EAI Endorsed Transactions
on Energy Web

Research Article **EAI.EU**

Research on Fault Diagnosis Method for Photovoltaic Array Based on XGBoost Algorithm

Zongyu Zhang[1*], Bodi Liu[2], Chun Xie[3] and Ermei Yan[4]

[1,2,3,4]Guizhou Power Grid Co., Ltd. Intelligent Operation Center, Guiyang, Guizhou, 551417, China

Abstract

INTRODUCTION: Photovoltaic (PV) energy sources frequently experience issues, including fragmentation, open-circuit, short-circuiting, and other common and hazardous problems. The current focus of PV research is on fault detection within solar arrays. Traditional models encounter challenges in identifying errors due to uncertainties in panel settings and the complex nature of the actual PV structure.
OBJECTIVES: This study aims to introduce a novel Extreme Gradient Boosting (XGBoost) approach for fault diagnosis in PV arrays.
METHODS: The XGBoost algorithm is trained using collected PV array defect data samples. Data preprocessing is performed to manage missing values and remove noisy data. Feature extraction is conducted using Linear Discriminant Analysis (LDA) to improve detection accuracy. To further enhance XGBoost's performance, the World Cup Optimization (WCO) approach is applied to select optimal features from the extracted data. Fault detection is then conducted using the XGBoost algorithm on the processed data. Various indicators are utilized for performance assessment within the Python environment.
RESULTS: The comparative analysis demonstrates that this research improves fault detection efficiency in PV arrays compared to existing methodologies.
CONCLUSION: The study presents an effective method for enhancing fault detection in PV systems, showcasing the advantages of the XGBoost and WCO-based approach over conventional methods.

Keywords: Faults, Detection, PV arrays, Linear discriminant analysis (LDA), World cup optimization (WCO), Extreme gradient boosting (XGBoost)

Received on 09 09 2024, accepted on 01 10 2024, published on 19 11 2024

doi: 10.4108/ew.7224

1. Introduction

Interest in solar energy as a sustainable power source has grown as conventional fossil resources are becoming depleted. As a vital component of solar power era buildings, photovoltaic (PV) arrays are extremely susceptible to damage due to the placement of their components in hazardous exterior environments [1]. PV array fault detection is an essential safety measure for providing great, commercially viable PV energy technology. Due to its many benefits, including consistent overall performance, long term viability zero emissions, and adaptability, solar PV generating is now widely employed in the renewable power sector [2]. The need to guarantee the uninterrupted functioning of PV systems is growing in conjunction with their installed capacity. One major operational issue with these systems is effectively identifying and managing crises. Fault detection is now crucial for raising revenue and preventive expensive maintenance, rather than only safeguarding the dependability and security of PV power plants [3].

Identifying disorders effectively is a significant approach, especially when it comes to PV systems DC side. There is now a significant void in the standard administration and

[*] Corresponding author. Email: 13984101521@163.com

protection of PV structures due to the incapacity of traditional protective devices to accurately identify and categorize problems [4]. Because PV arrays operate in a range of environmental conditions, they can eventually experience physical damage or degeneration. Defects such as short circuits or grounding problems can lead to electrical shocks, fries, or heating if they are not discovered and corrected promptly. Fault analysis increases the safety of consumers and nearby workers by identifying potential threats before they escalate. Solar PV structures require efficient failure detection and active repair. The probability of a fireplace increases during power outages, and ineffective fault detection can lead to several safety concerns. PV structures could face several issues [6]. Thus, before developing surveillance and fault identification algorithms, it is crucial to understand the failures observed in real world situations.

The aging process has several reasons why PV systems malfunction. The essential parts that transform the DC generated by solar panels into alternating current (AC) suitable for use in houses or on the grid are inverters. Factors such as thermal stress, aged components, or manufacturing errors can cause inverters to fail, which can lead to diminished performance or system breakdown as a whole [7]. The three main classifications of PV system defects are incipient, sudden, and intermittent based on the features of their time course. Some defects are transient or irregular and show up gradually, such as partial shadows and environmental stress [8]. Problems including disconnector loneliness, junction box failures, failures of open circuits and short circuits from line to line or line to ground that occur abruptly and repeatedly as a consequence of damage to the PV array are referred as "permanent" or "abrupt" flaws [9].

Hot spots are regarded as irreversible flaws. Since these are so minute and have a tendency to alter gradually over time, early detection of these flaws is thought to be the most difficult [10]. These PV array problems include open circuits, short circuits, hot spots, shadows, and aging. These PV array flaws have the potential to lower efficiency, shorten service life, and create a fire danger [11]. The solar array's nonlinear properties and the inverter's Maximum Power Point Tracking technologies prevent traditional safety systems from tripping under certain faults, hence decreasing system efficiency and raising fire hazard threats [12].

1.1 Aim of this study

Variations in weather, shading effects, and aging-related deterioration could impact on PV systems and cause weak or transient fault symptoms. Sophisticated methods or more sensors could be needed to increase fault detection accuracy and reliability as traditional diagnostic techniques can have trouble differentiating between real defects and typical operating changes. This study proposes a new XGBoost technique for fault detection in solar arrays to address common and serious PV issues.

1.2 Key contribution of this study

- Datasets were gathered from PV array defect data.
- The purpose of pre-processing data is to deal with missing values or eliminate noisy data.
- LDA for defect diagnosis in PV arrays effectively extracts discriminative features, enhancing classification accuracy.
- WCO provides accurate and efficient maintenance solutions by optimizing feature selection for accurate fault diagnosis in PV arrays.
- Photovoltaic array problems could be effectively identified with the XGBoost approach.

1.3 Structure of this paper

The remaining study aspects could be categorized as follows: In section 2, we will discuss the related work. Section 3 discusses the approaches. Section 4 presents the outcomes of the experiment. The discussion is presented in Section 5, and the conclusion is covered in Section 6.

2. Related works

An approach to failure diagnostics based on the PV array output deviation characteristics was presented in the research [13]. The analysis of the PV array output deviation characteristics under various array configurations and time series was done which is based on the PV array's current on the direct current (DC) side.It significantly contributes by proposing a deviation function that could be used to extract fault characteristics from PV arrays and by providing an easy-to-apply fault detection approach that used array current in PV plants.

An approach to defect diagnostics for solar arrays using a clustering technique and an array auto-encoder was provided in study [14]. The method was capable of extracting features and mining data collection features for fault detection using a limited amount of labeled data samples. The membership function was utilized for defect diagnostics, and the clustering procedure yields clusters and clustering centers. Moreover, the effectiveness of the suggested fault diagnostic approach was confirmed using both simulation and experimental information.

The current drop across two maximum power point tracking (MPPT) sampled instants was used as the basis for a sensor less detection approach that was suggested in study [15]. Simulations were used to confirm its ability to identify problems in a range of conditions, independent of misalignment and irradiance levels. Different voltages could be modified to distinguish between normal and faulty filament currents to easily discover problems.

An Adaptive Neuro-Fuzzy Inference System (ANFIS) approach-based intelligent PV fault identification system was demonstrated in research [16]. The Grid Partitioning (GP) and Subtracted Clustering (SC) algorithms were important for an effective PV fault identity and categorization system that could be implemented utilizing certain study data to train the ANFIS model. When it came to predicting and categorizing different PV system problems, the accuracy of the ANFIS SC methodology turned into better than that of the ANFIS GP technique

To identify line-line (LL) and open circuit (OC) defects in PV systems, research [17] provided an effective and automated fault detection method by figuring out and choosing traits for the duration of the training phase using an ensemble learning algorithm and a smaller dataset. The nonlinear nature of PV features, their dependence on the operating environment, and the MPPT methods could render it hard for traditional shielding devices to evaluate failures in PV systems. The recommended approach started by determining the PV arrays fundamental voltage and current operating parameters.

Through the use of semi-supervised graph signal processing, study [18] tackled the issue of defect classification in PV arrays. Extensive labelled data sets were needed for training in traditional fault detection and classification systems. Obtaining associated data for various failure classes in utility-scale solar arrays required a lot of resources. The suggested approach generally produced strong classification results. They showed notable improvements over previous approaches and test their methodology using a real-time dataset.

Research [19] suggested a flaw diagnosis system based on a genetic algorithm optimizing deep belief network (GA-DBN). The network's effectiveness and rate of convergence were raised by modifying the weight and bias of the network using the GA approach. The fitness function was built using the restricted Boltzmann machine recovering error. It showed how the suggested approach increases the detection accuracy of solar array faults and enhances the generalization capacity of conventional DBN.

The effectiveness of ML models appropriate for module-level embedding in inexpensive hardware was assessed in study [20] by discussing eight distinct PV module defects and how they affect PV module output. The flaws were replicated and applied to a real solar system, allowing measurement data to be collected and annotated at the panel level. A number of ML approaches were used to classify the mistakes with respect to the standard state.

In an attempt to develop a suitable ML framework, article [21] intended to automatically identify and diagnose common PV array problems. PV systems were becoming more and more common in contemporary electrical grids, which have raised concerns about how to diagnose faults in PV arrays. Furthermore, the optimization hyperparameters of the fault classifiers were selected using Bayesian optimization. To assess the performance of the described classifiers, simulated and experimental scenarios were performed.

A mechanism to use current-voltage characteristics, orI-V curves for PV defect diagnostics was established in study [22]. PV modules' I-V curves reveal a wealth of data regarding their condition. Diagnoses were made using a subset of the data from the I-V curve. Then the fault characteristics could be acquired directly from the resample current vectors, through a Gramian angular difference region, or through the recurrent plot. Along with measurement errors and background noise, it also included resilience. Three aspects of classifier analysis were feature dimension reduction, disturbance resilience, and transformation effect.

An intelligent fault detection model was presented in study [23] to identify and classify defects in PV systems. Throughout the experimental validation, a number of fault state and average state data were gathered over the winter in a variety of weather conditions. To efficiently identify defects in solar systems, it was vital to understand how the current/voltage (I/V) characteristics respond under different environmental settings. In some challenging circumstances of a PV system, especially in the winter, I/V characteristics were almost identical to those of normal states.

Developing a precise method for identifying errors in PV systems was the primary objective of study [24]. The two main stages of the conventional failure detection and diagnosis (FDD) method were feature extraction and fault diagnosis. The PV arrays maximum electricity point trackers, dependency on isolation efficiency, fault magnitudes, and non linear PV characteristics have made FDD more difficult. Multiple metrics and data sets gathered from different grid connected PV (GCPV) tool operating scenarios were used to evaluate the FDD performance.

Study [25] suggested a flawed diagnosis method for PV systems that relied on the application of ensemble learning (EL) technology. By aggregating the predictions of many ML systems, El approaches beat single ML algorithms in terms of quality and applicability prediction. By utilizing the appropriate features and ideal parameters for each unique learning technique and the EL model, the proposed strategy improved classification performance but also demonstrated outstanding generalization ability for identifying PV system defects.

2.1. Problem statement

The timely and effective diagnosis of issues with PV arrays is essential to maximizing solar energy's performance and durability. Current fault-detecting approaches are not scalable, inaccurate, or efficient enough to ensure the overall dependability and maintenance effectiveness of PV systems. The difficulty is in creating a reliable fault diagnostic technique that can use data from sensors or monitoring systems to precisely identify and indicate several kinds of defects including shading, module failures and wiring problems in real-time or almost real-time. This approach should be flexible to accommodate varying PV array designs and environmental factors while guaranteeing rapid action to prevent operational losses and performance deterioration.

3. Methodology

We collect defect data samples for PV arrays. The purpose of pre-processing data is to deal with missing values or eliminate noisy data. Feature extraction is done using the LDA to enhance detection performance. To improve XGBoost's efficacy, the WCO approach is applied to produce the best features from the recovered data. Then we utilize the XGBoost method to detect issues with the PV arrays. Figure 1 shows the overall flow.

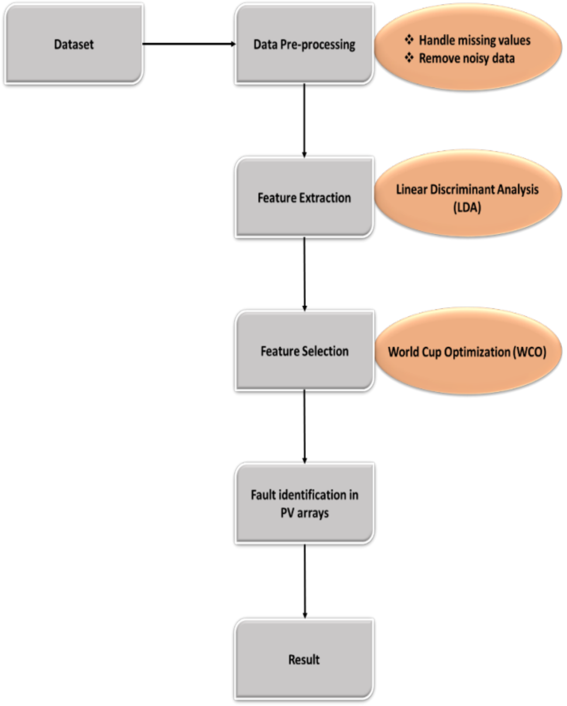


Figure 1. Overall flow

3.1. Data Collection

The 2×6 PV array's DC characteristics are detected to gather the fault data set. It is possible to ascertain using the $I - V$ curve detector of $IVT - 30 - 1000$ (IVT stands for IV Tester or IV Tracer, 30 indicates specific features, and 1000 denotes maximum current capacity), the DC characteristics of the 2×6 PV array. Additionally, the $IVT - 30 - 1000$ detector is capable of detecting the PV array's maximum power voltage (Vmp), open circuit (OC) current, maximum power current (Imp), and short circuit (SC) current.

The OC, SC, and lack of irradiance faults could be simulated using a 2×6 PV array, and a PV array defect data set can be gathered. Additionally, there are 325 fault data used to evaluate the fault diagnostic technique in addition to the 550 fault data used to train the fault diagnosis model. Moreover, the percentage of every fault data used to train the fault diagnostic model.

3.2. Data preprocessing

Preprocessing data with missing values or eliminating noisy data is essential for precise analysis and efficient maintenance in solar array defect diagnostic techniques. Imputation procedures or exclusion are required when there are missing values, as they could distort findings and make more difficult to identify errors. In similar, noisy data that is data including outliers or sensor errors can wrap trends, cause false alarms, or fail to detect real problems.

3.2.1. Identifying missing values

The dataset contains defect data from devices such as the IVT-30-1000 and features of a 2x6 PV array, therefore faults in measurement or equipment operation could result in missing outcomes.

Detection: Locate entries or columns with missing values. Thus, there could be missing values for characteristics like open circuit current, short circuit current, maximum power voltage, and maximum power current.

Handling: Choose methods to deal with information that is lacking. If there is a large amount of missing data that cannot be properly imputed, options include removing the rows and columns or imputation (changing missing values with approximated ones, such as mean or median).

3.2.2. Noise removal

There are several potential causes of noise in data, including errors in measurements, defective sensors, and external influences on the instruments.

3.3. Feature extraction

PV array defect diagnostics use LDA to extract discriminative features from the data. By using this technique, LDA intends to identify a feature set that maximizes the separability between various fault or condition categories inside the PV array. LDA determines the directions which distinguish different faults by projecting the data into a lower dimensional space.

3.3.1. LDA

A supervised method of reducing dimensionality is LDA. LDA works on the basis of projecting high-dimensional data into low-dimensional space. In the predicted space, similar categories will be nearer to one another, whereas data from other classes will be further away.

$Z = [z_1, z_2, .., z_j, ..]$ is the fault-related feature matrix, where z_j is the feature vectors of the j^{th} sample. Considering the mean vector of the i^{th}classsamples $\mu_j(i = 1,2, .., l)$ is defined as follows in equation (1).

$$\mu_i = \frac{1}{M_i} \sum_{w \in W_i} w\,, \quad i = 1,2, \dots, l \tag{1}$$

WhereM_i indicates thequantity of samples from the i^{th} class, W_i represents the set of samples from the i^{th} sample, and l is the number of sample classes in the training set.

The class samples' correlation matrix, orQ_i, is defined as follows in equation (2).

$$Q_i = \sum_{w \in W_i} (w - \mu_i)(w - \mu_i)^S, \quad i = 1,2, \dots l \tag{2}$$

Let us assume that the space with low dimensions projected by LDA has dimension r, and that the appropriate basis vector is$w_1, w_2 \dots., w_r$. X is a $n \times r$ matrix made up of the basis vectors. The optimization desire for the objective of the greatest between-class distance and the shortest within-class distance following LDA projection is as follows in equation (3-5).

$$\arg\max_x \left(\frac{X^S T_a X}{X^S T_x X} \right) \tag{3}$$

Where, the within-class scatter matrix is indicated byT_x.

$$T_x = \sum_{i=1}^{l} \sum_{w \in w_i} \quad (w - \mu_i)(w - \mu_i)^S \tag{4}$$

The between-class scatter matrix is shown by T_a.

$$T_a = \sum_{i=1}^{l} M_i (\mu_i - \mu)(\mu_i - \mu)^S \tag{5}$$

Where μ is the average vector.

An objective function for LDA optimization is defined as follows since the function under consideration is not scalar and could not be optimized as a scalar function (see equation (6)).

$$\arg\max_x I(X) = \frac{\prod_{diag} X^S T_a X}{\prod_{diag} X^S T_x X} \tag{6}$$

Where the product of A's primary diagonal components is indicated by the symbol $\prod_{diag} A$. The $I(X)$ optimization procedure could be converted into equation (7).

$$I(X) = \frac{\prod_{j=1}^{r} X_j^S T_a X_j}{\prod_{j=1}^{r} X_j^S T_x X_j} = \prod_{j=1}^{r} \frac{X_j^S T_a X_j}{X_j^S T_x X_j} \tag{7}$$

The eigenvectors $x_1, x_2, \dots x_r$ that correspond to the r biggest eigenvalues thus represent the projection matrix X. Consequently, the feature matrix following projection could be expressed as equation (8).

$$\hat{Z} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} [z_1 z_2 \dots.. z_j] = X^S Z \tag{8}$$

3.4. Feature selection

PV array defect diagnostic techniques are using WCO, a feature selection technique inspired by the competitive dynamics of soccer organizations. By carefully choosing the most pertinent characteristics, this novel method intends to improve the exactness and efficacy of problem detection in PV systems.

3.4.1. World Cup Optimization (WCO)

The World Cup Optimization Algorithm (WCO) draws inspiration from the global battle between teams competing for the World Cup title. Based on their Ranks, teams are divided into many seeds. Teams' previous performances on prior courses are used to determine their rank. Teams are arranged according to their positions in WCO algorithms, with the stronger teams going to the first seed, the weaker teams going to the next seed, and so on. In the first era, stronger teams persevere through the struggle to save and elevate themselves to a superior status. The challenge begins after the sowing phase.

Usually, a preparatory task is given to teams in small groups all over the world to begin the competition. After the preliminary contests, the best two teams from each group move on to the next phase, with the worst teams being eliminated. Before the global competition begins, teams in small groups are usually given an initial task. Following the preliminary matches, each group's top two teams advance to the next round, while the worst two teams are eliminated.

The first stage is to choose the teams and continents.

Teams from each of the continents currently taking part in the challenges are represented by an array of $1 \times M_{var}$in a M variable dimensional (M_{var}) optimization challenge with N number of continents. The following is an example of this array (equation (9 and 10)).

$$Continent = [Contry_1, Contry_2, \dots, Contry_{M_{var}}] \tag{9}$$

$$Contry_j = [w_1, w_2, \dots, w_{M_{var}}] \tag{10}$$

Where, w_i represent the country's i^{th} team. The values $(w_1, w_2, \dots, w_{M_{var}})$of the variable are floating point numbers. An approach to determine the rank of the countries is to value the rank function e_q at a continent of$(w_1, w_2, \dots, w_{M_{var}})$ (equation (11 and 12)).

$$Rank = e_q(continent) = e_q(w_1, w_2, \dots, w_{P_{var}}) \tag{11}$$

$$P = M \times N \tag{12}$$

Where, N is the number of continents and M is the dimensions of the variables. An interval is chosen, given random values, and divided into sections that constitute forward the continents to speed up the convergence process.

Cost-function Analysis

This stage involves calculating the geographical locations on the continent. The points attained could not be entirely evident since there could exist a continent where certain teams have the highest point total (optimum fitness) while other teams have weak points. As a result, this method additionally takes consideration of the continents' mean value and standard deviation (equations (13 and 14)).

$$\overline{W} = \frac{1}{m} \sum_{j=1}^{m} W_j \tag{13}$$

$$\sigma = \sqrt{\frac{1}{m-1} \sum_{j=1}^{m} \left(W_j - \overline{W}\right)^2} \tag{14}$$

Where t represents the continent $W's$ standard deviation, m characterizes its members, and W displays the continent $W's$ mean value.

Ranking

Teams are rated using the following equation (15 and 16).

$$W_5 = [W_{51}, .. W_{5m}]^S \qquad (15)$$

$$W_{Total} = [W_{11}, .. W_{1m}, W_{21}, .. W_{2m}, W_{51}, .. W_{5m}]^S \qquad (16)$$

Where, S is the transposition operator and m is the quantity of teams on every continent. After this procedure, the two best values from each continent were chosen and combined into a new vector (W_{Rank}) for the upcoming championship contests. The best values from W_{Total} determine who will win the first cup (see equation 17 and 18).

$$(W_{Rank}) = [W_{11}, W_{12}, W_{21}, W_{22}, \ldots. W_{51}, .. W_{52}]^S \qquad (17)$$

$$W_{champion} = min(W_{Total}) = min([W_{11}, .. W_{1m}, W_{21}, .. W_{2m}, W_{51}, .. W_{5m}]^S) \qquad (18)$$

Where the minimal value of the responses is described by $W_{champion}$

Repeat the team competition phase

After the initial team tournament, new continents and the teams that compete in them will be created again using the results of previous championship games and the team rankings. Two-part vectors can be used in this situation as follows in equation (19).

$$Pop = W_{Total} = [W_{Bet}, W_{rand}] \qquad (19)$$

Where the new population of size $(M \times N)$ is described by function Pop W_{Total}, W_{rand} in this case indicates a random value among the intervals of issue constraints, and W_{Bet} is an array according to equation (20-22).

$$K < W_{Bet} < V \qquad (20)$$

$$V = \frac{1}{2} \times bd \times (\mathrm{Ub} + \mathrm{Lb}) \qquad (21)$$

$$K = \frac{1}{2} \times bd \times (\mathrm{Ub} - \mathrm{Lb}) \qquad (22)$$

Where, the coefficient bd lies between the problem's low and high bounds, Lb and Ub.

Exploration and exploitation

W_{Best} is the process of looking at areas of a search space that are close to the sites that were visited first, and comprehensive looking through fresh areas of a search space is referred as W_{Rand}. Cross Point (CP) divides the sizes of W_{Best} and W_{Rand} is follows in equations (23, and 24). End the algorithm if the criteria are met if not, repeat the entire process.

$$W_{Rand} = pop(1{:}\,CP, N) \qquad (23)$$

$$W_{Best} = pop(CP + 1{:}\,M, N) \qquad (24)$$

3.5 Fault Diagnosis Method for PV Array based on XGBoost algorithm

ML is used in the XGBoost algorithm-based fault diagnostic technique for PV arrays to effectively detect and categorize defects in PV systems. XGBoost is well-known ensemble learning and gradient boosting method that improves accuracy in fault condition prediction by combining several decision trees, including shading, module degradation, and wiring problems. Through the use of data-driven insights, this technique enhances defect identification, allowing for proactive maintenance and PV array performance enhancement.

3.5.1. XGBoost

Among the most effective methods for supervised learning are gradient boosting machines (GBM). Xgboost is one of its uses. It is applicable to problems with regression as well as classification. Because of Xgboost's exceptional out-of-core compute execution speed (equation (25)).

$$\dot{B}_{\cdot j} = \phi(w_j) = \textstyle\sum_{l=1}^{L} e_k(w_j), e_k \in \mathcal{F} \qquad (25)$$

To solve the above equation, we must minimize both the loss and the regularization intended to identify the optimal collection of functions (equation (26)).

$$\mathcal{L}(\phi) = \textstyle\sum_j l(z_j, \dot{B}_{\cdot j}) + \sum_l \Omega(e_k) \qquad (26)$$

The disparity between the expected and actual results z_j, is represented by the symbol l, which stands for the loss function. While Ω is a measure of the model's complexity, this prevents the model from being over fit and the following equation (27) is used to compute it.

$$\Omega(e_k) = \gamma S + \frac{1}{2}\lambda||x||^2 \qquad (27)$$

The number of leaves on the tree is represented by S in the equation above, and the weight of each leaf is represented by$x$.

Boosting, which involves adding a new function e while the model continues to train, which is used in decision trees to lower the desired function. As a result, the following new function (tree) becomes available on the s^{th} iteration (equation (28-31)).

$$\mathcal{L}^{(s)} = \textstyle\sum_{j=1}^{m} 1\left(z_j B_{\cdot j}^{(s-1)} + e_s(w_j)\right) + \Omega(e_s) \qquad (28)$$

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{j\in J_K} h_j)^2}{\sum_{j\in J_K} h_j + \lambda} + \frac{(\sum_{j\in J_Q} h_j)^2}{\sum_{j\in J_Q} h_j + \lambda} - \frac{(\sum_{j\in J} h_j)^2}{\sum_{j\in J} h_j + \lambda}\right] - \gamma \qquad (29)$$

$$h_j = \partial_{\dot{B}.^{(s-1)}} J(z_j, \dot{B}.^{(s-1)} \qquad (30)$$

$$g_j = \partial^2_{\dot{B}^{(s-1)}} 1(z_j, \dot{B}.^{(s-1)}) \qquad (31)$$

4. Experimental results

4.1. Experimental Configuration

The recommended approach is implemented on a Windows 10 laptop with an Intel i7 core CPU and 8GB RAM using Python 3.10.1. Utilizing the training data, we trained our suggested model with Tensor Flow/Keras or Scikit-Learn.

4.2. Performance Metrics

4.2.1. Outcome of proposed method

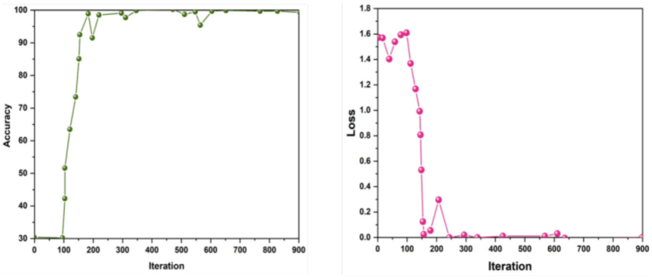
Accuracy

It is a reference to the accuracy with the fault diagnosis method locates and classifies malfunctions in the PV array system. It evaluates how well the approach can differentiate between various fault circumstances and normal operation.

Loss

The term "loss" usually describes the discrepancy between the solar array's actual output and its expected or ideal output under ordinary operating circumstances. It measures the reduction of performance based on errors like shading, malfunctioning modules, or bad connections.

Figures 2a and 2b show the accuracy and loss graphs for defect diagnosis in solar arrays, showing the efficiency of the diagnostic model across the training period.



(a) (b)

Figure 2. (a) accuracy (b) loss

4.2.2. Comparison Phase

Support vector machine (SVM) is an existing approach, whereas XGBoost is the method that is suggested [26]. The performance metrics are trained accuracy, rest accuracy, and difference.

The model can more precisely detect and categorize various kinds of defects inside the PV array due to these enhanced parameters. Then the model for fault diagnostics is trained using the preprocessed fault data, which has undergone through processes to manage missing values and eliminate noisy data. The model will be trained on high-quality data as a result, enabling to choose from the traits and patterns connected to every kind of failure.

Figure 3 and Table 1 provide a one-to-one method to show the accuracy of the classification methods. It shows the effectiveness of several parameter combinations in defect diagnostics. Thus, the combined accuracy rate of Vmp and Imp is 95%, indicating that these characteristics provide a strong basis for fault classification. In contrast, combined Vmp and OC voltage yield the highest accuracy of 96%, suggesting that both features are highly predictive of certain fault conditions. Vmp and SC voltage yields an accuracy of 91.5%. Other combinations, such as Imp with OC voltage and Imp with SC current, which have slightly lower accuracies of 93% and 91%, respectively, show that the model can classify problems. The high accuracy of 96% that is attained when the OC and SC parameters are combined, further demonstrates the model's robustness when using these specific parameter pairs for fault detection.

Table 1. Values for the accuracy of every one-to-one technique classification model

Number	Type of mixing	Accuracy (%)
1	Vmp & Imp	95
2	Imp & Open circuit	93
3	Imp & Short circuit	91
4	Vmp & Open circuit	96
5	Vmp & Short circuit	91.5
6	Open circuit & Short circuit	96

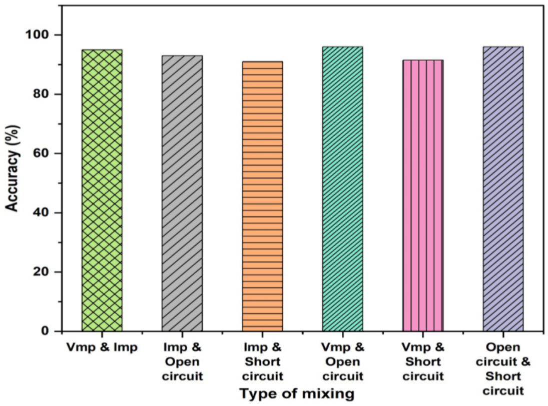


Figure 3. Accuracy for each one-to-one method performance

- Trained accuracy

The term “trained accuracy” describes a diagnostic model created using ML algorithms detects and categorizes problems following its training on a particular dataset in the context of a solar panel fault diagnostics technique. It provides an indication of the model learns and generalizes from the provided data to precisely identify problems in PV systems by quantifying the percentage of properly detected faults in relation to the overall quantity of error occurrences in the training set.
Table 2 and Figure 4 show the trained accuracy performance. The proposed method is XGBoost, it achieves 98.9 % when compared to the existing method SVM which achieves 98.72%. This shows that when it comes to detecting and differentiating fault states in solar arrays, XGBoost is slightly more successful than SVM.

Table 2. Values for R_m, R_t and $R_m - R_t$

Algorithm type	Trained accuracy (R_m) (%)	Rest accuracy (R_t) (%)	Difference (R_m- R_t)
R_{svm}	98.72	97	1.72
$R_{XGBoost}$- R_{svm}	0.18	0.5	0.32
$R_{XGBoost}$ [Proposed]	98.9	97.5	1.4

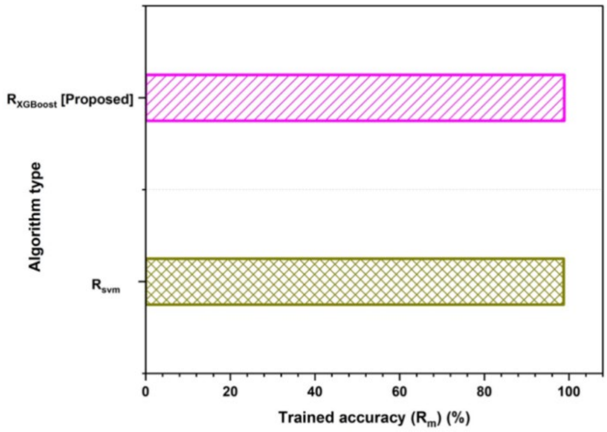


Figure 4. Trained accuracy performance

- Rest accuracy

"Rest accuracy" describes how well the system detects times when the solar array is running without any defects in the context of fault diagnostics. The metric quantifies the percentage of genuine negative situations, or fault-free, normal conditions, that the diagnostic technique accurately recognized among all the fault-free occurrences.
The rest accuracy performance is displayed in Figure 5 and Table 2. In comparison to the existing SVM approach, which achieves 97%, the suggested method, XGBoost, achieves 97.5%. This indicates that XGBoost outperforms SVM in terms of failure condition detection and differentiation in PV arrays.

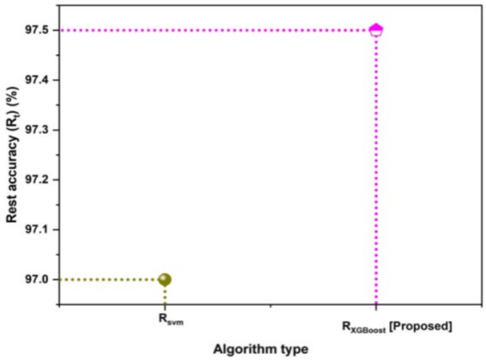


Figure 5. Rest accuracy performance

- Difference

Figure 6 and Table 2 show the difference of $R_m - R_t$. The proposed XGBoost achieves 1.4 and the existing method SVM achieves 1.72. XGBoost performs that much better than SVM. This indicates XGBoost works to increase the accuracy of defect identification when compared to the conventional SVM method.

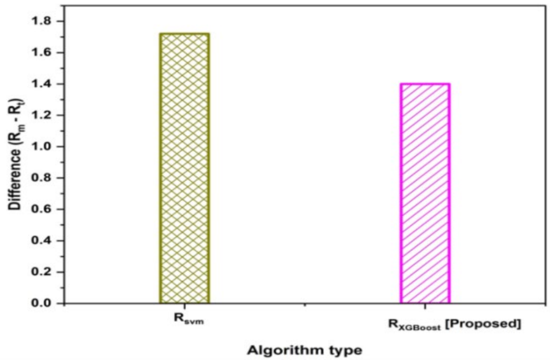


Figure 6. $R_m - R_t$ difference performance

Figure 7 and Table 2 show the difference $between\ R_{XGBoos\,t} - R_{SVM}$. The trained accuracy achieves 0.18, the rest accuracy achieves 0.5 and the difference achieves 0.32.

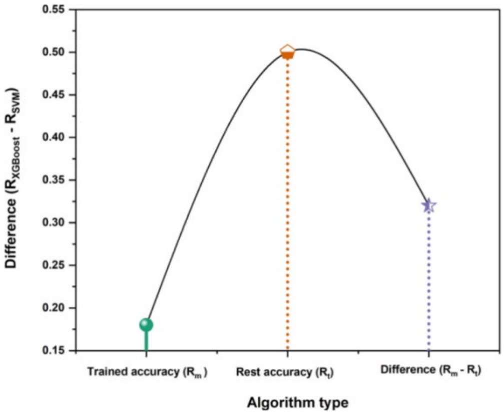


Figure 7. $R_{XGBoost} - R_{SVM}$ difference performance

5. Discussion

The existing method is SVM.Working with large datasets could make SVMs computationally and memory intensive, which is a significant challenge. SVMs could have trouble processing datasets that are very noisy or that cannot be linearly separated without careful kernel selection and modification, which could be challenging and domain-specific. In addition, the choice of kernel functions and parameters for SVM is critical and requires expertise since it influences the robustness and generalizability of the models. The suggested approach is XGBoost. When used for defect detection in solar arrays. Initially, since it specializes in handling intricate, non-linear correlations among input features and fault states, it is ideally suited to identify hidden trends in data that conventional techniques missed. Strong management of missing data and exceptions makes XGBoost more reliable in real-world scenarios where the quality of data could vary.

6. Conclusion

PV energy sources commonly encounter PV issues. In particular, issues like fragmentation, open-circuit short-circuiting, and others were quite common and hazardous. A new XGBoost technique for PV array defect diagnostics is presented in this study. The XGBoost algorithm is trained by collecting samples of PV array fault data. The purpose of pre-processing data was to deal with missing values or eliminate noisy data. Feature extraction is done using the LDA to enhance detection performance. To improve XGBoost's efficacy, the WCO approach is applied to produce the best features from the recovered data. Then we utilize the XGBoost method to effectively detect issues with the solar arrays. The proposed method is compared to the existing method as SVM in terms of trained accuracy (98.9%), rest accuracy (97.5%), and difference (1.4). The proposed method is better when compared to the existing methods.

6.1. Limitation and Future Scope

Limitation: The quality, comprehensiveness, and accuracy of the data used to train the algorithm have a major impact on the problem diagnosis's accuracy and dependability. Insufficient or skewed data could result in less-than-ideal performance or even a false diagnosis. The model's internal functioning could prove complicated and tricky to read, despite its excellent predicted accuracy. This makes it difficult to comprehend the underlying causes of certain diagnostic findings.

Future scope: The future study in this field appears to have numerous interesting directions. One way to increase data collection and raise the caliber of information provided is to integrate innovative sensor technologies with Internet of Things (IoT) devices. This could result in defect identification and diagnosis are more reliable and accurate. The adoption of real-time defect diagnostic systems for large-scale solar arrays could be made possible by advancements in computer resource efficiency and scalability, which would enhance maintenance procedures and overall system dependability.

Acknowledgements

Transmission line drone remote charging system （067500KK52210001）

References

[1] Sharma, V.K., Singh, R., Gehlot, A., Buddhi, D., Braccio, S., Priyadarshi, N. and Khan, B., 2022. Imperative role of photovoltaic and concentrating solar power technologies towards renewable energy generation. International Journal of Photoenergy, 2022(1), p.3852484.

[2] Dhanraj, J.A., Mostafaeipour, A., Velmurugan, K., Techato, K., Chaurasiya, P.K., Solomon, J.M., Gopalan, A. and Phoungthong, K., 2021. An effective evaluation of fault detection in solar panels. Energies, 14(22), p.7770.

[3] Lazzaretti, A.E., Costa, C.H.D., Rodrigues, M.P., Yamada, G.D., Lexinoski, G., Moritz, G.L., Oroski, E., Goes, R.E.D., Linhares, R.R., Stadzisz, P.C. and Omori, J.S., 2020. A monitoring system for online fault detection and classification in photovoltaic plants. Sensors, 20(17), p.4688.

[4] Fotopoulou, M., Rakopoulos, D., Trigkas, D., Stergiopoulos, F., Blanas, O. and Voutetakis, S., 2021. State-of-the-art low and medium voltage direct current (DC) microgrids. Energies, 14(18), p.5595.

[5] Mustafa, R.J., Gomaa, M.R., Al-Dhaifallah, M. and Rezk, H., 2020. Environmental impacts on the performance of solar photovoltaic systems. Sustainability, 12(2), p.608.

[6] Soomar, A.M., Hakeem, A., Messaoudi, M., Musznicki, P., Iqbal, A. and Czapp, S., 2022. Solar photovoltaic energy optimization and challenges. Frontiers in Energy Research, 10, p.879985.

[7] dos Santos, S.A.A., Torres, J.P.N., Fernandes, C.A. and Lameirinhas, R.A.M., 2021. The impact of aging of solar cells on the performance of photovoltaic panels. Energy Conversion and Management: X, 10, p.100082.

[8] Fairbrother, A., Quest, H., Özkalay, E., Wälchli, P., Friesen, G., Ballif, C. and Virtuani, A., 2022. Long-Term

Performance and Shade Detection in Building Integrated Photovoltaic Systems. Solar Rrl, 6(5), p.2100583.

[9] Lipták, R. and Bodnár, I., 2021. Simulation of fault detection in photovoltaic arrays. Analecta Technica Szegedinensia, 15(2), pp.31-40.

[10] Wang, A. and Xuan, Y., 2021. Close examination of localized hot spots within photovoltaic modules. Energy Conversion and Management, 234, p.113959.

[11] Karimi, M., Samet, H., Ghanbari, T. and Moshksar, E., 2020. A current-based approach for hotspot detection in photovoltaic strings. International Transactions on Electrical Energy Systems, 30(9), p.e12517.

[12] Chen, S.Q., Yang, G.J., Gao, W. and Guo, M.F., 2020. Photovoltaic fault diagnosis via semisupervised ladder network with string voltage and current measures. IEEE Journal of Photovoltaics, 11(1), pp.219-231.

[13] Zhao, J., Sun, Q., Zhou, N., Liu, H. and Wang, H., 2020. A photovoltaic array fault diagnosis method considering the photovoltaic output deviation characteristics. International Journal of Photoenergy, 2020(1), p.2176971.

[14] Liu, Y., Ding, K., Zhang, J., Li, Y., Yang, Z., Zheng, W. and Chen, X., 2021. Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with IV curves. Energy Conversion and Management, 245, p.114603.

[15] Li, C., Yang, Y., Zhang, K., Zhu, C. and Wei, H., 2021. A fast MPPT-based anomaly detection and accurate fault diagnosis technique for PV arrays. Energy Conversion and Management, 234, p.113950.

[16] Abbas, M. and Zhang, D., 2021. A smart fault detection approach for PV modules using Adaptive Neuro-Fuzzy Inference framework. Energy Reports, 7, pp.2962-2975.

[17] Eskandari, A., Aghaei, M., Milimonfared, J. and Nedaei, A., 2023. A weighted ensemble learning-based autonomous fault diagnosis method for photovoltaic systems using genetic algorithm. International Journal of Electrical Power & Energy Systems, 144, p.108591.

[18] Fan, J., Rao, S., Muniraju, G., Tepedelenlioglu, C. and Spanias, A., 2020, June. Fault classification in photovoltaic arrays using graph signal processing. In 2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS) (Vol. 1, pp. 315-319). IEEE.

[19] Tao, C., Wang, X., Gao, F. and Wang, M., 2020. Fault diagnosis of photovoltaic array based on deep belief network optimized by genetic algorithm. Chinese Journal of Electrical Engineering, 6(3), pp.106-114.

[20] Hojabri, M., Kellerhals, S., Upadhyay, G. and Bowler, B., 2022. IoT-based PV array fault detection and classification using embedded supervised learning methods. Energies, 15(6), p.2097.

[21] Badr, M.M., Hamad, M.S., Abdel-Khalik, A.S., Hamdy, R.A., Ahmed, S. and Hamdan, E., 2021. Fault identification of photovoltaic array based on machine learning classifiers. IEEE Access, 9, pp.159113-159132.

[22] Li, B., Delpha, C., Migan-Dubois, A. and Diallo, D., 2021. Fault diagnosis of photovoltaic panels using full I–V characteristics and machine learning techniques. Energy Conversion and Management, 248, p.114785.

[23] Basnet, B., Chun, H. and Bang, J., 2020. An intelligent fault detection model for fault detection in photovoltaic systems. Journal of Sensors, 2020(1), p.6960328.

[24] Hajji, M., Harkat, M.F., Kouadri, A., Abodayeh, K., Mansouri, M., Nounou, H. and Nounou, M., 2021. Multivariate feature extraction-based supervised machine learning for fault detection and diagnosis in photovoltaic systems. European Journal of Control, 59, pp.313-321.

[25] Kapucu, C. and Cubukcu, M., 2021. A supervised ensemble learning method for fault diagnosis in photovoltaic strings. Energy, 227, p.120463.

[26] Wang, J., Gao, D., Zhu, S., Wang, S., and Liu, H., 2023. Fault diagnosis method of photovoltaic array based on support vector machine. Energy sources, part a: recovery, utilization, and environmental effects, 45(2), pp.5380-5395.