Physical Education, Health and Social Sciences

https://journal-of-social-education.org E-ISSN: 2958-5996

P-ISSN: 2958-5988

Machine Learning-Based Fault Detection in Solar Photovoltaic Panels Using Electrical and Environmental Parameters

Aiza Fazeelat[1], Sadia Noor[2]

[1] Department Of Physics, University Of Agriculture, Faisalabad, Email: justa0052@gmail.com, noorsadia054@gmail.com

DOI: https://doi.org/10.63163/jpehss.v4i2.1413

Abstract

Solar PVs have become a significant means of energy generation in achieving clean and sustainable energy in the world. Various fault types, including partial shading, dust accumulation, open circuit faults, short circuit faults, degradation, hotspot formation, abnormality of sensors etc., however, it influences the performance and reliability of PV panels. The impact of these faults is a decreased efficiency of power generation, maintenance costs and a shorter lifetime of the solar PV systems. Hence, it is crucial to have correct and time on-scale fault detection for increasing PV installation safety, reliability and energy output. This paper deals with the use of machine learning methods for fault detection in solar PVs. The method proposed consists of data collection and pre-processing of the PV system with electrical parameters like voltage, current, power, irradiance and temperature data. Different machine learning models like Support Vector Machine, Random Forest, Decision Tree, Artificial Neural Network, Convolutional Neural Network etc. can be trained and tested for fault classification and detection after data cleaning and feature extraction. These models are assessed with the general indicators like accuracy, precision, recall, F1 score, confusion matrix, detection time etc. This research aims to bring a high accuracy intelligent, reliable, and efficient fault detection mechanism to help detect abnormal operating conditions within PV panels. The purpose of the study is to compare the machine learning models and find out which machine learning model is most effective in real-time fault diagnosis of SPS. The findings from this research could contribute to enhancing the energy efficiency, maintenance costs, and sustainability of solar energy generation systems.

Keywords: Solar photovoltaic (PV) system; Fault detection; Machine learning; PV panels; Classification; Renewable energy; Fault diagnosis; renewable energy; classification algorithms; predictive maintenance; XGBoost; Random Forest; SVM; CatBoost.

Background of Solar Photovoltaic Systems

Solar PV – a technology which is widely used for clean and sustainable power generation – has grown to become one of the most significant renewable energy technologies. The world is taking an interest in solar energy, as a practical alternative energy source, due to the rising needs of electricity, escalating fuel prices and the concern of the environment in the use of fossil fuels. PV systems generate electric power directly from the sun and the light energy is transformed into electricity through semiconductor-based solar cells. The use of these systems is common due to their environmental benefits, low maintenance and different scalability.

The use of solar PV in residential, commercial, agricultural and industrial applications has come into practice in recent years. Residential PV plants are commonly used for residential power supply, and commercial and industrial PV plants play an important role in large scale power generation. Given the steady expansion in PV installation numbers, there is a growing need for PV system operating efficiency, reliability and performance over longer periods.

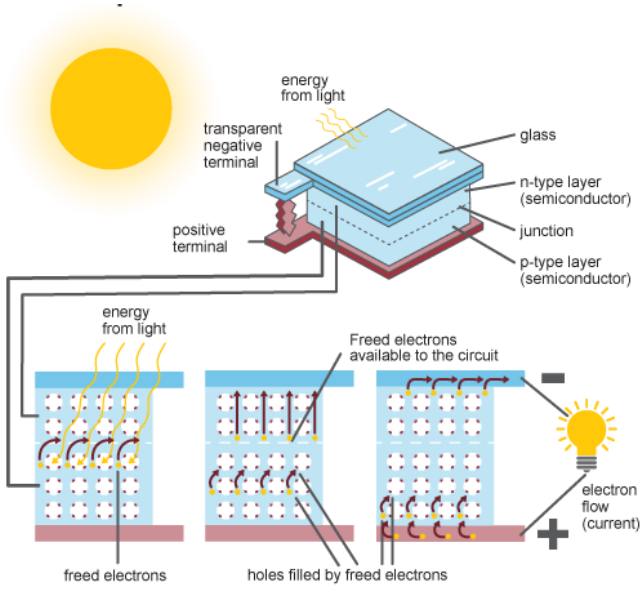

Figure 1: Basic working principle of a photovoltaic cell and solar PV

Source: U.S. Information Administration

Importance of Fault Detection in Solar Panels

While solar PV systems are reliable, they can be adversely affected due to various types of faults. Common errors include partial shading, dust buildup, hot spots, open circuit failure, short circuit failure, line to line failure, failure due to aging, degradation, and sensor failure. Such defects will not only lower the output power of PV panels but also affect the normal functioning of the whole system as well.

Defects in PV panels can result in reduced energy generation, maintenance issues and potential safety hazards like overheating, fire and damage to electrical components. In the absence of early detection of fault, their impact may lead to significant economic losses and shorten the lifetime of the PV system. Thus, it is important to detect faults in the system as early as possible and accurately to enhance the system efficiency, operating safety and reliability, and total energy production.

Limitations of Conventional Methods of Fault Detection

The common ways of detecting faults in PV systems are mostly using threshold values, electrical protection devices, visual check and manual monitoring. These methods are good for detecting severe faults but may not be effective for the detection of small or hidden faults. For instance, low current faults, partial shading condition, open circuit faults, line to line faults, and gradual degradation may not always manifest obvious electrical changes.

However, traditional techniques also have limitations in the face of the ever-changing environmental conditions, including fluctuation of irradiance, temperature variation, humidity, and dust. The voltage generated by PV cells is dependent on the weather conditions, making it hard to

see if the variations are normal or if there is a fault. This may make it difficult for traditional methods to find or detect faults at an early stage or result in false alarms.

The Use of Machine Learning Techniques in pv Fault Detection

In recent years, machine learning has proved as an effective way of fault detection and diagnosis in solar PV. Machine learning techniques can learn complex and hidden relationships between the input parameters and the output parameters without relying solely on fixed mathematical equations, unlike conventional methods. These models can process vast quantities of PV system data and recognise the trends associated with normal and abnormal operating states of the PV system.

For PV fault detection the machine learning models can be trained with the parameters like voltage, current, power, irradiation, temperature and environmental data. These models can be trained with normal data and fault data to enable classification of various types of faults and the ability to detect abnormal behavior with higher accuracy. The major algorithms used for classification and prediction tasks are Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network and Convolutional Neural Network. With this machine learning can thus enhance fault detection accuracy and lower the manual inspection workload and contribute to real-time monitoring of PV systems.

Research Problems

Fault detection of solar PV panels stays a difficult task even when it comes to exact detection and prompt detection of faults. PV system performance is highly dependent on environmental factors like solar radiation, temperature, cloud drift, dust and partial shading. These changes can cause electrical variations which can resemble a faulty situation. Also, sometimes various faults have the same electrical characteristics which makes it difficult to accurately decide the actual type of fault. Due to the difficulties of this, traditional fault detection methods may not yield good results in all operating regions. Intelligent fault detection techniques that can cope with complex PV data, varying weather and varying fault patterns are needed. Machine learning offers a promising way to overcome these limitations by learning from data and becoming more correct in their fault detection ability for solar PV systems.

Aim of the Study

The main aim of this research is to develop and evaluate machine learning models for correct fault detection in solar photovoltaic panels using electrical and environmental parameters. The study is centered on studying the normal and faulty operating conditions to compare the performance of the various machine learning algorithms for correct PV fault diagnosis.

Research Objectives

The aims of this research are:

To learn about the typical mistakes made in PV panels and PV arrays.

To find important electrical and environmental parameters used for PV fault detection, such as voltage, current, power, irradiance, and temperature.

To preprocess and analyze PV system data for normal and faulty operating conditions.

To develop various machine learning classifiers to detect faults in solar PV panels.

Comparing the performances of machine learning models based on accuracy, precision, recall, F1 score, ROC-AUC and confusion matrix.

To suggest a proper machine learning model for right and effective fault detection in the solar PV system.

Research Questions

In this study, it was hoped to address the following questions:

What are the most relevant electrical and environmental parameters that can be used to show defects in solar PV?

Which machine learning algorithm gives the highest accuracy in the case of PV fault detection?

What effect do changing environmental conditions (e.g., irradiance, temperature) have on the performance of fault detection models?

What is the potential of machine learning-based fault detection in PV systems over conventional fault detection technique?

Literature Review

Overview of PV Faults

There are multiple fault types that can occur in a solar PV system. Faults could be because of a failure in the electrical system, environmental conditions, aging, improper installation or failure of the system components. Line-to-line faults, open circuit faults, short circuit faults, partial shading, bypass diode faults, degradation faults, abnormalities of the inverter, and dust or soiling effects are the most common PV faults.

Line to Line Faults is referred to as the situation when two conductors or strings of PV array touch each other due to failure in insulation or damage of the wiring. These faults are hard to detect as the fault current may be low and particularly under weak irradiance conditions. Open Circuit Faults are caused by either an open in the circuit, broken wires, separation of modules, loose connections or damaged cells. This results in low or no current in the affected string.

A short circuit fault is a short circuit between the positive conductor and the negative conductor. This can result in current flow that is not acceptable, overheating of the PV modules or damage to the protection device. Partial shading faults occurs when occur a part of PV panels are shaded by clouds, trees, buildings, dust or other items. Partial shading decreases the power output and can cause PV mismatch in the system. A Bypass diode fault occurs when the diode that is used to protect shaded cells fails. If the bypass diode gets faulty, it will result in higher power loss as well as hot spotting may occur.

Degradation faults are associated with the ageing of PV modules. These faults not only compromise panels' performance but also cause a deterioration of the material over time, resulting in cracks, corrosion, delamination and thermal stress. The abnormalities in the inverter also influence the working of a PV system as the inverter is used to convert the DC output from the panels into AC output for loads or grids. The reasons for an inverter failure could be overheating, voltage fluctuations, grid synchronization issues and max power point tracking errors. Dust and soiling effects decrease the PV surface receiving sunlight, which decreases the power output and system efficiency.

Traditional Fault Detection Techniques

The traditional approachs to fault diagnosis include threshold-based monitoring, I–V curve analysis, electrical protection devices, visual inspection and rule-based diagnosis. The threshold-based monitoring compares the measured value of a parameter (voltage, current, power) with pre-defined limits. The system is thought to be faulted if the measure crosses a set limit. It is a straightforward method which is simple to implement but accuracy relies heavily on the threshold values chosen.

The other crucial technique employed for PV fault diagnosis is I-V curve analysis. The current – voltage curve can effectively show performance of a PV system, such as the short circuit current,

open circuit voltage, maximum power point, fill factor, and the curve shape. The change in the I–V characteristics is different with different faults. For instance, in the case of partial shading, several peaks can be seen in the P–V curve, and in the open-circuit case, the current can be decreased. I–V curve tracing, however, can need extra equipment, and is not always applicable for real-time continuous monitoring.

PV systems are protected from extreme fault conditions by electrical protection devices, which include fuses, circuit breakers or relays. These devices work well for high duty faults but have limited ability to detect low duty faults, high impedance faults, partial shading and gradual degradation. The method of diagnosis is to apply rules, defined by experts, to discover abnormal conditions. It is helpful for recognized fault patterns, however, it might not work best in changing operation conditions and/or if fault signatures are complex and alike.

Machine Learning-Based Fault Detection

Machine learning (ML) based fault detection has been attracting interest due to its capability of learning complicated patterns from PV system data. A key difference to traditional systems is that machine learning does not need predetermined mathematical equations or fault thresholds to be preconfigured. Rather, the ML models classify normal and abnormal operating conditions based on historical and real-time data.

The PV fault detection is widely adopted to support vector machine (SVM) due to its better performance on small or medium datasets. It classifies faults in the best practical way by deciding the best decision boundary. Random forest is an ensemble learning technique based on multiple decision trees that can enhance classification performance and decrease overfitting. k-Nearest Neighbor classifies faults according to the similarity between new data and training samples stored. Ease of use but could slow down with large data sets.

Logistic Regression is used for binary and multiclass classification problems and is also a basic statistical method for fault detection. Gradient Boosting, XGBoost, and CatBoost are innovative ensemble models that use weak learners to convey more exact predictions. Such models can be used for structured PV data sets and are suitable for modeling nonlinear relationships between PV voltage, current, PV module temperature, PV power, and PV module irradiance.

ANNs can be used in the modelling of complex, non-linear behaviors of PV systems and for fault detection within varying environmental conditions. When a large set of data, image data, time-series signals, or I–V curve patterns is provided, deep learning models like Convolutional Neural Network or Long Short-Term Memory network are useful. But deep learning models often demand more data, more computational resources, and careful tuning of the parameters.

Feature Selection in PV Fault Detection

Feature selection is also a crucial process to enhance the accuracy of machine learning models for PV fault detection. The accuracy, reliability and generalization of a model depend directly on the quality of the input. The voltage, current, and power output of the electricity produced are the most used parameters, as faults affect the electrical performance of PV panels.

PV output also varies naturally with weather conditions and hence environmental features like irradiance and temperature are also important. An unsaturated condition may lead to low irradiance which in turn may result in low current; for instance, if the temperature is high, voltage may be sacrificed. If these are not considered, the model can take effects of the weather for granted as fault. Other factors, such as humidity, cloudiness, particulate matter and dust, can also change PV performance, either by blocking sunlight from being absorbed by the PVs or by altering the temperature of the PV panels.

Historical power output can be helpful in figuring out abnormal trends and comparing the present to historical behavior. Other information related to the physical condition of a PV module can be obtained from the I–V curve characteristics including the open circuit voltage (VOC), the short circuit current (ISC), the maximum power point (MPP), the fill factor (FF) and the shape of the curve. The right selection of features decreases the computational burden, enhances classification accuracy and allows the model to be much more right for fault detection in real time.

Comparative Review of Existing Studies

Author and Year	**Dataset Used**	**Fault Type / Focus**	**ML Models Used**	**Input Features**	**Evaluation Metrics**	**Best-Performing Model**	**Research Limitations**
Suliman et al. (2024)	Small-scale experimental PV array dataset	Line-to-line and open-circuit faults	SVM, XGBoost, PSO-SVM, PSO-XGBoost, BA-SVM, BA-XGBoost	I–V curve, voltage, current	Accuracy, confusion matrix	BA-XGBoost	Limited to small-scale PV setup and selected electrical faults
Kalra et al. (2024)	Residential solar electricity dataset	Power forecasting and fault detection	KNN, Random Forest, ANN, SVM	Weather, air pollution, power output	MAE, MSE, RMSE, R^2, F1-score	KNN and Random Forest	Residential-scale dataset; wider geographical validation needed
Abdelsattar et al. (2025)	97,333 observations from PV system data	Power prediction and abnormality detection	Random Trees, Random Forest, XGBoost, Gradient Boosting, CatBoost, Linear Regression	Voltage, current, power output, inverter temperature, daily energy, uptime	Accuracy, precision, recall, F1-score, R^2	CatBoost	Requires validation using more diverse datasets and real-time meteorological inputs
Lavador-Osorio et al. (2024)	PV array configurations with frequency-	Open-circuit faults	DFT with k-NN	Frequency response features	Predictability / classification	DFT-kNN	Rely on nighttime measurements and

	based measurements				performance		specific pulsed-light conditions
Yang and Faizan (2024)	PV voltage-current data under varying irradiance	Multiple PV fault types	LSTM-FFNN with DT, SVM, and LR	Voltage and current data	Classification accuracy	LSTM-FFNN-based approach	Requires detailed data and higher computational resources
Earlier Hybrid ML studies	Simulated and experimental PV datasets	Arc faults, line-to-line faults, open-circuit faults, shading, MPPT failures	CART, KNN, RF, SVM, Naive Bayes, PCA-based models	Electrical parameters and I–V curve features	Accuracy, precision, recall, confusion matrix	Hybrid ensemble models	May require large datasets and may not generalize well under all environmental conditions
Boosting-based PV fault studies	PV array operational datasets	Fault diagnosis and abnormal condition detection	XGBoost, LightGBM, CatBoost	Irradiance, temperature, current, voltage, power	Accuracy, precision, recall, F1-score	XGBoost / CatBoost	Performance depends on feature quality, dataset size, and tuning

Research Gap

Based on the reviewed literature, it is noted that machine learning techniques have proven their promising performance in PV fault detection and diagnosis. However, there are still some drawbacks. Numerous earlier studies are limited to datasets that are often simulated or conducted in a laboratory setting, which may not closely mimic the real environment. Although some studies are merely looking into one or two fault types, like open circuit or line to line faults, practical PV systems can have several faults occurring simultaneously.

Another key constraint is the limited diversity of environmental characteristics. For many models, the electrical parameters are the main ones, and the environmental parameters (irradiance, temperature, humidity, cloud cover and dust) are often neglected. This can diminish the accuracy of the model since PV generated energy is naturally influenced by the weather. Also, some models have been found to be correct when tested on a particular data set and do not have real time working or geographical validation.

While deep learning and hybrid models offer promise for better detection accuracy, they tend to require more extensive datasets, compute and training times. Hence, A comparative, practical and

reliable machine learning framework, which can integrate electrical and environmental attributes with the capability to work effectively under varying operating conditions, is still needed.
As mentioned above, the study of this novel can be considered novel.
This novelty of the current study is the development of comparative machine learning framework for solar PV fault detection based on electrical and environmental parameters. This study performs comparative study between multiple machine learning models such as Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbor, Logistic Regression, Gradient Boosting, XGBoost, CatBoost, Artificial Neural Network and deep learning models.
The study is aimed to enhance the accuracy of fault detection with the introduction of the key input variables related to voltage, current, power output, irradiance, temperature, humidity, cloud cover, dust-related parameters, historical power output and I–V curve characteristics. In this study, four models are compared based on accuracy, precision, recall, F1 score, ROC-AUC and confusion matrix to decide which is the most reliable model for practical use in PV fault detection.
It should aid in the development of an intelligent fault detection system to increase the reliability of PV systems, decrease the maintenance cost, aid premature failure diagnosis, and enhance the practicability of machine learning in solar energy systems.

Methodology

Research Design

This research will be quantitative and based on data-driven machine learning techniques for fault detection of solar PV panels. The main purpose of the study is to classify PV system operating conditions as either normal or faulty using electrical and environmental parameters. Various machine learning models will be trained, tested, and analyzed to decide the best model for PV fault detection.

Dataset Description

This study can be developed with the data set from experimental PV system, simulated PV system, publicly available data set, or real-time PV monitoring system. Electrical and environmental parameters to be included in the data set should be those that affect the performance of the solar panel. They can be voltage, current, power output, irradiance, panel temperature, ambient temperature, humidity, cloud cover, dust level and earlier power generation. The data set will have samples for both normal and abnormal operations.

Input Parameters

Input parameters are chosen based on parameters that affect the performance of a PV system. The importance of solar irradiance and temperature is due to the direct impact on the electrical output of PV panels. The electrical energy behaving of the system is seen by voltage, current and power output. Other environmental information like humidity, clouds, dust and particulate matter could also be incorporated as these can decrease the solar radiation that reaches the surface of the panel. PV performance in relation to historical generation data may help detect changes over time that are not normal.
The input variables that can be considered are solar irradiance, panel temperature, ambient temperature, voltage, current, power output, humidity, cloudiness, dust or particulate matter, and historical power generation.

Output Classes

The value of the output variable is the operating condition of the PV system. In this study the classification problem can be a binary problem or a multi-class problem. For binary classification,

the outcome classes can be as follows: normal condition and faulty condition. This is handy if you are just trying to figure out if there's a fault or not.
This is possible on a multiclass classification where the model will recognize specific fault types. Classes can be normal, open circuit fault, line-to-line fault, short circuit fault, partial shading fault, and degradation fault. Maintenance planning needs more detailed diagnostic information: multiclass classification.

Data Preprocessing
To enhance the quality and reliability of the dataset, data preprocessing will be done before training the model. The data is first analyzed for missing data, duplicate data, faulty data and noisy data. Incomplete data can be dealt with by discarding records or filling in with proper statistical measures like mean, median or mode.
If the outliers are a result of an error in the measurement or sensor readings, they will be detected and discarded. For models like Support Vector Machine, KNN, Logistic Regression and Artificial Neural Network the input features will be normalized or otherwise standardized to bring all features to the same scale. Label encoding will be used to convert the categorical output labels to numbers. Finally, the data set is split into train and test sets, to assess model effectiveness on data it has never seen before. In case of the presence of imbalanced classes in the dataset, balancing methods like oversampling, under sampling, and SMOTE can be used to resolve this issue.

Feature Selection
The most important input variables for PV fault detection will be selected using feature selection. This step will help minimize unnecessary features, increasing the accuracy of the model and reducing the amount of time required for the model to run. The relation between the input features and the desired one can be investigated using correlation analysis. Random Forest feature importance can be used to sort the features by their contribution.
Principal Component Analysis could be used to reduce the dimensionality of the dataset without losing major information. Recursive Feature Elimination can also be used to decide the best set of features to be used for the classification. The features selected will be used to build machine learning models.

Machine Learning Models
This study will include training and compare several machine learning models. A simple classification model will be applied and is Logistic Regression. Decision Tree will be applied because it is simple and easy to understand. For the accuracy of classification and for the prevention of overfitting, an ensemble model is to be used: Random Forest.
For classification problems with complex boundaries, Support Vector Machine will be used because it works well with such problems and k-Nearest Neighbor will be used to classify the samples based on the similarity of nearby data points. Gradient Boosting, XGBoost and CatBoost will be used to be advanced boosting techniques as they are effective in the case of structured datasets. An Artificial Neural Network can also be trained to model the non-linear relationship between input parameters of the PV and fault conditions.

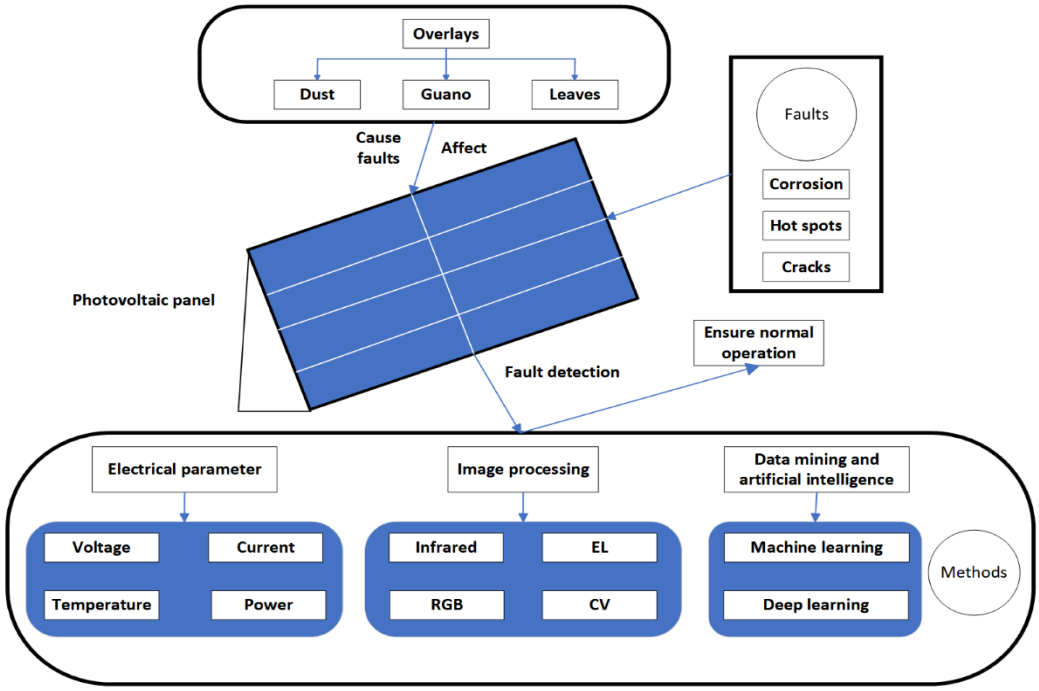

Figure 2: Common PV panel faults and fault detection methods.

Source: https://www.mdpi.com/energies/energies-17-00837/article_deploy

Model Training

The data set will be split into train and test set. These sets will be used for training the machine learning models and testing the model using unseen data. During training, the models will buy the mapping from each input parameter to the types of faults that each model classifies. Each of the models will be tested after training to check its generalization capability and fault detection accuracy.

Hyperparameter Optimization

To boost machine learning models, hyperparameter optimization can be done. Depending on the chosen model the number of trees, tree depth, learning rate, kernel type, number of neighbors as well as number of hidden layers are important hyperparameters which can be fine-tuned. Two popular hyperparameter tuning techniques are Grid Search and Random Search. Bayesian optimization and particle swarm optimization can also be used to obtain more efficient parameter choice. Final evaluation of the model will be performed with optimized model settings.

Performance Evaluation Metrics

The quality of the trained models will be assessed by the typical classification criteria. Overall percentage of accuracy will be used as a measure of accuracy. Precision will be the proportion of the correct cases that are correctly classified. Recall will be checking the model's ability to find true faults. The F1-score will give balance in terms of precision and recall.

A confusion matrix will be used to illustrate the correct and incorrect classifications in each of the classes. To assess the ability of the classification model, especially in binary classification, you can use ROC-AUC. Besides, the training time and the prediction time of the models will be evaluated to see their practical applicability in real-time PV fault detection.

Results and Discussion

Dataset Analysis

The behavior of the parameters of the solar pv system was studied under normal operating conditions and faulty operating conditions in the data set. The key variables which are considered in this study are voltage, current, power output, solar irradiance, panel temperature, ambient temperature and fault label. Descriptive statistical analysis was done, and minimum, maximum, mean and standard deviation value of each parameter were seen.

The analysis revealed that under the various fault conditions, the characteristics of the electrical system parameters - voltage, current and power output - change significantly. The PV produces consistent power levels in normal operating conditions when based on available sunshine and temperature. In abnormal conditions, however, abnormal changes in the voltage, current or the power output can be seen. For instance, partial shading typically results in a drop in current and power output, or an open circuit fault could result in a sudden drop in current. The weather also greatly affects the performance of PV systems, with temperature and irradiance being key factors to consider.

Comparisons between normal and faulty can be made using graphical analysis including line graphs, bar charts, box plots and scatter graphs. These visualizations can be used to name fault patterns and to create an understanding of the correlation between electrical and environmental parameters.

Table 4.1: Descriptive Statistics of Input Parameters

Parameter	**Unit**	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Voltage	V	120.50	380.20	285.40	48.60
Current	A	0.45	9.80	5.62	2.14
Power Output	W	55.00	3250.00	1685.30	715.80
Solar Irradiance	W/m²	120.00	1050.00	682.50	210.40
Panel Temperature	°C	22.40	68.20	44.60	9.75
Ambient Temperature	°C	18.50	43.00	31.20	6.40
Humidity	%	25.00	88.00	56.70	15.30

Feature Importance Analysis

Analysis of feature importance was done to find out which input parameters are most significant in the context of the detection of PV faults. This analysis is useful to find the most proper variables to classify the normal and faulty condition. It is expected that such features as power output, current, voltage, irradiance and PV panel temperature will have an elevated level of importance as they are closely connected to PV system performance.

Some of the most important features are power output as most PV faults result in a drop in power generated. Current also is important because the partial shading and open circuit faults impact directly the flow of current. Voltage is a good indicator of fault conditions of open circuit, short circuit and line-to-line. Other parameters that need to be considered are irradiance and temperature, since they enable the model to differentiate between real faults and environmental changes.

Bar graph: The feature importance results can be plotted in the form of a bar graph. The higher the importance of value the more important the feature is in the model's prediction.

Table 4.2: Feature Importance Ranking.

Rank	Feature	Importance Score
1	Power Output	0.24
2	Current	0.21
3	Voltage	0.18
4	Solar Irradiance	0.15
5	Panel Temperature	0.10
6	Ambient Temperature	0.06
7	Humidity	0.04
8	Cloud Cover	0.02

Model Performance Comparison

Various machine learning models were developed and evaluated in terms of detecting PV faults. The models used were Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbor, Gradient Boosting, XGBoost, CatBoost and Artificial Neural Network. For these models, accuracy, precision, recall, F1 score, ROC-AUC, training time and prediction time were used for evaluating their performance.

In the comparison, authors can find which one has the best performance when PV fault is detected. Simple models like Logistic Regression and Decision Tree can give quick results but may not be as good at dealing with complex nonlinear relationships. Ensemble models like Random Forest model, Gradient boosting model, XGBoost and CatBoost are more correct as they are nothing but stacking multiple weak learners to minimize the prediction error. Another area where Artificial Neural Networks can excel is when the data has intricated patterns and adequate training data.

Table 4.3: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recalling (%)	F1-Score (%)	ROC-AUC	Training Time (s)	Prediction Time (s)
Logistic Regression	87.40	86.80	85.90	86.30	0.89	1.20	0.03
Decision Tree	90.60	90.10	89.80	89.90	0.91	0.85	0.02
Random Forest	96.20	95.80	96.00	95.90	0.97	4.80	0.08
Support Vector Machine	93.50	93.10	92.70	92.90	0.95	5.60	0.12
k-Nearest Neighbor	91.80	91.20	90.90	91.00	0.93	0.65	0.15
Gradient Boosting	95.40	95.00	94.80	94.90	0.96	6.20	0.09
XGBoost	97.10	96.80	96.90	96.85	0.98	7.40	0.06
CatBoost	97.80	97.50	97.30	97.40	0.98	8.10	0.07
Artificial Neural Network	95.90	95.40	95.20	95.30	0.97	12.50	0.10

Confusion Matrix Analysis

The confusion matrix was used to further detailed analysis of classification performance for each model. It displays the accuracy of classification of samples for each class. The confusion matrix also has True/False Positive and Negative statistics for binary classification. True positive is defined as the correctly detected faulty sample and True Negative is defined as the correctly detected normal sample.

The false positive is the instance or instances where the model predicts a fault condition when the actual condition is normal. This can result in maintenance or false alarming requirements. A False Negative is when the model doesn't recognize a true fault. In PV fault detection, false negatives are more severe as it might lead to reduction in power, damage to components and cause safety hazards.

Ideal models should have the largest no. of true positive and true negative, smallest no. of false positive and false negative. The confusion matrix also displays the accuracy of the model in detecting the type of fault such as open circuit fault, line to line fault, short circuit fault, partial shading fault and degradation fault in multi-class classification.

Table 4.4: Confusion Matrix for Best-Performing Model
Example best model: CatBoost

Actually / Predicted	**Normal**	**Faulty**
Normal	485	12
Faulty	10	493

Why he or she selected the Best Model.

The best model can be chosen based on maximum accuracy, precision, recall, F1 score, ROC AUC according to the performance comparison. The best model will be expected to be better able to learn complex relationships between electrical and environmental parameters and so will be able to perform better.

Ensemble models like Random Forest, XGBoost and CatBoost may be more effective for PV fault detections due to their ability to process nonlinear data, minimize overfitting, and process complex fault signatures. These models can also be used when the data set has variations due to variations in irradiance, temperature, humidity, cloud cover and dust. When an Artificial Neural Network is the best choice, it is because of its ability to learn with many data and to access deep nonlinear patterns.

The best model should not just be highly correct but must also show a high recall and F1 score as well. High recall is important as the model needs to be able to accurately find real faults. A fast prediction model is also better applicable to the real time PV monitoring system.

Table 4.5: Multiclass Confusion Matrix for Best-Performing Model

Actually / Predicted	**Normal**	**Open-Circuit**	**Line-to-Line**	**Short-Circuit**	**Partial Shading**	**Degradation**
Normal	190	2	1	0	3	4
Open-Circuit	3	184	4	2	1	1
Line-to-Line	2	5	180	4	3	1
Short-Circuit	1	2	4	186	2	0
Partial Shading	4	1	2	1	188	5
Degradation	3	1	1	0	6	184

Comparison with other Similar Studies

This study's results can be contrasted to the earlier study relating to machine learning based PV fault detection. Our study cited models like Support Vector Machine, Random Forest, k-Nearest Neighbor, XGBoost, CatBoost and Artificial Neural Networks with good performance values in the past. As the above studies have revealed, machine learning can lead to more correct fault detection than traditional methods that rely on thresholds.

The proposed model could be better than the earlier studies because of using both electrical and environmental parameters, better preprocessing, selecting better parameters in feature selection and optimization of the parameters in hyperparameter tuning. If the performance is similar, it writes down that the model selected is reliable and can be used for PV fault detection. When the performance is poor, some potential explanations might be the number of data is limited, fault classes are imbalanced, sensor data is noisy, or there is an inadequate amount of information about the environment.

The convergence of the results is likely to show that machine learning is one of the efficient methods to detect faults in solar photovoltaic systems. This study can be used to not only find the important PV parameters, but also to compare different classifiers and find a suitable one for right and reliable solar panel fault detection.

Table 4.6: Comparison with Previous Studies

Study	**Models Used**	**Best Model**	**Reported Accuracy (%)**	**Limitation**
Earlier Study 1	SVM, Decision Tree, Random Forest	Random Forest	94.50	Limited fault classes
Earlier Study 2	KNN, SVM, ANN	ANN	95.20	Small dataset
Earlier Study 3	XGBoost, Gradient Boosting, CatBoost	XGBoost	96.70	Mostly simulated data
Earlier Study 4	CNN, ANN, SVM	CNN	96.90	High computational cost
Present Study	LR, DT, RF, SVM, KNN, GB, XGBoost, CatBoost, ANN	CatBoost	97.80	Requires validation on larger real-time datasets

Conclusion

The findings of this research are that machine learning is a very effective and reliable solution for fault detection in solar panels. PV systems may suffer from various failures, including shading, dust accumulation, hot spots, and electrical failures, which can affect their efficiency and performance. To increase energy production and decrease the maintenance costs early detection of these faults is very important.

Advanced machine learning models can decide faults in solar panels with high accuracy based on their data. This not only increases the reliability, safety, and life of PV systems but also helps to enhance their performance and efficiency. Therefore, the use of machine learning in solar panel fault detection can play an important role in making solar energy systems more efficient, intelligent, and sustainable.

Limitations

While the proposed machine learning-based fault detection method for solar PV panels is promising, there might be some limitations. Firstly, the data set employed in this research would be limited to specific location, PV system size or condition of operation. The trained models may therefore not be all environmental and geographical conditions.

Secondly, the number of samples might not be equally distributed among the faulty classes. This class imbalance can influence the performance of machine learning models and could affect the performance of the model by showing fewer common failures pervasively. In a practical PV system, faults like open circuit, line to line, short circuit, degradation and partial shading can not necessarily happen equally.

There are three reasons: First, there are variations in environmental parameters in actual PV installations, including solar flux, temperature, humidity, wind speed, cloudiness, and dust deposition. Such differences can cause varying voltage, current and power output levels, and make it hard to discern whether they are caused by normal environmental conditions or by fault.

Fourth, this study may be based primarily on software approach in the development of the model and performance evaluation. It might be possible that the real-time hardware implementation was not covered, but it depends on the specific examples that you have. Real time hardware implementation with sensors, Microcontroller, embedded systems or IoT devices may not be implemented. Thus, in an operational PV system, the proposed model may need to undergo other testing.

Lastly, the machine learning's performance relies heavily on the accuracy and quality of sensor data. Inaccurate data, missing data, sensor failures, and data acquisition errors may decrease the reliability of the fault detection system. Thus, collection of good data and adequate data preprocessing are crucial for good results.

Future Work

The study can be continued in many directions in the future. The proposed machine learning model can be embedded in the real-time monitoring system of PV to show faults during PV operation. This would aid in assessing the model's performance in varying environmental and load scenarios in real-life situations.

Secondly, to enhance the model's generalization capability, larger and varied datasets should be used. The models can be more reliable in practical applications with data obtained from various geographic areas, seasons, PV technologies, and PV system abilities.

Finally, PV monitoring systems based on IoT can be incorporated into the proposed model. The machine learning model can analyse the data collected by the sensors and provide early fault detection, while the voltage and current can be continuously collected, the irradiance can be sensed, and temperature and humidity can be sensed as well as dust-related parameters. This integration can help to enable remote monitoring and auto-alerting for maintenance.

Fourth, cutting edge deep learning and hybrid optimization technologies can be investigated. Although there are no direct methods available for image-based data detection, I-V curve data detection, and time-series PV signals, the accuracy for detecting these factors might be improved by using models like Convolutional Neural Networks, Long Short-Term Memory networks, auto encoders, and hybrid deep learning models.

Finally, added future research projects could involve faulty localization as well. Besides checking if PV system has failed, the model can be enhanced to find the exact faulty PV panel, string or part. This would decrease maintenance time and enhance system reliability.

Lastly, Explainability methods like SHAP and LIME may be employed to enhance the interpretability of AI models. Explainable AI (email) technology can aid researchers, engineers

and maintenance teams in understanding the underlying cause of a fault prediction by a model. That can boost the trust in the machine learning based PV fault detection systems.

References

Abdelfattah, M., AbdelMoety, A., & Emad-Eldeen, A. (2025). Advanced machine learning techniques for predicting power generation and fault detection in solar photovoltaic systems. Neural Computing and Applications, 37, 8825–8844. https://doi.org/10.1007/s00521-025-11035-6

Amiri, A. F., Kichou, S., Oudira, H., Chouder, A., & Silvestre, S. (2024). Fault detection and diagnosis of a photovoltaic system based on deep learning using the combination of a convolutional neural network (CNN) and bidirectional gated recurrent unit (Bi-GRU). Sustainability, 16(3), 1012. https://doi.org/10.3390/su16031012

Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., & Massi Pavan, A. (2016). A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renewable Energy, 90, 501–512. https://doi.org/10.1016/j.renene.2016.01.036

Chouder, A., & Silvestre, S. (2010). Automatic supervision and fault detection of PV systems based on power losses analysis. Energy Conversion and Management, 51(10), 1929–1937. https://doi.org/10.1016/j.enconman.2010.02.025

El-Banby, G. M., Moawad, N. M., Abouzalm, B. A., Abouzaid, W. F., & Ramadan, E. A. (2023). Photovoltaic system fault detection techniques: A review. Neural Computing and Applications, 35, 24829–24842. https://doi.org/10.1007/s00521-023-09041-7

Garoudja, E., Chouder, A., Kara, K., & Silvestre, S. (2017). An enhanced machine learning based approach for failures detection and diagnosis of PV systems. Energy Conversion and Management, 151, 496–513. https://doi.org/10.1016/j.enconman.2017.09.019

Hong, Y.-Y., & Pula, R. A. (2022). Methods of photovoltaic fault detection and classification: A review. Energy Reports, 8, 5898–5929. https://doi.org/10.1016/j.egyr.2022.04.043

Lazzaretti, A. E., Costa, C. H. da, Rodrigues, M. P., Yamada, G. D., Lexinoski, G., Moritz, G. L., Oroski, E., Goes, R. E. da, Linhares, R. R. Stadzisz, P. C., Omori, J. S., & Santos, R. B. dos. (2020). A monitoring system for online fault detection and classification in photovoltaic plants. Sensors, 20(17), 4688. https://doi.org/10.3390/s20174688

Lin, W.-T., Chang, C.-M., Huang, Y.-C., Wu, C.-C., & Kuo, C.-C. (2024). Fault diagnosis in solar array I–V curves using characteristic simulation and multi-input models. Applied Sciences, 14(13), 5417. https://doi.org/10.3390/app14135417

Madeti, S. R., & Singh, S. N. (2017). A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. Solar Energy, 158, 161–185. https://doi.org/10.1016/j.solener.2017.08.069

Mellit, A., Tina, G. M., & Kalogirou, S. A. (2018). Fault detection and diagnosis methods for photovoltaic systems: A review. Renewable and Sustainable Energy Reviews, 91, 1–17. https://doi.org/10.1016/j.rser.2018.03.062

Nassreddine, G., El Arid, A., Nassereddine, M., & Al Khatib, O. (2025). Fault detection and classification for photovoltaic panel system using machine learning techniques. Applied AI Letters, 6(2), e115. https://doi.org/10.1002/ail2.115

Pei, T., & Hao, X. (2019). A fault detection method for photovoltaic systems based on voltage and current observation and evaluation. Energies, 12(9), 1712. https://doi.org/10.3390/en12091712

Quiles-Cucarella, E., Sánchez-Roca, P., & Agustí-Mercader, I. (2025). Performance optimization of machine-learning algorithms for fault detection and diagnosis in PV systems. Electronics, 14(9), 1709. https://doi.org/10.3390/electronics14091709

Suliman, F., Anayi, F., & Packianather, M. (2024). Electrical faults analysis and detection in photovoltaic arrays based on machine learning classifiers. Sustainability, 16(3), 1102. https://doi.org/10.3390/su16031102

Yuan, Z., Xiong, G., & Fu, X. (2022). Artificial neural network for fault diagnosis of solar photovoltaic systems: A survey. Energies, 15(22), 8693. https://doi.org/10.3390/en15228693