum measured power. Module temperature and irradiance were then added to the electrical variables. Consequently, each processed observation contained seven features: V, I, P, V_{mpp}, I_{mpp}, T, and G.The processing procedure resulted in 3,600 experimental observations, including 1,800 clean and 1,800 partial-shading observations. Figure 6 presents representative measured I-V curves from the two module technologies.

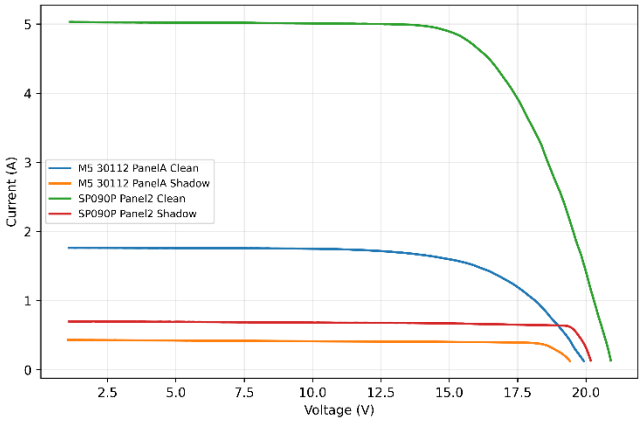


Figure 6: Representative measured I-V curves of HYBRYTEC-M5-30/12 and SP090P photovoltaic modules under clean and partial-shading conditions.

Before applying PCA, the seven variables were standardized using parameters estimated exclusively from the training panels. PCA was subsequently fitted only to the training part of each validation fold. Figure 7 presents the cumulative explained variance of the principal components, while Figure 8 displays the measured observations in the space of the first two principal components.

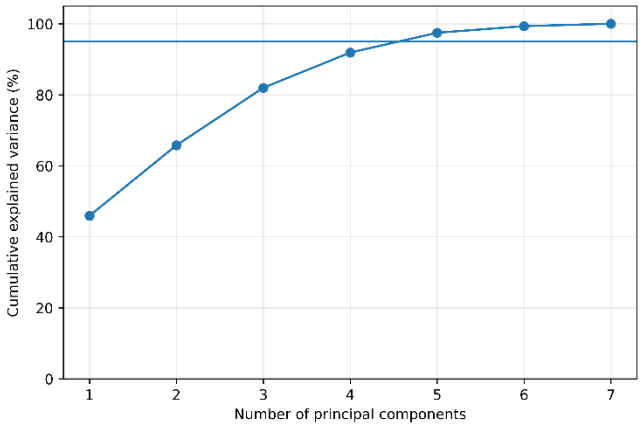


Figure 7: Cumulative explained variance of PCA for the external measured PV dataset. The minimum number of components preserving at least 95% of the total variance was retained.

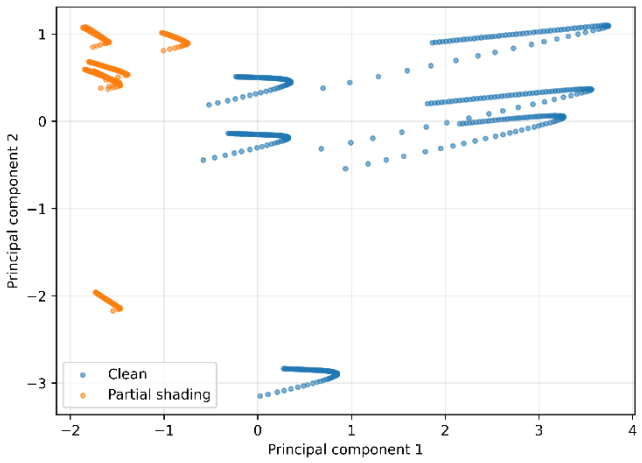


Figure 8: Projection of the measured clean and partial-shading observations onto the first two principal components.

2.9 Benchmark Classifiers and External Validation Protocol

Six classifier families were evaluated using the external measured dataset: SVM, Gaussian Naive Bayes, Random Forest, XGBoost, LightGBM, and a shallow multilayer perceptron. Each classifier was evaluated once using the original features and once using PCA-transformed features. Identical classifier parameters were used within each original/PCA pair. Therefore, the measured difference within each pair reflects the effect of PCA preprocessing rather than a change in classifier settings.Random Forest combines the predictions of multiple decision trees trained using bootstrap samples and random feature subsets. XGBoost and LightGBM are gradient-boosting methods that sequentially construct decision trees to reduce previous classification errors. The MLP classifier contains two hidden layers with 32 and 16 neurons and uses the ReLU activation function and Adam optimizer.

Because 300 observations were extracted from each measured I-V curve, a conventional random point-wise train-test split could place strongly related observations from the same physical panel in both subsets. This could introduce information leakage and produce overly optimistic results. Therefore, leave-one-panel-out cross-validation was employed.In each of the six folds, all clean and partial-shading observations belonging to one physical panel were used as the independent test set. The observations from the remaining five panels were used for training. Each fold therefore contained 3,000 training observations and 600 test observations.

Feature standardization and PCA were fitted exclusively using the training panels in each fold and were subsequently applied to the held-out panel. Aggregated out-of-fold predictions were used to calculate accuracy, Macro-Precision, Macro-Recall, Macro-F1, Weighted-F1, Cohen's kappa, and confusion matrices. Table 3 summarizes the classifier settings.

Table 3: Hyperparameter settings of the classifiers used for external validation.

Classifier	Main settings
SVM	RBF kernel, (C=10, γ =0.25)
Naive Bayes	Gaussian Naive Bayes
Random Forest	120 trees, square-root feature sampling, balanced class weights
XGBoost	100 trees, maximum depth 3, learning rate 0.05, subsample 0.8
LightGBM	100 estimators, 15 leaves, learning rate 0.05
MLP	Hidden layers 32 and 16, ReLU, Adam, early stopping
PCA	Standardization followed by components preserving at least 95% of variance

3 Results and Discussion

This section presents two complementary evaluations. First, SVM, Naive Bayes, PCA-SVM, and PCA-Naive Bayes are compared using the ten-class MATLAB/Simulink dataset. Second, external validation is performed using measured clean and partial-shading data. The external analysis evaluates six classifier families before

and after PCA using leave-one-panel-out validation. The results are discussed in terms of overall performance, class-specific errors, the effect of PCA, statistical significance, and practical limitations.

3.1 Fault Classification Results Without PCA

First, the original input features are directly used to train and test the SVM and Naive Bayes classifiers. The confusion matrix of the SVM classifier is shown in Figure 9. This matrix compares the actual labels and the predicted labels for the ten operating conditions considered in this study. The diagonal elements show the correctly classified samples, while the off-diagonal elements show the misclassified samples.

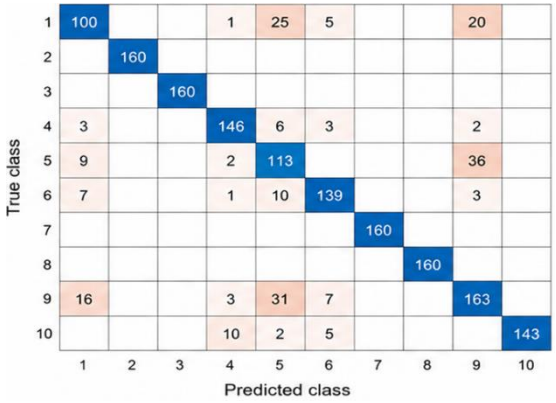


Figure 9: Confusion matrix of the SVM classifier.

SVM achieves an overall classification accuracy of 86.50%. The concentration of samples near the main diagonal indicates that the RBF-SVM classifier can model much of the nonlinear relationship between the electrical and environmental variables and the ten operating conditions. Nevertheless, several misclassifications remain among operating conditions with similar electrical effects, particularly partial-shading, open-circuit, and normal operating conditions. The additional precision, recall, and F1-based measures are reported in Table 4.

Naive Bayes achieves an accuracy of only 43.125%, which is substantially lower than that of SVM. The confusion matrix contains numerous off-diagonal elements, and several actual classes are assigned to a limited number of predicted classes. This weak performance can be attributed to the conditional-independence assumption of Naive Bayes. Voltage, current, power, temperature, and irradiance are physically correlated, and therefore the original feature space does not adequately satisfy this assumption.

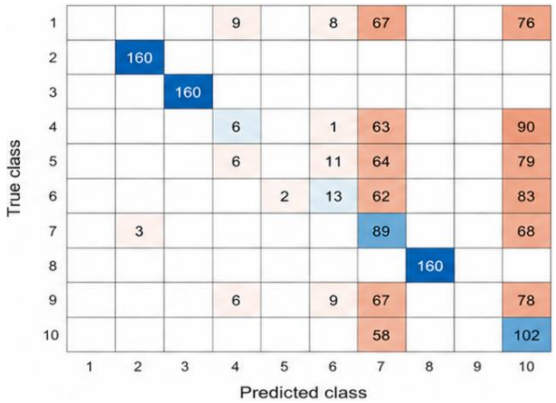


Figure 10: Confusion matrix of the Naive Bayes classifier.

Table 4: Simulation-based classification results of the four investigated methods.

Classifier type	Accuracy	Mean squared error	Macro-Precision	Macro-Recall	Macro-F1
SVM	86.5%	3.48	86.84%	86.5%	85.57%
Naive Bayes	43.125%	38.5	38.8%	43.13%	37.43%
PCA-SVM	97.93%	0.685	98.12%	97.94%	97.96%
PCA-Naive Bayes	83.2%	3.1425	85.64%	83.19%	83.46%

3.2 Fault Classification Results With PCA

To evaluate the effect of feature preprocessing, PCA is applied to the input dataset before classification. After PCA transformation, the reduced feature matrix is used to train and test the same classifiers. Figure 11 shows the confusion matrix obtained by the PCA-SVM method.

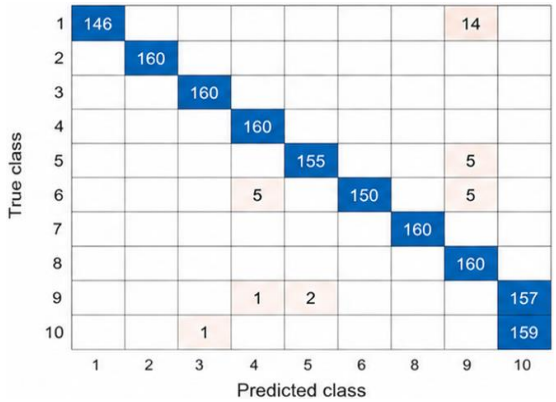


Figure 11: Confusion matrix of the PCA-SVM classifier.

PCA-SVM achieves the highest simulation-based accuracy of 97.93%. Compared with SVM without PCA, its accuracy increases by 11.43 percentage points, while its MSE decreases from 3.48 to 0.685. The substantial reduction in off-diagonal elements indicates that the PCA-transformed feature space enables the RBF-SVM classifier to construct more effective nonlinear decision boundaries among the ten operating conditions.

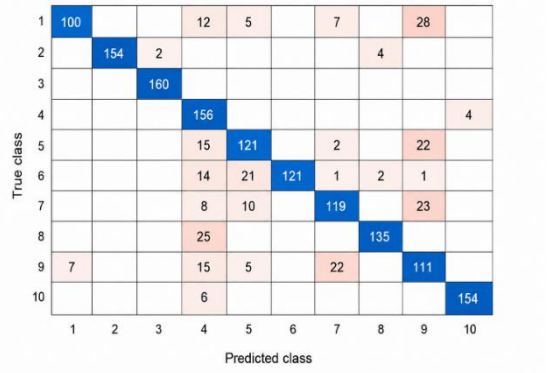


Figure 12: Confusion matrix of the PCA-Naive Bayes classifier.

PCA-Naive Bayes achieves an accuracy of 83.20%, representing an improvement of 40.075 percentage points compared with Naive Bayes in the original feature space. Its MSE also decreases from 38.50 to 3.1425. This result indicates that transforming the correlated original variables into orthogonal principal components

substantially reduces the adverse effect of the original feature dependencies. Nevertheless, PCA-Naive Bayes remains less accurate than PCA-SVM.

3.3 Comparative Analysis of the Simulation-Based Classification Methods

The numerical comparison of the four simulation-based methods is presented in Table 4. The table reports accuracy, mean squared error, Macro-Precision, Macro-Recall, and Macro-F1 for SVM, Naive Bayes, PCA-SVM, and PCA-Naive Bayes.

Table 4 shows that SVM performs considerably better than Naive Bayes when the original input variables are used. SVM achieves an accuracy of 86.50%, whereas Naive Bayes achieves only 43.125%. The corresponding MSE values are 3.48 and 38.50, respectively. These findings indicate that the nonlinear RBF decision boundaries of SVM are more suitable for separating the considered PV operating conditions than the probabilistic independence assumptions of Naive Bayes.

PCA substantially improves both classifiers in the ten-class simulation study. The accuracy of SVM increases from 86.50% to 97.93%, while its MSE decreases from 3.48 to 0.685. The accuracy of Naive Bayes increases from 43.125% to 83.20%, while its MSE decreases from 38.50 to 3.1425. The improvement obtained by PCA-SVM can be explained by the interaction between feature decorrelation and nonlinear margin-based classification. The original voltage, current, power, temperature, and irradiance variables contain correlated information. PCA transforms these variables into orthogonal components and concentrates most of the variance into a smaller feature space. The RBF-SVM classifier can subsequently construct more effective nonlinear decision boundaries using this transformed representation. PCA also considerably improves Naive Bayes because the transformation removes linear correlation among the principal components and partially mitigates the mismatch between the original correlated measurements and the classifier's conditional-independence assumption. However, uncorrelated variables are not necessarily statistically independent. Therefore, PCA-Naive Bayes remains less accurate than PCA-SVM.

Overall, PCA-SVM provides the highest accuracy, the highest F1-based performance, and the lowest MSE in the simulation-based ten-class study. However, this superiority is limited to the considered simulation dataset. The external validation in the following sections examines whether PCA provides similar benefits across different classifier families and measured PV data..

3.4 External Validation Using Measured PV Data

The external validation was conducted to evaluate whether the selected electrical and environmental features could distinguish clean and partial-shading conditions beyond the MATLAB/Simulink environment. The leave-one-panel-out procedure ensured that all observations belonging to the tested physical panel remained completely unseen during model training.

Table 5 presents the aggregated out-of-fold results of the twelve evaluated configurations. Random Forest achieves the highest accuracy of 97.53%, a Macro-F1 of 97.53%, and a Cohen's kappa coefficient of 0.951. Naive Bayes and LightGBM both achieve accuracies of 91.67%. Among the PCA-based benchmark models, PCA-LightGBM and PCA-XGBoost achieve the highest accuracies of 90.89% and 90.03%, respectively.

Figure 13 provides a visual comparison of the external-validation accuracies. The classifier ranking differs from that of the ten-class simulation study. PCA-SVM achieves the strongest result in the simulation analysis, whereas Random Forest provides the highest performance for the measured binary classification task. This difference indicates that model performance depends on dataset characteristics, module technology, the number of classes, and fault complexity.

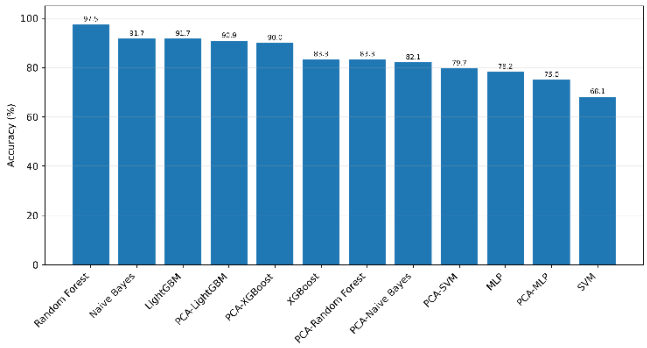


Figure 13: External-validation accuracy of the six classifier families with and without PCA under leave-one-panel-out validation.

Figure 14 presents the aggregated Random Forest confusion matrix. All 1,800 clean observations are correctly classified. Among the 1,800 partial-shading observations, 1,711 are correctly identified and 89 are assigned to the clean class. Therefore, Random Forest achieves a partial-shading recall of 95.06% without producing false partial-shading alarms for clean observations.

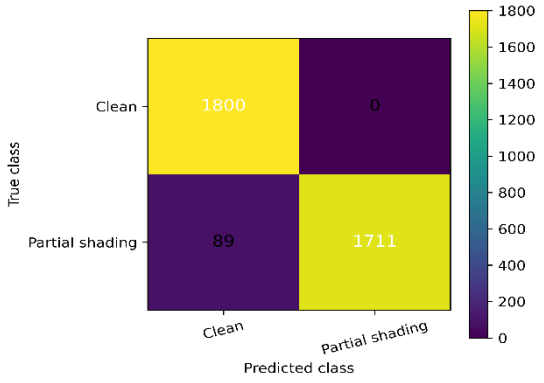


Figure 14:Aggregated leave-one-panel-out confusion matrix of the Random Forest classifier using measured clean and partial-shading PV data.

3.5 Effect of PCA Across Different Classifier Families

Table 6 presents a paired comparison of each classifier family before and after PCA. Identical classifier parameters and validation folds

Table 5: External-validation performance of the classifiers using measured clean and partial-shading PV data.

Classifier	Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	Weighted-F1 (%)	Kappa
Random Forest	**97.53**	**97.64**	**97.53**	**97.53**	**97.53**	**0.951**
Naive Bayes	91.67	92.86	91.67	91.61	91.61	0.833
LightGBM	91.67	92.86	91.67	91.61	91.61	0.833
PCA-LightGBM	90.89	92.12	90.89	90.82	90.82	0.818
PCA-XGBoost	90.03	91.67	90.03	89.93	89.93	0.801
XGBoost	83.33	83.33	83.33	83.33	83.33	0.667
PCA-Random Forest	83.33	87.50	83.33	82.86	82.86	0.667
PCA-Naive Bayes	82.08	86.48	82.08	81.53	81.53	0.642
PCA-SVM	79.69	79.85	79.69	79.67	79.67	0.594
MLP	78.25	78.58	78.25	78.19	78.19	0.565
PCA-MLP	75.00	75.00	75.00	75.00	75.00	0.500
SVM	68.14	68.15	68.14	68.13	68.13	0.363

were used within each pair. Therefore, the reported changes isolate the effect of PCA preprocessing.

Table 6: Effect of PCA on external-validation accuracy and panel-level statistical comparison.

Family	Without PCA (%)	With PCA (%)	Change (points)	(p)-value
SVM	68.14	79.69	**+11.56**	0.500
XGBoost	83.33	90.03	**+6.69**	0.500
LightGBM	91.67	90.89	−0.78	0.625
MLP	78.25	75.00	−3.25	0.750
Naive Bayes	91.67	82.08	−9.58	0.250
Random Forest	97.53	83.33	−14.19	0.500

PCA increases SVM accuracy from 68.14% to 79.69%, corresponding to an improvement of 11.56 percentage points. PCA also increases XGBoost accuracy from 83.33% to 90.03%, corresponding to an improvement of 6.69 percentage points. These results indicate that the orthogonal PCA representation can improve feature separability for the margin-based SVM and the boosting-based XGBoost classifier.

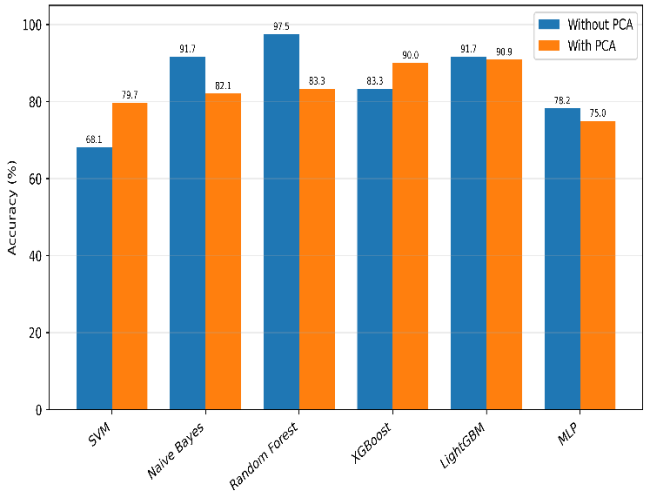


Figure 15: Paired comparison of classification accuracy before and after PCA for SVM, Naive Bayes, Random Forest, XGBoost, LightGBM, and MLP.

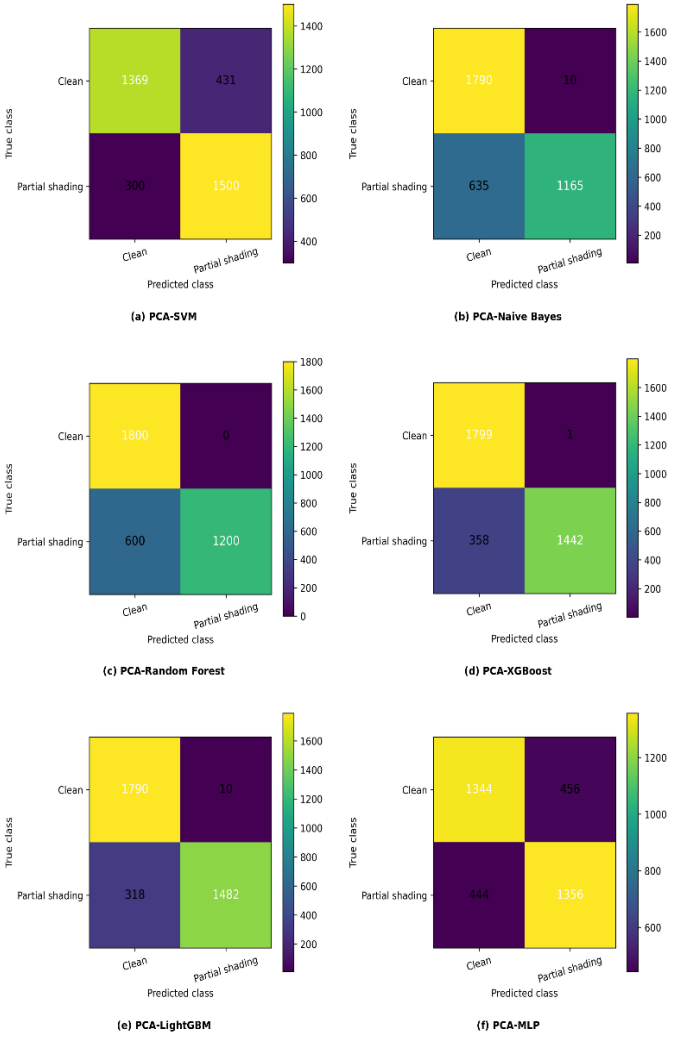


Figure 16: Aggregated leave-one-panel-out confusion matrices of the PCA-based classifiers using measured PV data: (a) PCA-SVM, (b) PCA-Naive Bayes, (c) PCA-Random Forest, (d) PCA-XGBoost, (e) PCA-LightGBM, and (f) PCA-MLP.

In contrast, PCA does not improve all classifier families. Random Forest accuracy decreases from 97.53% to 83.33%, while Naive Bayes accuracy decreases from 91.67% to 82.08%. Tree-based classifiers can directly exploit physically meaningful thresholds in voltage, current, power, temperature, and irradiance. PCA transforms these original variables into linear combinations and may therefore weaken useful axis-aligned splitting structures. Similarly, orthogonal principal components do not necessarily produce more suitable class-conditional distributions for Gaussian Naive Bayes. The MLP model also does not benefit from PCA because it can learn nonlinear internal representations directly from the standardized original variables. Figure 16 presents the confusion matrices of the six PCA-based models. The different error distributions confirm that the effect of PCA depends on the learning mechanism of the classifier. Therefore, PCA should not be regarded as an automatically beneficial preprocessing method for every classification model.

3.6 Statistical Significance Analysis

Exact paired sign-flip permutation tests were applied to the leave-one-panel-out accuracy values to examine whether the PCA-related performance changes were statistically supported. The six physical panels were treated as independent experimental units. The 300 operating points extracted from each measured I-V curve were not treated as independent statistical experiments because observations originating from the same curve are correlated. A point-level significance test would therefore produce artificially small p-values. PCA increases the average accuracies of SVM and XGBoost by 11.56 and 6.69 percentage points, respectively. However, none of the PCA/non-PCA comparisons reaches the significance threshold of $\alpha = 0.05$. The p-values are 0.500 for SVM, 0.500 for XGBoost, 0.625 for LightGBM, 0.750 for MLP, 0.250 for Naive Bayes, and 0.500 for Random Forest.

These results indicate that the observed differences represent descriptive performance changes on the available experimental panels. The limited number of six independent panels does not provide sufficient statistical power to establish universal superiority. A larger number of independently measured panels is required for stronger statistical conclusions.

3.7 Limitations of the Study

Although measured experimental data were included in the revised study, several limitations remain. First, the external dataset contains only six independent physical panels and covers only clean and partial-shading conditions. Therefore, the external analysis does not validate all electrical fault classes considered in the ten-class simulation study. Second, the measurements were obtained from individual PV modules under outdoor experimental conditions rather than from a utility-scale operational PV plant. The external results should therefore be interpreted as module-level experimental validation rather than complete field verification.

Third, the 300 observations extracted from each measured I-V curve are correlated and do not represent 300 independent physical experiments. The leave-one-panel-out procedure prevents panel-level information leakage, but the statistical power remains limited by the availability of only six independent panels. Finally, the different classifier rankings obtained for the simulation and measured datasets indicate that dataset shift, module technology, fault complexity, and environmental variability can influence model selection. Additional measured datasets containing open-circuit, short-circuit, bypass-diode, bridge, and series-resistance faults are required for comprehensive validation of the complete ten-class framework.

4 Conclusion

This study investigated machine learning-based fault detection and classification in photovoltaic arrays using both simulation-generated and measured experimental data. In the ten-class MATLAB/Simulink study, SVM achieved an accuracy of 86.50%, while Naive Bayes achieved 43.125%. PCA substantially improved both classifiers. PCA-SVM achieved the highest simulation-based accuracy of 97.93% and the lowest MSE of 0.685. PCA-Naive Bayes increased the accuracy from 43.125% to 83.20% and reduced the MSE from 38.50 to 3.1425. External validation was subsequently conducted using 3,600 observations extracted from twelve measured I-V curves of six physical PV panels under clean and partial-shading conditions. Random Forest achieved the highest external-validation accuracy of 97.53%. PCA increased SVM accuracy from 68.14% to 79.69% and XGBoost accuracy from 83.33% to 90.03%. However, PCA reduced the performance of several other classifier families. These results demonstrate that PCA can provide a more separable feature representation for some learning methods, but its effectiveness depends on the classifier architecture and data distribution. The measured-data results confirm the usefulness of the selected electrical and environmental variables for partial-shading diagnosis. Nevertheless, the external dataset contains only six independent panels and one fault type that directly overlaps with the simulation study. Therefore, the results should not be interpreted as comprehensive field validation of all ten operating conditions. Future work will extend the external validation to a larger number of physical panels and measured fault types, including open-circuit, short-circuit, bypass-diode, bridge, and series-resistance faults. Transfer-learning approaches will be investigated to adapt trained classifiers across different module technologies and operating sites. Online and incremental learning will also be considered to address module aging, seasonal variation, and changing environmental conditions. Furthermore, explainable artificial intelligence methods, such as SHAP and LIME, and low-cost edge-computing implementation will be investigated for transparent real-time PV condition monitoring.

5 Acknowledgement

Not applicable.

Disclosure of Potential Conflicts of Interest

The authors declare that there is no conflict of interest.

Reference

[1] International Renewable Energy Agency. Renewable Capacity Statistics 2026. Abu Dhabi: IRENA; 2026. Available: https://www.irena.org//media/Files/IRENA/Agency/Publication/2026/Mar/IRENA_DAT_RE_capacity_statistics_2026.pdf.

[2] Rahman MM and Srivastava AK. Detection, classification, and localization of faults and failures in photovoltaic arrays: a review, comparative analysis, and path forward. Discover Energy 2026;6:26. doi: 10.1007/s43937-026-00136-5.

[3] Madeti SR and Singh SN. A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. Solar Energy 2017;158:161–185. doi: 10.1016/j.solener.2017.08.069.

[4] Pillai DS and Rajasekar N. A comprehensive review on protection challenges and fault diagnosis in PV systems. Renewable and Sustainable Energy Reviews 2018;91:18–40. doi: 10.1016/j.rser.2018.03.082.

[5] Zbib B and Al Sheikh H. Fault detection and diagnosis of photovoltaic systems through I–V curve analysis. 2nd International Conference on Electrical, Communication, and Computer Engineering 2020:1–6. doi: 10.1109/ICECCE49384.2020.9179390.

[6] Madeti SR and Singh SN. Modeling of PV system based on experimental data for fault detection using kNN method. Solar Energy 2018;173:139–151. doi: 10.1016/j.solener.2018.07.038.

[7] Chen Z et al. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energy Conversion and Management 2018;178:250–264. doi: 10.1016/j.enconman.2018.10.040.

[8] Fadhel S et al. PV shading fault detection and classification based on I–V curve using principal component analysis: Application to isolated PV system. *Solar Energy* 2019;179:1–10. doi: 10.1016/j.solener.2018.12.048.

[9] Chen LC et al. Fault diagnosis and classification for photovoltaic arrays based on principal component analysis and support vector machine. IOP Conference Series: Earth and Environmental Science 2018;188:012089. doi: 10.1088/1755-1315/188/1/012089.

[10] Yi Z and Etemadi AH. A novel detection algorithm for line-to-line faults in photovoltaic arrays based on support vector machine. IEEE PES General Meeting 2016:1–5. doi: 10.1109/PESGM.2016.7742026.

[11] Vahidi A and Golkar MA. Fault detection and classification in PV arrays using machine learning algorithms in the presence of noisy data. 9th Iranian Conference on Renewable Energy and Distributed Generation 2022:1–7. doi: 10.1109/ICREDG54199.2022.9804510.

[12] Nedaei A, Eskandari A and Aghaei M. Photovoltaic fault detection and classification: Reconsideration of classic machine learning and dataset shrinkage techniques for simplification. *Results in Engineering* 2025;27:106356. doi: 10.1016/j.rineng.2025.106356.

[13] Hassan MS, Chin VJ and Gopal L. Accurate diagnosis of concurrent faults in photovoltaic systems using CONMI-based feature selection and support vector machines. Energy Conversion and Management 2025;344:120293. doi: 10.1016/j.enconman.2025.120293.

[14] Pillai PK, Devaraj D and Selvam S. Smart diagnostics of AI-powered IoT solutions for solar grid reliability. Discover Electronics 2025;2:72. doi: 10.1007/s44291-025-00113-7.

[15] Ali YM and Ding L. Fault diagnosis in photovoltaic arrays: A robust and efficient approach using feature engineering and 1D-CNN. International Journal of Electrical Power & Energy Systems 2025;170:110863. doi: 10.1016/j.ijepes.2025.110863.

[16] Sabati A, Bayindir R and Rahebi J. Photovoltaic panels fault detection with convolutional neural network and Bitterling Fish Optimization algorithm. International Journal of Computational Intelligence Systems 2025;18:239. doi: 10.1007/s44196-025-00984-4.

[17] Navid Q, Xia Y, Ali A, Hassan A, Ali A, Manzoor T and Ali S. Fault diagnosis framework for photovoltaic system using fully connected neural network-based back surface thermal analysis. Solar Energy 2026;303:114139. doi: 10.1016/j.solener.2025.114139.

[18] Seghiour A et al. Fault detection and diagnosis in photovoltaic systems using artificial intelligence and time-frequency analysis. Scientific Reports 2026;16:10056. doi: 10.1038/s41598-026-39386-7.

[19] Talaat FM, Salem M and Shaban WM. AI-driven fault detection and classification in photovoltaic systems using deep learning techniques. Scientific Reports 2026;16:8727. doi: 10.1038/s41598-026-40246-7.

[20] Vahidi A, Toulabi M, Pourmoosavi MA and Amraee T. Semi-supervised fault detection and classification in photovoltaic arrays using a modified generative adversarial networks architecture. Energy Reports 2026;16:109516. doi: 10.1016/j.egyr.2026.109516.

[21] Vahidi A, Taghizadeh A and Toulabi M. Fault detection and classification in induction motors: An explainable convolutional neural networks approach. 2025 5th International Conference on Electrical Machines and Drives 2025. doi: 10.1109/ICEMD68435.2025.11441724.

[22] Vahidi A, Nikoofard A and Toulabi M. Data-driven incomplete information Stackelberg game approach for economic load dispatch in power grid. Electric Power Systems Research 2026;259:113024. doi: 10.1016/j.epsr.2026.113024.

[23] Restrepo-Cuestas BJ, Guarnizo-Lemus C, Montoya-Marín JA and Montano J. Dataset of photovoltaic panel performance under different fault conditions: cracks, discoloration, and shading effects. Data in Brief 2025;59:111392. doi: 10.1016/j.dib.2025.111392.

[24] Restrepo-Cuestas B, Guarnizo-Lemus C, Rojas-Montano JJ, Montoya-Marin JA and Ibañez-Diaz AA. Experimental data for mismatching faults of photovoltaic modules. Mendeley Data, Version 2; 2024. doi: 10.17632/xjs42j8dtf.2.