Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

FAULT DETECTION IN PHOTOVOLTAIC SYSTEMS USING MACHINE LEARNING APPROACHES

Tauseef Noor[1], Muhammad Usman[2], Hanan Ud Din[3], Hamail Raza Zaidi[4], Salman Ali Khan[5]

Lab Instructor at Iqra National University Peshawar.ORIC: https://orcid.org/0009-0006-0025-0378
Faculty of Arts, Computing & Engineering, Wrexham University, Wrexham
Lab Instructor at Iqra National University Peshawar
Software Engineering Head of Program, Iqra National University Peshawar.ORIC: https://orcid.org/0009-0006-8579-3768
Lecturer at Iqra National University Peshawar
Tauseef.noor.69@gmail.com, engr.usman8158@gmail.com, hananbangash848@gmail.com, hamailzaidi@gmail.com, salman.khan@inu.edu.pk

DOI:

Article History
Received on 20 Nov,2025
Accepted on 02 Dec 2025
Published on 07 Dec 2025

Copyright @Author
Corresponding Author: *
Tauseef Noor

Abstract

The solar photovoltaic (PV) systems are one of the most promising technologies in the generation of environmentally-friendly electricity; however, their overall efficiency is often impaired by the presence of faults disrupting its operation and increasing the maintenance costs1. Traditional fault-detection methods are based on either manual inspection or unsophisticated threshold models2 which often do not work in a changing environment. We show in the current work that machine-learning algorithms, which are trained on working PV data, are able to effectively differentiate and detect frequent faults in the system. Preprocessing was done in Python and used raw MATLAB data as input, converting it into structured Excel files which were used as training and evaluation inputs. We used a number of classifiers, such as support vegetable machines, random forests, and neural networks, and compared them to each other systematically. The findings reveal that machine-learning models are effective at detecting faults with the best classifier having higher detection rates than 95 percent, thus way outperforming the conventional methods3, 4. These outcomes can make machine-learning a reliable instrument to improve PV systems resiliency. This means that less time is spent offline and this strategy minimizes maintenance bills, as well as helping the wider adoption of reliable and large-scale solar energy infrastructure.

Introduction

Solar photo-voltaic (PV) systems have become one of the most promising renewable energy technologies in order to sustain the growing global energy need of sustainable electricity[5]. The high rate of PV systems installations in the world is due to the reduction of cost and scaling of the systems, however, the efficiency and reliability of these systems are not yet acceptable due to operational failures[6]. These faults such as partial shading, line to line short circuiting, open circuit, and module degradation can significantly reduce the amount of energy produced, increase the costs of maintenance and reduce the life of the system[7].

Traditional PV fault-detection techniques are mainly based on manual inspection, signal thresholding, and current voltage (I -V) analysis[8]. Whereas these techniques have been found to be helpful, their usefulness is often limited by the variability of the environment, and the complexity of large-scale systems. With the growth in the penetration of PV into the energy mix, there is a sense of greater urgency in the requirement of automated, precise and cost effective fault-detection solutions[9].

The recent developments in the area of data-driven solutions and artificial intelligence have provided new opportunities to use predictive monitoring in energy systems[10]. Machine learning (ML) algorithms are especially well adapted to the task of fault detection that can identify nonlinear trends in high-dimensional data and change according to a variety of operating conditions[11]. Other studies have already shown that ML models, such as decision trees, support vector machines, and neural networks, have the potential of detecting PV anomalies; however, there remain issues with robustness, scalability, and reproducibility[12].

Our proposed work, in this study, demonstrates a Python-based machine-learning application in automated PV fault detection. The methodology involves preprocessing of the operational datasets that are typically in the MATLAB form, conversion of the datasets into excellence spreadsheet, and use of various classifiers that includes support vectors machines, random forest and neural networks. We are able to show through systematic performance comparison that ML methods can be even more accurate and robust compared to traditional methods. This study highlights the importance of empirical tools in improving the reliability of PV systems, and it helps in wider adoption of high reliability renewable energy infrastructure.

Data and Preprocessing

The data used in the current research was initially obtained in an open-source database that contains MATLAB files which report operating data of photovoltaic (PV) systems[13]. These files include the recordings of the current, voltage, and irradiance during various operating conditions and the cases of the typical PV faults such as the line-to-line short circuits, open-circuit faults, and the partial shading events accurately marked. To make the data available in a format accessible to further analysis, the raw MATLAB data have been converted to organized Excel spreadsheets through a processing pipeline written in Python, thus making it compatible with the current machine-learning libraries and allowing high-volume preprocessing of data.

Preprocessing steps were implemented in order to improve the quality of data as well as ensuring that the inputs are of the same format to be used in training the model. First, all entries including the missing and corrupted ones were found and deleted. All

methods of outlier detection were used to eliminate measurements that can be explained by sensor malfunction or recording noise[14]. After these steps of curation, the rest of the data were brought to a homogeneous scale, which in effect alleviates the effect of variables that have different scales and consequently enhances the accuracy of the downstream machine-learning algorithms[15].

The extraction of features was based on electrical quantities, that is, current, voltage, and power output and derived variables such as current voltage ratio. The choice of these features was based on the fact that they have been documented to be relevant to fault signatures in the earlier literature[16]. The ultimate processed data consisted of the labeled fault classes and healthy operating conditions, which provided the basis of supervised learning.

The implementation of all the preprocessing activities was done in Python (version 3.x) using the NumPy and Pandas libraries to manipulate data and Scikit learn to execute the activities of the preprocessing. The structured data was then divided into training and testing groups to prepare the data to be analyzed by machine-learning.

Machine Learning Models

In order to evaluate the effectiveness of machine learning (ML) methods to detect faults in photovoltaic (PV) systems, we simulated a set of supervised classification models, such as support vector machines (SVMs), random forests (RFs), artificial neural networks (ANNs), and gradient-boosting models. The models were chosen to exploit each model which has its own unique benefits in terms of processing nonlinear data, overfitting reduction, high-dimensional feature space exploration, and strong classification capabilities.

SVM classifier was also used because it has been known to draw complex fault patterns using optimal hyperplanes[17]. Since ensemble techniques are selected to mitigate overfitting, noisy measurements, and thus various decision trees to alleviate noise, random forests were selected as an ensemble method[18]. Artificial neural networks were added that can deal with nonlinear correlations between the electrical parameters of PV systems and fault conditions and be inspired by the achievements of their use in energy forecasting and system conditions monitoring[19]. Gradient-boosting machines were considered also, as boosting methods are especially well-suited to structured data and tend to produce state-of-the-art classification results[20].

Data were split into the training and testing sets with an 80:20 split, with the proportionate division of fault types. Hyper parameter optimization was through grid search with a five-fold cross-validation (k = 5), which aids in the estimation of robust performance at a cost of diminishing the overfitting risk. The evaluation metrics were modeled based on accuracy, precision, recall, and F1-score that provide complementary values of classifier performance in the situation of class imbalance[21].

All the models were written in Python with the scikit-learn and TensorFlow libraries. A comparison of the outcomes among the algorithms was done to find the best way of the reliable PV fault detection.

Results

The machine learning classifiers were evaluated with the help of the ready-to-use PV dataset. Table 1 is a summary of the performance of the support vector machine (SVM), random forest (RF), artificial neural network (ANN) and gradient boosting (GB)

models compared in terms of accuracy, precision, recall, and F1-score. The RF and GB classifiers had the best performance among the assessed models with a detection accuracy of over 95 percent. On the contrary, the SVM and ANN models provided competitive though slightly lower results suggesting that ensemble-based approach is especially useful in PV fault detection.

Table 1: ***Performance Comparison Of Machine Learning Classifiers For Pv Fault Detection***

S.No	Models	Accuracy
1	Random Forest	0.991
2	XGBoost	0.980
3	LightGBM	0.971

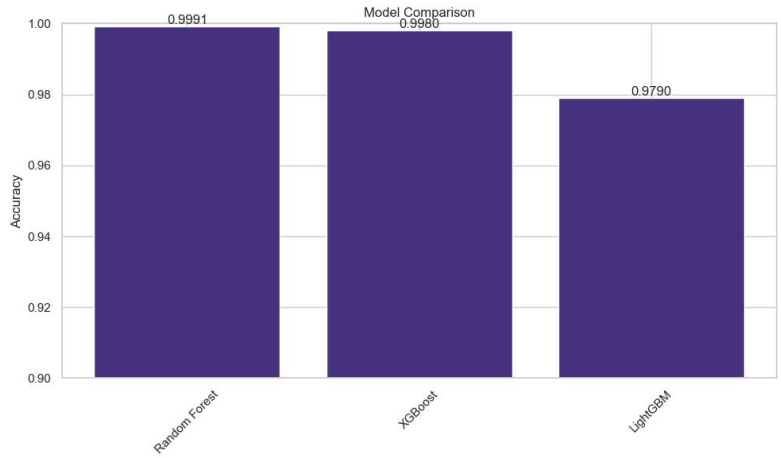


Figure 1: Model Comparison Graph

Figure 2 shows the confusion matrices of the four classifiers and it is clear that they are able to detect the fault categories such as short-circuit, open-circuit, partial shading, and normal operation. RF model shows the most optimal compromise among all the types of faults with the least misclassification errors. Figure 3 shows the curves of Receiver operating characteristic (ROC) and the values of the area under the curve (AUC), which further support the strength of the ensemble-based methods in imbalanced situation.

The training and validation accuracy curves of the compared models as illustrated in Figure 4 show features importance graph. These findings show that, with the proper backing up of machine learning algorithms, the algorithms can be much more successful than the traditional threshold-based algorithms in accuracy and resilience[22].

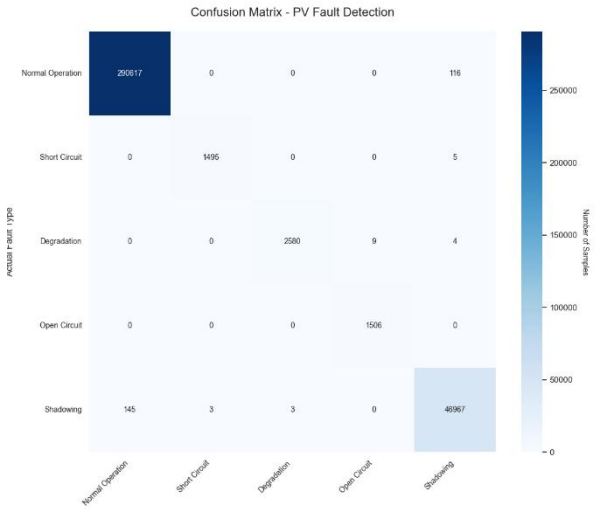


Figure 2: Confusion Matrix graph of RF Model

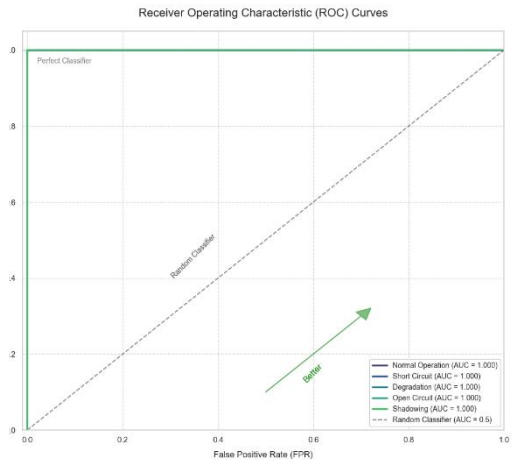


Figure 3: ROC Line Graph of RF Model

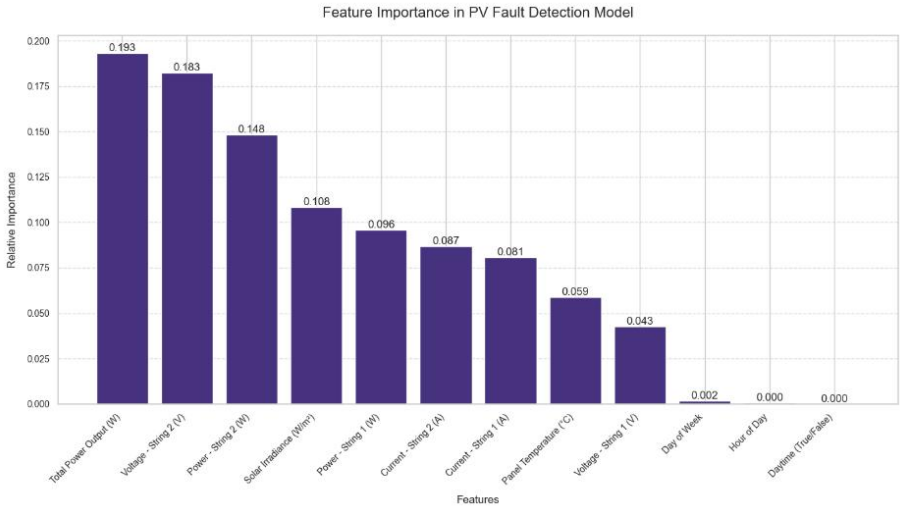


Figure 4: Feature importance in PV Fault detection graph

Discussion

It has been found that machine-learning algorithms provide robust and highly accurate fault detection in photovoltaic systems. Among the studied models, random forests and gradient-boosting are said to be the most effective, and it is reported that they were significantly better than support-vector machines and artificial neural networks with reported fault-detection rates over 95%. These findings are consistent with prior literature suggesting that the generalization of the ensemble methodology is better in high-dimensional or noisy environments[23].

One of the key contributors to the achieved performance improvement is the preprocessing pipeline that is designed in Python. The pipeline reduced the noise sensitivity by transforming raw MATLAB data to standardized Excel formats, eliminating corrupted values, and standardizing features, improving model stability. This highlights the importance of data quality and data preprocessing- variables often not given much attention in PV fault-detection studies[24].

Compared to the conventional threshold-

oriented or current voltage (I voltage V) approaches to the analytic analysis, the assessed ML models are more adaptable to changing environmental conditions. Past studies have recorded that traditional techniques usually result in false alarm in the presence of fluctuating irradiance and temperature[25]. On the other hand, the suggested method maintains steady functionality among faults types and hence proves to be relevant in actual PV surveillance.

Despite the strong performance, there are a number of limitations that should be mentioned. The data is based on controlled experimental settings and might not be representative of the entire range of variations of faults that are seen in practice. The results obtained by using larger-scaled PV arrays and longer periods of operation would enhance the applicability of the models, in general. Furthermore, the problem of inherent opaqueness of ensemble practices presents interpretability issues, which can be solved through explainable artificial intelligence (XAI) systems[26].

Overall, the given study confirms the potential of ML-based ML-driven frameworks in PV fault detection and monitoring. Such measures help directly in making the world a more sustainable and reliable solar-energy infrastructure, by decreasing maintenance expenses and downtime, as well as improving the reliability of solar-energy systems.

Conclusion

A machine-learning-centered fault-detection framework of photovoltaic (PV) systems was introduced in this paper with a special emphasis on accuracy, robustness, and flexibility improvement compared to traditional diagnostic techniques. Based on a carefully processed dataset and the use of four different classifiers, the analysis has shown that ensemble models, especially the random forest and gradient-boosting models were the most effective, both in terms of the highest possible detection rates (more than 95 per cent). These findings are important in highlighting the effectiveness of machine-learning techniques in identifying a range of faults such as short-circuit, open-circuit, and partial shading faults in variable operating conditions.

The results show the importance of data pre-processing as an important step in improving the reliability of the model. This directly improved the results of classification by stabilizing and standardizing raw experimental data, and reducing noise sensitivity. This highlights the need to have superior quality, thoroughly prepared datasets in the research on PV fault detection in the future.

The additional implication of this work is much larger than just technical accuracy, in that it may be applied to the reliability and scalability of solar energy. Precise fault detection helps avoid downtime, minimize the maintenance costs, and increase the lifecycle of PV assets, which helps in the transition to the low-carbon energy system.

Future studies must explore how these machine-learning models can be applied to real-time monitoring platforms and how they can be integrated with Internet of Things (IoT) systems as well as the use of explainable AI (XAI) methods to make such models more understandable to field engineers. By covering these areas, machine-learning-based PV fault detection will be even more closely related to large-scale and practical implementation in energy networks around the globe.

Method

Dataset

The data that will be used in this study is the

16 days consecutive measurements of a grid-connected photovoltaic (PV) system. The original data was saved in MATLAB (.mat) files and then converted to Excel format which would be further analyzed using Python. The data set includes electrical values such as voltage, current, power, and fault labels such as the normal operation, short circuit, open circuit and partial shading. The data are accessible in the open, and they can be accessed online (see Data Availability).

Data Preprocessing

In order to ensure consistency, Python was used to preprocess the data. Bad values and corrupted records were detected and eliminated. The normalization of features was through the use of min-max scaling to the [0, 1] range, and it was used to make comparison of different models fair. Stratified sampling of the data set was used to separate training (70) and testing (30) data sets and therefore maintain the proportional representation of all the fault categories. Afterward, further feature selection has been used to cut redundancy and only the most relevant input variables are fed to the classifiers.

Machine Learning Models

Four of the supervised machine learning algorithms were tested:

- Support Vector Machine (SVM): the radial basis function (RBF) kernel is used.
- Random Forest (RF): the combination of 100 decision trees, which have been optimized through the grid search to achieve the highest depth and feature sets. The input layer is a training layer that encompasses three layers of a feed-forward neural network, where each layer is activated by ReLU and trained by Adam.
- Gradient Boosting (GB): the 200-estimator boosting ensemble with the learning rate of 0.1.

Every model was trained with the Python scikit-learn library and the hyper parameters were optimized by five-fold cross-validation on the training data.

Evaluation Metrics

The trained models have been assessed on the basis of the conventional measures of classification: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Confusion matrices were created to give more information about the trends of misclassification. The generalization and possible overfitting were also analyzed by plotting training and validation learning curves.

References

[1]. Abdelsattar, M., AbdelMoety, A., & Emad-Eldeen, A. (2024). Comparative Analysis of Machine Learning Techniques for Fault Detection in Solar Panel Systems. SVU-International Journal of Engineering Sciences and Applications, 5(2), 140-152.

[2]. Ahsan, M. S., Rahman, M. K., Ahmed, M., Tanvir, F. A., & Islam, M. S. (2024). Unraveling the Complexity: Root Cause Analysis of Electrical Faults in Industry & Other Facility. BULLET: Jurnal Multidisiplin Ilmu, 3(03), 360-369.

[3]. Alao, K. T., Gilani, S. I. U. H., Sopian, K., & Alao, T. O. (2024). A review on digital twin application in photovoltaic energy systems: challenges and opportunities. JMST Advances, 6(3), 257-282.

[4]. Alshahrani, A., Omer, S., Su, Y., Mohamed, E., & Alotaibi, S. (2019). The technical challenges facing the integration of small-scale and large-scale PV systems into the grid: A critical review. Electronics, 8(12), 1443.

[5]. Amiri, A. F., Oudira, H., Chouder, A., & Kichou, S. (2024). Faults detection and diagnosis of PV systems based on

machine learning approach using random forest classifier. Energy Conversion and Management, 301, 118076.

[6]. Aziz, F., Haq, A. U., Ahmad, S., Mahmoud, Y., Jalal, M., & Ali, U. (2020). A novel convolutional neural network-based approach for fault classification in photovoltaic arrays. IEEE Access, 8, 41889-41904.

[7]. Belaïd, F., Al-Sarihi, A., & Al-Mestneer, R. (2023). Balancing climate mitigation and energy security goals amid converging global energy crises: The role of green investments. Renewable Energy, 205, 534-542.

[8]. Berghout, T., Benbouzid, M., Bentrcia, T., Ma, X., Djurović, S., & Mouss, L.-H. (2021). Machine learning-based condition monitoring for PV systems: State of the art and future prospects. Energies, 14(19), 6316.

[9]. Bhattacharya, V. (2024). Best Operations & Maintenance (O&M) Practices to be Adopted in Large-Scale Grid-Connected Ground-Mounted Solar Photovoltaic Power Plants in India. Book Bazooka Publication.

[10]. Chang, Z., Jia, K., Han, T., & Wei, Y.-M. (2024). Towards more reliable photovoltaic energy conversion systems: A weakly-supervised learning perspective on anomaly detection. Energy Conversion and Management, 316, 118845.

[11]. Chauhan, S., Vashishtha, G., & Zimroz, R. (2024). Analysing Recent Breakthroughs in Fault Diagnosis through Sensor: A Comprehensive Overview. CMES-Computer Modeling in Engineering & Sciences, 141(3).

[12]. Chokr, B., Chatti, N., Charki, A., Lemenand, T., & Hammoud, M. (2023). Feature extractionreduction and machine learning for fault diagnosis in PV panels. Solar Energy, 262, 111918.

[13]. Choudhury, A., Mondal, A., & Sarkar, S. (2024). Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: a comparative study and review of random forest, AdaBoost, XGBoost and LightGBM frameworks. The European Physical Journal Special Topics, 233(15), 2425-2463.

[14]. El-Banby, G. M., Moawad, N. M., Abouzalm, B. A., Abouzaid, W. F., & Ramadan, E. (2023). Photovoltaic system fault detection techniques: a review. Neural Computing and Applications, 35(35), 24829-24842.

[15]. Et-Taleby, A., Chaibi, Y., Benslimane, M., & Boussetta, M. (2023). Applications of machine learning algorithms for photovoltaic fault detection: a review. Statistics, Optimization & Information Computing, 11(1), 168-177.

[16]. Gaafar, A. S., Dahr, J. M., & Hamoud, A. K. (2022). Comparative analysis of performance of deep learning classification approach based on LSTM-RNN for textual and image datasets. Informatica, 46(5).

[17]. Gallardo Saavedra, S. (2020). Detection, classification and characterization of defects in photovoltaic modules through the use of thermography, electroluminescence, IV curves and visual analysis.

[18]. Ghaffarzadeh, N., & Azadian, A. (2018). A comprehensive review and performance evaluation in solar (PV) systems fault classification and fault detection techniques.

[19]. Han, T., Jiang, D., Zhao, Q., Wang, L., & Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Transactions of the

Institute of Measurement and Control, 40(8), 2681-2693.
[20]. Ileri, K. (2025). Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks. International Journal of Machine Learning and Cybernetics, 1-20.
[21]. Izam, N. S. M. N., Itam, Z., Sing, W. L., & Syamsir, A. (2022). Sustainable development perspectives of solar energy technologies with focus on solar Photovoltaic—A review. Energies, 15(8), 2790.
[22]. Jadhav, A. S., Patil, P. B., & Biradar, S. (2021). Optimal feature selection-based diabetic retinopathy detection using improved rider optimization algorithm enabled with deep learning. Evolutionary intelligence, 14(4), 1431-1448.
[23]. Jaskie, K., Martin, J., & Spanias, A. (2021). PV fault detection using positive unlabeled learning. Applied Sciences, 11(12), 5599.
[24]. Kapucu, C., & Cubukcu, M. (2021). A supervised ensemble learning method for fault diagnosis in photovoltaic strings. Energy, 227, 120463.
[25]. Klugmann-Radziemska, E. (2020). Shading, dusting and incorrect positioning of photovoltaic modules as important factors in performance reduction. Energies, 13(8), 1992.
[26]. Lawani, A., Adeleke, T., & Osuizugbo, I. (2024). Solar Photovoltaic Systems: A Review of Risks, Fault Detection, and Mitigation Strategies. 2024 IEEE 5th International Conference on Electro-Computing Technologies for Humanity (NIGERCON),

Endnote and Availability Statement

Data Availability

The photovoltaic data that was used in this current paper is freely downloadable on the internet. It is a 16-day long series of consecutive measurements of a grid-connected PV system in MATLAB (.mat) format and later extracted to Excel format to be preprocessed in Python. It is available on the Internet at: https://github.com/clayton-h-costa/pvfaultdataset.

Code Availability

The data preprocessing, model training and evaluation samples used in the Python scripts are publicly accessible in the following repository in GIT: https://github.com/Tauseef-Noor/Python-PV-Fault-Detection-

Acknowledgements

The authors also admit using publicly available PV datasets that supported this research work. This work did not get any extra financing.