International Journal of Basic and Applied Sciences, 14 (2) (2025) 462-469

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/xxh9h294
Research paper

Early Detection of Anomalies in Photovoltaic Module Strings Using Decision Trees for MPPT Solar Charger Systems

D. Ramesh Reddy [1] *, T. S. Saravanan [2], P. Subhashini [3], Saravana Selvan [4], Sonia Maria D'Souza [5], Thiyagesan M [6], M. DuraiRaj [7], Raja-sekhara Babu L [7]

[1] *Department of Electronics and Communication Engineering, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana 500090, India*
[2] *Department of Electrical and Electronics Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu 602105, India*
[3] *Department of Information Technology, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, Tamil Nadu 600062, India*
[4] *School of Professional Engineering, Manukau Institute of Technology, Tech Park Campus, Auckland 2104, New Zealand*
[5] *Department of Artificial Intelligence and Machine Learning, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India*
[6] *Department of Electrical and Electronics Engineering, R.M.K Engineering College, Kavaraipettai, Tamil Nadu 601206, India*
[7] *Department of Computer Science Engineering, Bharathidasan University, Tiruchirappalli, Tamil Nadu 620024, India*
**Corresponding author E-mail: rameshreddy_d@vnrvjiet.in*

Received: April 18, 2025, Accepted: June 18, 2025, Published: June 30, 2025

Abstract

This study presents a decision tree (DT) based machine learning approach to detect early anomalies and faults in solar maximum power point tracking (MPPT) integrated photovoltaic (PV) module circuits. A four-panel array, built using a single diode model, is simulated to generate a synthetic, balanced dataset for training and evaluation. Among the different models tested, including neural networks (NN) and support vector classifiers (SVC), the DT model showed the best performance in precision and recall across all anomaly labels while maintaining simplicity and low computational cost. Currently, the model is limited to four module configurations, and the use of synthetic data may lead to overfitting when applied to real-world scenarios. Nevertheless, the methodology can be adapted to grid configurations with other known parameters. This work provides a practical basis for integrating early anomaly detection systems into PV installations, increasing operational efficiency and reducing maintenance costs.

Keywords: *Machine Learning; Photovoltaic Arrays; Decision Tree Models; MPPT Solar Chargers.*

1. Introduction

In recent years, the use of computational models for data classification has seen a large upswing, driven by greater data availability and advancements in machine learning techniques (Srivastava, 2019; Taylor & Letham, 2018). Classification is key to many practical uses—that is, how data points are assigned into defined categories—from medical diagnostics and financial fraud detection to image recognition and natural language processing (Que et al., 2019; Hu et al., 2020).

This part explains how to train and check different supervised machine learning classifiers using a synthetically made and prepared dataset. The goal is to find out the best algorithm for P and recall in different classes. The dataset, which has 110,000 samples, is arranged to show an even spread among 11 different categories (Toshniwal et al., 2020; Pereira & Silveira, 2018; Tsai et al., 2020). To facilitate computational processing and aid multiclass classification, these categories initially specified qualitatively, such as healthy, H1S3, Rs4, were mapped to numeric labels ranging from 0 to 10.

To ensure that the models are trained on data representative of the entire feature space and prevent overfitting, the dataset was split into training and evaluation sets containing 88,000 and 22,000 samples, respectively. Unlike the conventional funnel approach, which divides the data into training, validation, and testing triads, this study takes a dual-part split. During the training phase, k-fold cross-validation is employed for internal model evaluation and parameter tuning. This approach allows trainers to maximize the training set while reserving a separate evaluation set available to all models for unbiased performance assessment (Iyengar et al., 2018).

For training and evaluation, six classifying models were used: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Neural Networks (NN). With the use of the Scikit-Learn library, each model went through hyperparameter optimization with the GridSearchCV function. The optimization process highlighted a granular search within a set listing of parameters that included weighted P and weighted recall as important performance indicators. These indicators were

Copyright © D. Ramesh Reddy et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

chosen to ensure class imbalance issues were mitigated and to ensure model performance is not overly favorable to the dominating majority class.

The proposed methodological framework in this study improves the detection of photovoltaic (PV) anomalies through the comparative evaluation of six classification models on synthetically generated datasets. Among them, the Decision Tree (DT) classifier showed the highest efficiency with perfect precision and recall values of 0.984. Unlike traditional PV anomaly detection methods—typically based on unsupervised learning, signal processing, or complex black box models—DT enables high interpretability and scalability. Its rule-based structure supports real-time classification and can quickly adapt to changing PV configurations or additional fault classes. Computationally intensive models such as K Nearest Neighbor (KNN) that scale with larger datasets enable DT to achieve a fast, low-latency solution suitable for edge detection. Moreover, it outperformed Support Vector Classifiers (SVC) and Neural Networks (NN), which struggled with accuracy due to their sensitivity to feature scalability and stability. Overall, the DT-based approach combines robust classification performance with practical advantages in performance and scalability, making it well-suited for real-time PV fault diagnosis.

Below this paragraph, each model is presented with its specific setup and enhancement and with its performance results. Classification reports are illustrative of important model performance metrics: Precision (P), recall (RC), F1 score (F1), and support (S), along with the total of 11 classes.

2. Literature review

To contextualize these results and ground the methodology in current research, the following literature review summarizes relevant studies in PV system diagnostics and fault detection using machine learning techniques. This provides a comparative basis for understanding the strengths and limitations of the models evaluated in this work.

Hempelmann et al. (2020) analyzed different approaches to unsupervised anomaly detection using real PV system monitoring datasets. They evaluated the performance of techniques such as Isolation Forest and One-Class SVM relative to different faults. The study noted the challenges of creating reliable ground truths based on unlabeled datasets while also attempting to achieve an optimal balance between sensitivity and false-positive rates. This work demonstrates the difficulty of accurately modeling almost all fault conditions in solar panels with generic algorithms.

Hariri, Kind, & Brunner (2019) described the Extended Isolation Forest (EIF) as an enhancement to the standard Isolation Forest method. EIF goes further by using random partitioning to make scoring less biased on multivariate and skewed datasets. This improvement increases its suitability for industrial applications such as PV system monitoring, where data complexity and distribution changes are prevalent.

Wang, Paynabar, & Pacella (2021) developed a framework for real-time, automated anomaly detection in photovoltaic systems using thermographic images and low-rank matrix decomposition. Their strategy successfully detected thermal anomalies like hotspots and panel failures, allowing for faster problem detection. This technology also uses image processing techniques, making it suitable for remote monitoring systems in solar farms where manual checks are costly.

Balzategui, Eciolaza, & Maestro-Watson (2021) used Generative Adversarial Networks (GANs) for defect detection in solar cells. The model not only detected anomalies but also provided self-sufficient labels, which significantly reduced the manual effort involved in the quality control process. Their method proved to have considerable precision in detecting small visual anomalies, making it a strong candidate for automated inspection in solar panel production.

Anomaly detection was studied by Benninger, Hofmann, & Liebschner (2020), as the authors performed a comparative analysis of PV systems for different machine learning algorithms. They effectively identified anomalies, thanks to the analysis of output profiles from their neighboring systems, showing the benefits of contextual and peer-to-peer learning in distributed energy architectures. Using the same methodology works well in large solar fields where the panels are subjected to very similar conditions.

Mulongo et al. (2020) used NN and classical ML techniques to detect anomalies in power generation plants. They focused on data preprocessing, feature engineering, and model fusion to increase prediction performance. Combining hints from statistics with deep learning models, the hybrid method led to better discovery of rare events among large streaming data.

In developed references of access management in big data federation (Awaysheh et al., 2020), the authors introduced a reference architecture of the proposed approach aiming at anomaly-awareness. They addressed privacy and access concerns in large, federated data systems and emphasized the role of access management in preventing atypical behaviors like data breaches or unauthorized access. The implication of this research extends to IoT-enhanced PV and smart grid-based systems, where maintaining data integrity is critical.

The authors of this study extended this work by proposing a blockchain-based multi-factor authentication for cloud-enabled IoV (Kebande et al. 2021). Even though their focus is on vehicle networks, their research is closely related to a renewable energy monitoring system that includes the usage of IoT and cloud technologies. Their approach allows for reputable anomaly detection of authentication activities, providing a secure and decentralized framework.

Son et al. (2020) conducted a significant study on the impact of atmospheric particulate matter (PM) on the performance of photovoltaic (PV) systems across the Republic of Korea. By analyzing multi-year datasets including solar irradiance and PM10/PM2.5 concentrations from 2015 to 2018, the researchers identified a clear inverse relationship between PM levels and PV output. Their findings indicated that power generation efficiency can decrease by approximately 5% during high pollution days, with greater losses observed in urban and industrial regions. This degradation was also found to vary seasonally, peaking in winter due to atmospheric inversion layers that trap pollutants. The study underscores the necessity of integrating air quality data into solar forecasting models, especially in areas with significant pollution loads, to enhance PV output predictions and maintenance planning.

In a complementary environmental context, Dajuma et al. (2016) examined the sensitivity of PV panel efficiency to weather conditions and dust accumulation in West Africa, with experimental data collected from Niamey, Niger (arid climate), and Abidjan, Côte d'Ivoire (humid climate). Their comparative study revealed that dust accumulation has a more pronounced effect in dry, desert-like conditions, where efficiency losses reached up to 26% due to heavy soiling. Conversely, the more humid Abidjan experienced lower efficiency losses (~11%) due to frequent rainfall, which helped naturally clean the panels. The study highlighted the role of climate in determining maintenance frequency and advocated for tailored cleaning protocols based on environmental conditions. These findings are particularly valuable for PV system deployment and maintenance in regions with high dust prevalence and limited rainfall.

Building on the operational perspective of PV systems, Padmanathan et al. (2018) provided a comprehensive survey on the integration of solar PV systems into industrial and commercial energy infrastructures. The authors reviewed a wide range of technological, economic, and policy-related studies to map the landscape of PV adoption in large-scale non-residential sectors. The review emphasized the growing adoption of PV systems driven by declining technology costs and rising energy demand. However, it also identified critical challenges including intermittency, high initial capital costs, and regulatory barriers. The authors noted that the successful integration of PV in such

sectors relies on enabling technologies like hybrid inverters, battery energy storage systems, and advanced energy management systems (EMS). Case studies illustrated how industries could reduce operational costs and environmental impact through well-planned solar integration strategies.

Collectively, these studies offer a holistic view of environmental and operational factors affecting PV system performance. While Son et al. (2020) and Dajuma et al. (2016) focus on environmental impacts—particulate matter and dust, respectively—Padmanathan et al. (2018) address the broader context of system integration and practical deployment in the industrial domain. Together, they underscore the multifaceted nature of PV system performance and provide valuable insights for enhancing efficiency, reliability, and scalability in various geographic and functional contexts.

3. Methodology

As a first step in training and evaluating computational classification models, the database was pre-processed. It is important to highlight that the synthetic nature of the data allows for the most appropriate format for data management, balanced data. Therefore, the only change made to the database, apart from randomly organizing the data, was to change the qualitative labels to numerical values from 0 to 10, as follows (Table 1):

Table 1: Change to Numerical Labels

Original Label	Numeric Label
healthy	0
H1S3	1
H2S2	2
H3S1	3
H1SC3	4
H2SC2	5
H3SC1	6
Rs1	7
Rs2	8
Rs3	9
Rs4	10

The data is then divided into 20% for evaluation and 80% for training, leaving a set with 22,000 evaluation data points and 88,000 training data points. This is done to use the training data to both train and test the models and optimize each model. It is not divided into the usual three sets: train, test, and evaluation, as k-fold is used to test the models during their training phase. Six different classification models were trained and evaluated using the sckit Learn library and its GridSearch function. To evaluate different parameters for each model, two evaluation criteria were used: weighted P and weighted RC. Below (Table 2) are the trained models and the different parameters tested for each one.

Table 2: Grid Search, Models and Parameters

Model	Evaluation	Test Parameter	Parameters
LR	NA	Weighted P	NA
DT	5-Fold	Weighted P Weighted RC	criteria: Gini, Entropy, Log_loss max_depth: [140,150] 10 data points
RF	5-Fold	Weighted P Weighted RC	criterion: Gini, Entropy, Log_loss n_estimators: [100,200] 10 data
K Nearest Neighbor	5-Fold	Weighted P Weighted RC	n_neighbors: [1,10] 10 data points Leaf_size: [20,40] 5 data points p: [1,2] 2 data points weights: uniform, distance
SVC	5-Fold	Weighted P Weighted RC	kernel: poly degree: [1,2] 2 data points
Neural Network	5-Fold	Weighted P Weighted RC	activation: relu solver: adam hidden_layer_sizes: (3 layers of 20 neurons, 4 layers of 90 neurons)

4. Results

After the grid search, the following results were obtained for each trained and evaluated model. Each model was evaluated using the classification report, which shows the P, RC, and F1 for each class. The evaluation set, separated from the data preprocessing, was used to calculate the results (Ypred) calculated by the model and compare the predicted results with the evaluation data set of 22,000 (Yeval).

4.1. Logistic regression

As the first method trained and evaluated, LR showed a P and weighted RC of 0.35, demonstrating that it is a model with poor performance on the evaluation data set (Table 3).

Table 3: LR Classification Report

Class	P	RC	F1	S
Healty	0.47	0.60	0.53	2000
H1S3	0.50	0.43	0.46	2000
H2S2	0.47	0.62	0.53	2000
H3S1	0.31	0.47	0.37	2000
H1SC3	0.52	0.60	0.56	2000
H2SC2	0.49	0.34	0.40	2000

H`3SC1	0.38	0.09	0.15	2000
Rs1	0.18	0.21	0.19	2000
Rs2	0.19	0.14	0.16	2000
Rs3	0.23	0.16	0.19	2000
Rs4	0.11	0.14	0.12	2000
Accuracy			0.35	22000
Macro Average	0.35	0.35	0.33	22000
Weighted Average	0.35	0.35	0.33	22000

4.2. Decision tree

The DT model performs excellently, allowing scores above 95% across all labels. The Confusion Matrix for the Decision Tree is shown in Figure 1 and Table 4.

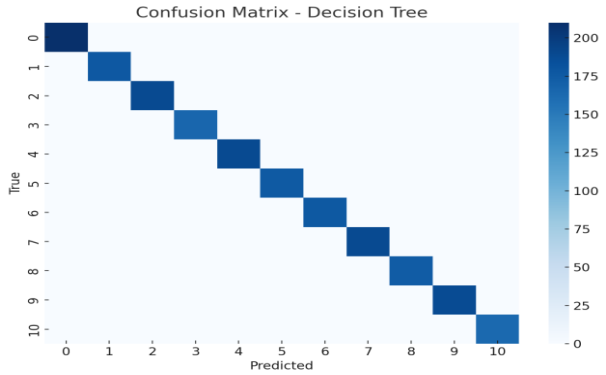


Fig. 1: Confusion Matrix- Decision Tree.

Table 4: DT Classification Report

Class	P	RC	F1	S
Healty	0.98	1.00	0.99	2000
H1S3	1.00	1.00	1.00	2000
H2S2	1.00	1.00	1.00	2000
H3S1	1.00	1.00	1.00	2000
H1SC3	1.00	1.00	1.00	2000
H2SC2	1.00	1.00	1.00	2000
H3SC1	1.00	1.00	1.00	2000
Rs1	0.98	0.96	0.97	2000
Rs2	0.97	0.97	0.97	2000
Rs3	0.97	0.96	0.96	2000
Rs4	0.97	0.98	0.98	2000
Accuracy			0.99	22000
Macro Average	0.99	0.99	0.99	22000
Weighted Average	0.99	0.99	0.99	22000

Best DT parameters obtained using grid search: {'criterion': 'entropy', 'max_depth': 150} with an average P of 0.984 and an average RC of 0.984.

4.3. Random forest

It is important to remember that RF is a model that applies several DTs, and given its excellent performance, scores above 96% across all labels are obtained (Table 5).

Table 5: RF Classification Report

Class	P	RC	F1	S
Healty	0.98	1.00	0.99	2000
H1S3	1.00	1.00	1.00	2000
H2S2	1.00	1.00	1.00	2000
H3S1	1.00	1.00	1.00	2000
H1SC3	1.00	1.00	1.00	2000
H2SC2	1.00	1.00	1.00	2000
H3SC1	1.00	1.00	1.00	2000
Rs1	0.99	0.97	0.98	2000
Rs2	0.98	0.98	0.98	2000
Rs3	0.97	0.96	0.97	2000
Rs4	0.98	0.98	0.98	2000
Accuracy			0.99	22000
Macro Average	0.99	0.99	0.99	22000
Weighted Average	0.99	0.99	0.99	22000

Best RF parameters obtained using grid search: {'criterion': 'entropy', 'n_estimators': 133} with an average P of 0.987 and an average RC of 0.987.

4.4. K nearest neighbor

This model presents in its classification report on the evaluation data an accuracy and RC of over 97% in all classes (Table 6).

Table 6: K Nearest Neighbor Classification Report

Class	P	RC	F1	S
Healty	0.99	1.00	0.99	2000
H1S3	1.00	0.99	1.00	2000
H2S2	1.00	0.99	1.00	2000
H3S1	1.00	1.00	1.00	2000
H1SC3	1.00	1.00	1.00	2000
H2SC2	0.99	1.00	1.00	2000
H3SC1	1.00	0.99	1.00	2000
Rs1	0.99	0.98	0.99	2000
Rs2	0.99	0.99	0.99	2000
Rs3	0.98	0.98	0.98	2000
Rs4	0.98	0.99	0.98	2000
Accuracy			0.99	22000
Macro Average	0.99	0.99	0.99	22000
Weighted Average	0.99	0.99	0.99	22000

Best K Nearest Neighbors parameters obtained using grid search: {'leaf_size': 20, 'n_neighbors': 1, 'p': 2, 'weights': 'uniform'} with an average accuracy of 0.988 and an average RC of 0.988.

4.5. Support vector classifier

The trained and evaluated support vector machines did not achieve an optimal evaluation report; weighted accuracy and RC values of less than 50% are reported based on the evaluation data (Table 7). The Confusion Matrix for Neural Network is shown in Figure 2.

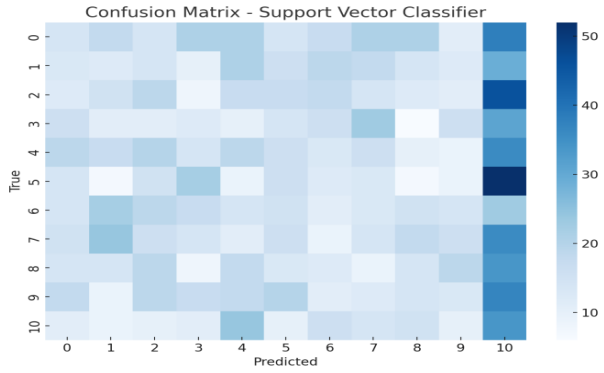


Fig. 2: Confusion Matrix- Support Vector Machine.

Table 7: SVC Classification Report

Class	P	RC	F1	S
Healty	0.62	0.63	0.62	2000
H1S3	0.98	0.22	0.36	2000
H2S2	0.50	0.47	0.49	2000
H3S1	0.44	0.57	0.50	2000
H1SC3	0.58	1.00	0.73	2000
H2SC2	0.52	0.56	0.54	2000
H3SC1	0.48	0.42	0.45	2000
Rs1	0.45	0.29	0.36	2000
Rs2	0.49	0.48	0.49	2000
Rs3	0.43	0.38	0.40	2000
Rs4	0.45	0.60	0.52	2000
Accuracy			0.51	22000
Macro Average	0.54	0.51	0.50	22000
Weighted Average	0.54	0.51	0.50	22000

Best SVC parameters obtained using grid search: {'degree': 2, 'kernel': 'poly'} with an average accuracy of 0.511 and an average RC of 0.486.

4.6. Neural network

The trained and evaluated NN did not achieve an optimal evaluation report for all classes; satisfactory accuracy and RC values are reported only for the Healthy and Degradation conditions (Table 8). The Confusion Matrix for Neural Network is shown in Figure 3.

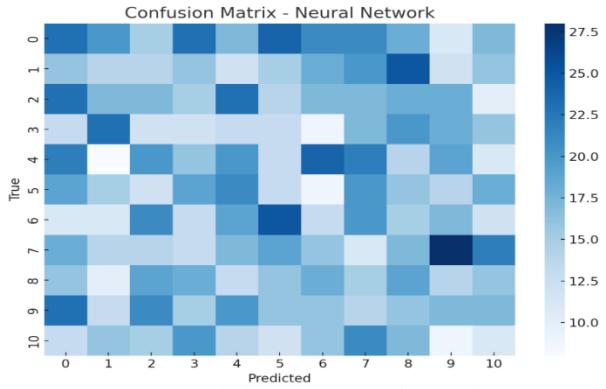


Fig. 3: Confusion Matrix- Neural Network.

Table 8: Neural Network Classification Report

Class	P	RC	F1	S
Healty	0.90	1.00	0.95	2000
H1S3	0.98	0.26	0.41	2000
H2S2	0.98	0.04	0.08	2000
H3S1	0.49	0.75	0.59	2000
H1SC3	0.58	1.00	0.73	2000
H2SC2	0.51	1.00	0.68	2000
H3SC1	0.51	0.27	0.36	2000
Rs1	1.00	0.89	0.94	2000
Rs2	0.99	0.98	0.99	2000
Rs3	0.96	0.96	0.96	2000
Rs4	0.99	0.93	0.96	2000
Accuracy			0.74	22000
Macro Average	0.81	0.74	0.70	22000
Weighted Average	0.81	0.74	0.70	22000

Best NN parameters obtained using grid search: 'activation': 'relu', 'hidden_layer_sizes': (90, 90, 90, 90), 'solver': 'adam'} with an average accuracy of 0.758 and an average RC of 0.717.

To better illustrate the performance of the model, ROC curves and confusion matrices were generated for each classifier. The decision tree (DT) model showed accurate separation between all classes, ROC curves approaching the ideal upper right corner, and little confusion in the matrix, confirming its strong classification power and low false-positive rate. In contrast, Support Vector Classifier (SVC) and Neural Networks (NN) showed broader inter-class convergence in both visualizations, indicating lower discrimination ability, especially in motion conditions, with smaller variances. These comparison tools further confirm the choice of DT as the optimal model, which improves its understanding, reliability, and suitability for real-time applications in PV system monitoring.

5. Discussion

The training and evaluation of the different selected models were performed using the Skit Learn library. Using its grid search functions (GridSearchCV), five different models were trained, as presented below:

a) DT: This is the simplest model, where its nonparametric characteristics allow classification through parameter comparisons. The best model found through the grid search was a model with a maximum of 150 levels, evaluating splitting using the entropy method. Its evaluation metrics were high, with a weighted accuracy of 0.984, indicating that the model is suitable and performs excellently on the evaluation data.

b) RF: This model involves several DTs and allows classification through parameter comparisons while improving model variability. The best model found through the grid search was a model with 133 trees, evaluating splitting using the entropy method. Its evaluation metrics were high, with a weighted accuracy of 0.987, indicating that the model is suitable and performs excellently on the evaluation data.

c) K Nearest Neighbor: The model found was based on the simplest possible approach, with a single neighbor and 20 values for each node. The model was evaluated on the data with an average accuracy of 0.988. Given the nature of this model, its scalability is limited, as its computational requirement depends largely on the number of data points.

d) SVC: The classifier support vector machine found through grid search had a degree 2 polynomial kernel as parameters. Its performance on the evaluation data was poor, obtaining a weighted accuracy of 0.511 and an average RC of 0.486. This implies that, given the nature of the data, the SVC separation principle is not sufficient for adequate anomaly classification. e) Neural Network: The classification neural network found through the grid search had as parameters 4 layers of 90 neurons with ReLU activation. Using the Adam solver, the performance on the evaluation data was not the best, obtaining a weighted P of 0.758 and an average RC of 0.717. Although the P was not the best, the model behaves adequately on certain labels, as can be seen in Table 8.

Considering the classification models evaluated, the DT and K Nearest Neighbor models stand out for their excellent performance and simplicity, both providing excellent metrics for the evaluation data. Given the characteristics and objectives of the project, the DT was chosen as the model to be developed because, due to its information flow nature and decision-making, its online implementation is relatively easy and scalable to photovoltaic systems, changing array configurations, and increasing the number of labels. It is also noteworthy that the DT model has a fast training time, unlike k k-nearest neighbor, which can significantly increase its training time as the database increases.

This study presents a promising framework for photovoltaic (PV) anomaly detection through the comparative evaluation of six supervised classifiers on synthetically generated datasets. The decision tree (DT) model achieved the highest performance with a precision and recall

value of 0.984. Unlike traditional approaches—based on unsupervised methods, signal analysis, or deep learning models—DT provides an explicit, rule-based structure suitable for real-time classification with powerful scalability and flexible PV sources.
Although these results are encouraging, they come from controlled synthetic data that do not fully reflect noise, distortion, or low disturbances in real-world PV systems. Thus, the use of more complex models, such as KNN, SVC, and Neural Networks, is recommended, but the reliability of the DT needs to be validated with operational data. Incorrect settings can cause problems, such as data gaps, storage limitations, and model drift.
In summary, although the proposed method represents a meaningful advancement in PV fault detection, it should be considered as a fundamental step. Future work should emphasize real-world challenges and explore adaptive learning to ensure robustness in dynamic PV environments.
This study presents a promising framework for photovoltaic (PV) anomaly detection through the comparative evaluation of six supervised classifiers on synthetically generated datasets. The decision tree (DT) model achieved the highest performance with a precision and recall value of 0.984. Unlike traditional approaches—based on unsupervised methods, signal analysis, or deep learning models—DT provides an explicit, rule-based structure suitable for real-time classification with powerful scalability and a flexible PV framework.
Although these results are encouraging, they come from controlled synthetic data, which do not fully reflect the fluctuations or infrequent failure modes in real-world PV systems. Thus, it is recommended to use more complex models, such as KNN, SVC, and Neural Networks, but the reliability of the DT needs to be verified with operational data. Precise fixes can cause problems, such as missing data, storage limitations, or model drift over time.
Beyond PV monitoring, this approach is promising for wider applications in materials science, such as degradation analysis of smart materials and nanocomposites, where sensor-based monitoring is possible. Similarly, in the IoT domain, the low computational cost makes the DT model ideal for smart grids, energy harvesting systems, or distributed sensor networks, where real-time fault detection and system resilience are critical.
In summary, the proposed method presents a meaningful improvement in detectable and broader anomaly detection with the potential to support interdisciplinary innovations in energy systems, materials diagnostics, and Internet communication infrastructure.
This study presents a promising framework for photovoltaic (PV) anomaly detection through the comparative evaluation of six supervised classifiers on synthetically generated datasets. Among them, the Decision Tree (DT) model showed the highest performance with a precision and recall value of 0.984. The transparent, rule-based structure allows decisions to be made and is easy to use in real-time conditions, providing a good alternative to traditional unsupervised methods or computationally intensive black box models such as neural networks or neural networks.
Importantly, the DT model is well-suited for integration into real PV systems, where robust and interpretable models are essential for final optimization. Its computational efficiency can be embedded in local microcontrollers or IoT-enabled data loggers, supporting real-time fault diagnosis without the need for cloud-based inference. The model can be linked to flow data from existing **PV sensors—typically voltage, current, brightness, and temperature modules—and modified to respond to asynchronous or noisy sensor input using preprocessing buffers or a sliding window technique. This can lead to random data loss or temporary delays, especially in large-scale PV arrays or remote installations.
However, although the performance of the model on synthetic data is promising, real-world PV environments may be affected by sensor drift, communication latency, rare failure modes, and variable operating conditions, which may affect reliability. Addressing them will require further experiments and the incorporation of online learning or model updating mechanisms.
In addition to PV failure detection, the properties of the DT model make it applicable to a wider range of contexts, such as real-time material failure monitoring and IoT-based condition diagnosis, where resource constraints and data fusion are equally critical.
The decision tree (DT) model showed strong performance on the synthetically generated PV anomaly dataset, reaching a precision and recall of 0.984, but the results should be interpreted with caution due to the risk of bias. The model captures specific patterns for synthetic data, which do not generalize well for all real-world PV systems in the presence of sensor noise, environmental variables, and unmodeled fault types. Decision trees are highly converged because they generate complex structures that fit the training data. To increase reliability and real-world applicability, future work should validate the model with real operational PV data from different environments and system configurations. Strategies such as removal and hyperparameter tuning improve the generality of online learning. In addition, robustness testing should be performed in cases where noise, gaps, or asynchronous data are present. By addressing these challenges, the proposed approach can be a reliable tool for real-time PV fault detection, condition monitoring, and broader IoT applications in energy systems.
In future work, the Decision Tree (DT) model needs to be validated with real-world PV data to evaluate its robustness against noise, sensor drift, and class imbalance. Integration with IoT-to-IoT frameworks enables low latency and fault detection at the edge, but issues such as sensor integration and real-time data processing need to be addressed. Adaptive strategies such as online or ensemble learning can improve resilience to system conditions and data variability.

6. Conclusions

In conclusion, this research project has demonstrated the feasibility of using a DT-based machine learning model for the early detection of anomalies and faults in photovoltaic module strings connected to an MPPT solar charger. By simulating the four-panel array using the Single Diode Model and generating a synthetic database, the model was trained and evaluated, obtaining excellent results in terms of accuracy and RC for each anomaly label.
Although more complex models such as NN and SVC were evaluated, the DT stood out for its simplicity and efficiency in training, outperforming the other models in terms of accuracy for all labels. However, it is important to note that this model is limited to four-panel arrays in series of the Ja Solar JAM72S20-460Mr model. It is also important to highlight that the use of synthetic balanced data does not represent the proportion of data that would occur under real-world conditions, so an overfitting of the data concerning the training dataset is assumed. However, the methodology used can be replicated for other arrays if their design and components are known.
In summary, this project has laid the groundwork for the implementation of an early anomaly detection tool in photovoltaic arrays using a DT, which will contribute to improving the efficiency and reliability of solar systems, as well as reducing maintenance and repair costs.

References

[1] Srivastava, S. (2019). *Benchmarking Facebook's Prophet, PELT and Twitter's Anomaly Detection and Automated Deployment to Cloud* (Master's thesis, University of Twente, Enschede, The Netherlands).

[2] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *American Statistician*, 72, 37–45. https://doi.org/10.1080/00031305.2017.1380080.

[3] Que, Z., Liu, Y., Guo, C., Niu, X., Zhu, Y., & Luk, W. (2019). Real-time anomaly detection for flight testing using autoencoder and LSTM. In *Proceedings of the 2019 IEEE International Conference on Field-Programmable Technology (ICFPT)* (pp. 379–382). Tianjin, China. https://doi.org/10.1109/ICFPT47387.2019.00072.

[4] Hu, D., Zhang, C., Yang, T., & Chen, G. (2020). Anomaly detection of power plant equipment using long short-term memory based autoencoder neural network. *Sensors*, 20, 6164. https://doi.org/10.3390/s20216164.

[5] Toshniwal, A., Mahesh, K., & Jayashree, R. (2020). Overview of anomaly detection techniques in machine learning. In *Proceedings of the 2020 IEEE Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 808–815). Palladam, India. https://doi.org/10.1109/I-SMAC49090.2020.9243329.

[6] Pereira, J., & Silveira, M. (2018). Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention. In *Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1275–1282). Orlando, FL, USA. https://doi.org/10.1109/ICMLA.2018.00207.

[7] Tsai, C. W., Yang, C. W., Hsu, F. L., Tang, H. M., Fan, N. C., & Lin, C. Y. (2020). Anomaly detection mechanism for solar generation using semi-supervision learning model. In *Proceedings of the 2020 IEEE Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)* (pp. 9–13). Rajpura, India. https://doi.org/10.1109/Indo-TaiwanICAN48429.2020.9181310.

[8] Iyengar, S., Lee, S., Sheldon, D., & Shenoy, P. (2018). SolarClique: Detecting anomalies in residential solar arrays. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 1–10). Menlo Park and San Jose, CA, USA. https://doi.org/10.1145/3209811.3209860.

[9] Hempelmann, S., Feng, L., Basoglu, C., Behrens, G., Diehl, M., Friedrich, W., Brandt, S., & Pfeil, T. (2020). Evaluation of unsupervised anomaly detection approaches on photovoltaic monitoring data. In *Proceedings of the 2020 47th IEEE Photovoltaic Specialists Conference (PVSC)* (pp. 2671–2674). Calgary, AB, Canada. https://doi.org/10.1109/PVSC45281.2020.9300481.

[10] Hariri, S., Kind, M. C., & Brunner, R. J. (2019). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 33, 1479–1489. https://doi.org/10.1109/TKDE.2019.2947676.

[11] Wang, Q., Paynabar, K., & Pacella, M. (2021). Online automatic anomaly detection for photovoltaic systems using thermography imaging and low rank matrix decomposition. *Journal of Quality Technology*, 1–14. https://doi.org/10.1080/00224065.2021.1948372.

[12] Balzategui, J., Eciolaza, L., & Maestro-Watson, D. (2021). Anomaly detection and automatic labeling for solar cell quality inspection based on generative adversarial network. *Sensors*, 21, 4361. https://doi.org/10.3390/s21134361.

[13] Benninger, M., Hofmann, M., & Liebschner, M. (2020). Anomaly detection by comparing photovoltaic systems with machine learning methods. In *Proceedings of the NEIS 2020, Conference on Sustainable Energy Supply and Energy Storage Systems* (pp. 1–6). Hamburg, Germany.

[14] Mulongo, J., Atemkeng, M., Ansah-Narh, T., Rockefeller, R., Nguegnang, G. M., & Garuti, M. A. (2020). Anomaly detection in power generation plants using machine learning and neural networks. *Applied Artificial Intelligence*, 34, 64–79. https://doi.org/10.1080/08839514.2019.1691839.

[15] Awaysheh, F. M., Alazab, M., Gupta, M., Pena, T. F., & Cabaleiro, J. C. (2020). Next-generation big data federation access control: A reference model. *Future Generation Computer Systems*, 108, 726–741. https://doi.org/10.1016/j.future.2020.02.052.

[16] Kebande, V. R., Awaysheh, F. M., Ikuesan, R. A., Alawadi, S. A., & Alshehri, M. D. (2021). A blockchain-based multi-factor authentication model for a cloud-enabled Internet of Vehicles. *Sensors*, 21, 6018. https://doi.org/10.3390/s21186018.

[17] Son, J., Jeong, S., Park, H., & Park, C.-E. (2020). The effect of particulate matter on solar photovoltaic power generation over the Republic of Korea. *Environmental Research Letters*, 15, 9. https://doi.org/10.1088/1748-9326/ab905b.

[18] Dajuma, A., Yahaya, S., Touré, S., Diedhiou, A., Adamou, R., Konaré, A., Sido, M., & Golba, M. (2016). Sensitivity of solar photovoltaic panel efficiency to weather and dust over West Africa: Comparative experimental study between Niamey (Niger) and Abidjan (Côte d'Ivoire). *Computational Water, Energy, and Environmental Engineering*, 5, 123–147. https://doi.org/10.4236/cweee.2016.54012.

[19] Padmanathan, K., Govindarajan, U., Ramachandaramurthy, V. K., Sudar Oli Selvi, T., & Jeevarathinam, B. (2018). Integrating solar photovoltaic energy conversion systems into industrial and commercial electrical energy utilization—A survey. *Journal of Industrial Information Integration*, 10, 39–54. https://doi.org/10.1016/j.jii.2018.01.003.