Solar Energy
And Sustainable Development
Refereed, biannual scientific journal issued by:
The Libyan Center for Solar Energy Research and Studies

Comparative Analysis of AI-Driven Machine Learning Models for Fault Detection and Maintenance Optimization in Photovoltaic Systems

Abdellahi Moulaye Rchid[1], Moussa Attia[2*], Mohamed Elmamy Mahmoud[3], Vatma Elvally[4], Zoubir Aoulmi[5], Abdelkader Ould Mahmoud[6].

[1,3,6]Applied Research Unit for Renewable Energies in Water and Environment (URA3E), University of Nouakchott, Nouakchott BP 880, Mauritania.
[2,5]Environment Laboratory, Institute of Mines, Echahid Cheikh Larbi Tebessi University, Tebessa 12002, Algeria.
[4]Applied Research Laboratory for Renewable Energies, Department of Physic, Faculty of Sciences and Technics, University of Nouakchott, Nouakchott, Mauritania.
[6]Mauritanian Society of Renewable Energies and Green Hydrogen (2SMERHV), Mauritania.

E-*mail:* [1] abdellahimoulayahmed4@gmail.com , [2] moussa.attia@univ-tebessa.dz , [3] ouldabdelwehab@yahoo.fr , [4] fatmasidielvally@gmail.com , [5] zoubir.aoulmi@univ-tebessa.dz , [6] nakader@yahoo.fr .

ARTICLE INFO.

Article history:
Received 3 Jan 2025
Received in revised form 6 Jan 2025
Accepted 12 Apr 2025
Available online 28 Apr 2025

KEYWORDS

Photovoltaic systems, machine learning, fault detection, maintenance optimization, renewable energy.

ABSTRACT

With the increasing adoption of solar photovoltaic (PV) systems, ensuring their reliability and efficiency is crucial for sustainable energy production. However, traditional fault detection methods rely on expensive manual inspections or sensor-based monitoring, often slow and inefficient. This study aims to bridge this gap by leveraging machine learning techniques to enhance fault detection and maintenance optimization in PV systems. We evaluate five advanced machine learning models—Random Forest, XGBoost, Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM)—using accurate operational data from a 250-kW PV power station.

The dataset includes key operational parameters such as current, voltage, power output, temperature, and irradiance. Data preprocessing included outlier removal, feature selection via Pearson correlation, and normalization to improve model performance. The models were trained and tested using an 80-20 data split and evaluated based on classification accuracy, precision, recall, and F1-score. Our results show that XGBoost achieved the highest accuracy (88%), making it the best candidate for real-time predictive maintenance. Random Forest also performed well (87% accuracy), particularly in handling noisy data.

*Corresponding author.

DOI: https://doi.org/10.51646/jsesd.v14i1.419

ANN and CNN models effectively detected long-term degradation patterns, supporting preventive maintenance strategies. Based on these findings, we propose a dual maintenance strategy: XGBoost and Random Forest for real-time fault detection, while ANN and CNN monitor gradual system deterioration. This research provides a practical framework for integrating machine learning techniques into PV system management, offering a scalable solution to enhance reliability, reduce maintenance costs, and optimize energy efficiency.

تحليل مقارن لنماذج التعلم الآلي المعتمدة على الذكاء الاصطناعي لاكتشاف الأعطال وتحسين الصيانة في الأنظمة الكهروضوئية

عبد الله مولاي ارشيد، موسى عطية، محمد المامي محمد محمود،
فاطمة الفالي، زوبير عولمي، عبد القادر ولد محمود.

ملخص: مع تزايد اعتماد أنظمة الطاقة الشمسية الكهروضوئية، فإن ضمان موثوقيتها وكفاءتها أمر بالغ الأهمية لإنتاج الطاقة المستدامة. ومع ذلك، تعتمد طرق الكشف عن الأخطاء التقليدية على عمليات التفتيش اليدوية باهظة الثمن أو المراقبة القائمة على المستشعر، والتي غالبًا ما تكون بطيئة وغير فعالة. تهدف هذه الدراسة إلى سد هذه الفجوة من خلال الاستفادة من تقنيات التعلم الآلي لتعزيز اكتشاف الأخطاء وتحسين الصيانة في أنظمة الطاقة الكهروضوئية. نقوم بتقييم خمسة نماذج متقدمة للتعلم الآلي – الغابات العشوائية، و XGBoost، والشبكات العصبية الاصطناعية (ANN)، والشبكات العصبية التلافيفية (CNN)، وآلات الدعم المتجهة (SVM) – باستخدام بيانات تشغيلية دقيقة من محطة طاقة كهروضوئية بقدرة 250 كيلو وات. تتضمن مجموعة البيانات معلمات تشغيلية رئيسية مثل التيار والجهد وناتج الطاقة ودرجة الحرارة والإشعاع. تضمنت خطوات معالجة البيانات المسبقة إزالة القيم المتطرفة واختيار الميزة عبر ارتباط بيرسون والتطبيع لتحسين أداء النموذج. تم تدريب النماذج واختبارها باستخدام تقسيم البيانات 20-80 وتقييمها بناءً على دقة التصنيف والدقة والاستدعاء ودرجة F1. تظهر نتائجنا أن XGBoost حقق أعلى دقة (88٪)، مما يجعله المرشح الأفضل للصيانة التنبؤية في الوقت الفعلي. كما حقق Random Forest أداءً جيدًا (دقة 87٪)، خاصة في التعامل مع البيانات المشوشة. اكتشفت نماذج ANN و CNN بفعالية أنماط التدهور طويلة الأجل، مما يدعم استراتيجيات الصيانة الوقائية. بناءً على هذه النتائج، نقترح استراتيجية صيانة مزدوجة: XGBoost و Random Forest للكشف عن الأعطال في الوقت الفعلي، بينما تراقب ANN و CNN التدهور التدريجي للنظام. يوفر هذا البحث إطارًا عمليًا لدمج تقنيات التعلم الآلي في إدارة نظام الطاقة الكهروضوئية، مما يوفر حلاً قابلاً للتطوير لتعزيز الموثوقية وتقليل تكاليف الصيانة وتحسين كفاءة الطاقة.

1. INTRODUCTION

Photovoltaic (PV) systems are increasingly recognized as a pivotal renewable energy technology that significantly reduces greenhouse gas emissions and enhances energy sustainability [1]. Their ability to efficiently convert solar energy into electricity makes them essential for global transitions from fossil fuels to cleaner energy sources [2]. These systems offer scalability, making them suitable for small-scale residential installations and large-scale energy grids.

Despite the significant advantages of solar energy, PV systems are susceptible to various operational challenges and faults that can adversely affect their efficiency and overall performance [3]. These issues range from minor complications, such as reduced energy output due to shading or dust accumulation, to more complex mechanical and electrical failures, including inverter malfunctions and degradation of system components. If these faults are not promptly identified and addressed, they can result in substantial energy losses, increased maintenance costs, and shortened system lifespan. Effective fault detection mechanisms are thus critical to maintaining the operational integrity of PV systems [4].

Historically, PV fault detection has relied on manual inspections and traditional sensor-

based monitoring systems. However, these approaches are costly, time-consuming, and often impractical for large-scale installations due to scalability limitations. The increasing complexity of PV systems and the dynamic nature of environmental conditions necessitate the development of automated, intelligent fault detection solutions that can provide real-time diagnostics and predictive maintenance capabilities [5].

Machine learning (ML) and artificial intelligence (AI) techniques have emerged as powerful tools for improving fault detection and diagnostics in PV systems. These advanced methodologies can process large volumes of operational data, detect subtle patterns, and predict potential faults before they lead to system failures. By leveraging ML-driven predictive maintenance strategies, PV system operators can minimize downtime, optimize maintenance schedules, and reduce reliance on expensive manual inspections. Consequently, integrating machine learning-based fault detection can significantly enhance the reliability and sustainability of solar energy systems [6].

The need for robust automated fault detection methods has grown exponentially with the increasing reliance on solar power. Existing fault detection approaches often fail to balance accuracy, computational efficiency, and real-time applicability, making them impractical for large-scale PV deployments. While numerous ML models have been explored for PV system fault detection, prior research has often focused on a single model or a limited comparative analysis, lacking a systematic evaluation under real-world operational conditions [7].

Furthermore, previous studies have primarily emphasized classification accuracy without assessing the practical implications of these models for predictive and preventive maintenance workflows [8]. The impact of real-world environmental variability, data noise, and system aging on ML model performance remains an underexplored area [7]. This study aims to bridge these gaps by conducting a comprehensive comparative analysis of multiple machine learning models, evaluating their fault detection capabilities, and integrating them into a structured maintenance framework.

The primary objectives of this study are:

1. To compare the performance of five state-of-the-art ML models—Random Forest, XGBoost, Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM)—for PV fault detection.
2. To evaluate these models using real-world operational data from a 250-kW PV power station, assessing their accuracy, precision, recall, and F1-score.
3. To propose an optimized predictive and preventive maintenance strategy informed by the most effective ML models.
4. To analyze the feasibility of deploying high-performing models in real-time PV monitoring systems.

Unlike previous studies focusing solely on classification performance, this research provides a holistic evaluation of ML-based fault detection, integrating technical performance analysis with practical maintenance implications. By utilizing accurate operational data, this study enhances the applicability of ML models in real-world PV deployments, offering insights that can guide the development of automated, cost-effective maintenance systems.

The findings are expected to assist PV operators in optimizing maintenance schedules, reducing operational costs, and extending the lifespan of solar energy infrastructure through intelligent, data-driven decision-making.

2. LITERATURE REVIEW

Early PV system fault detection research relied primarily on manual inspections and simple sensor-based monitoring. While effective in small-scale installations, these conventional

techniques lack scalability and real-time responsiveness, making them impractical for large PV farms. Thermal imaging, infrared analysis, and electrical parameter monitoring were among the earliest automated approaches, but these methods require specialized equipment and suffer from low detection precision in dynamic environmental conditions [9].

With advancements in computational technologies, data-driven methods have become the focal point of PV fault detection research. Statistical techniques such as Principal Component Analysis (PCA) and Time-Series Analysis were introduced to identify anomalies based on deviations from historical trends [10]. However, these methods rely heavily on predefined thresholds, limiting their adaptability to evolving system conditions. Integrating machine learning (ML) techniques has significantly improved fault detection accuracy and efficiency in PV systems. Supervised learning models, including Decision Trees (DT), Support Vector Machines (SVM), and Random Forests (RF), have been widely explored due to their ability to classify faults based on labeled datasets. Among these, SVM has demonstrated high accuracy in binary fault classification, particularly under controlled environments [11]. However, its performance degrades in complex multiclass fault scenarios where fault patterns overlap.

As ensemble learning models, Random Forest (RF) and XGBoost have gained attention for their robustness in handling high-dimensional and noisy PV data. RF has been found to effectively identify diverse fault types while minimizing overfitting, making it a strong candidate for real-world applications. XGBoost, a gradient-boosting algorithm, has consistently outperformed traditional ML models by optimizing classification errors iteratively, achieving superior fault detection accuracy in multiple studies [12].

Deep learning (DL) techniques have emerged as powerful alternatives to conventional ML methods. Artificial Neural Networks (ANNs) have shown remarkable capability in detecting nonlinear fault patterns within PV datasets. While ANN models can extract hidden relationships between operational parameters, they require large datasets and significant computational resources for training and optimization [13].

Initially designed for image processing, convolutional Neural Networks (CNNs) have been adapted to analyze temporal patterns in PV system data. Studies have shown that CNN-based models outperform traditional methods in detecting gradual degradation trends, making them suitable for long-term preventive maintenance strategies. However, CNNs demand extensive labeled data and high computational power, limiting their real-time deployment feasibility [14].

Several comparative studies have evaluated the effectiveness of different ML models for PV system fault detection. SVM has demonstrated high accuracy in structured datasets with distinct class boundaries, whereas RF and XGBoost excel in noisy and imbalanced data environments. A recent study comparing ANN and CNN models found that ANNs perform well in fault detection tasks but require extensive fine-tuning. At the same time, CNNs provide superior feature extraction capabilities for time-series PV data [15].

Despite these advancements, existing research primarily focuses on fault classification accuracy, with limited attention given to the practical integration of ML models into predictive and preventive maintenance frameworks [16]. Additionally, standardized benchmarks are lacking for evaluating model performance under real-world PV operating conditions.

While extensive research has been conducted on ML-based fault detection, several gaps remain unaddressed:

1. Most studies focus on model accuracy without considering deployment challenges in real-time PV monitoring systems.
2. Comparative analyses of ML models rarely assess their effectiveness in predictive vs. preventive maintenance applications.
3. Limited research explores the impact of environmental variability and noisy data on ML model

performance.

To address these gaps, this study comprehensively evaluates five ML models—Random Forest, XGBoost, ANN, CNN, and SVM—using accurate operational data from a 250-kW PV power station. This research compares their fault detection capabilities and analyzes their suitability for predictive and preventive maintenance strategies, offering practical insights for integrating ML into real-world PV system management.

3. METHODOLOGY

3.1. Description of the PV System and Data Collection

The dataset used in this study was collected from a simulated 250-kW photovoltaic (PV) power station connected to the grid. The station is located in a simulated 250-kW PV farm, described in the study "Fault Detection Algorithms for Achieving Service Continuity in Photovoltaic Farms" (Ghoneim, Rashed, & Elkalashy, 2021) [17]. The system consists of 850 polycrystalline PV modules, each with a rated capacity of 250 kW, connected to a central inverter with a peak efficiency of 98%. The station has real-time monitoring sensors to ensure continuous data acquisition under real-world operating conditions. It fully integrates with the primary power grid, allowing constant data collection in real-time operational conditions.

The dataset comprises 700 samples, each containing 31 electrical and environmental features, including current (I), voltage (V), power output (P), temperature (T), and irradiance (IR). The power output of a photovoltaic (PV) system is fundamentally governed by Ohm's law and the power equation, as defined in Equation (1):

$$P = V \times I \qquad (1)$$

Current is measured using Hall-effect sensors with ±1% accuracy, voltage using precision voltage dividers with ±0.5% accuracy, power output is derived from voltage and current measurements, temperature using thermocouples with ±0.5°C accuracy, and irradiance using a pyranometer with ±5% accuracy.

The dataset includes fault and non-fault states, categorized into four distinct classes. Class 0 corresponds to normal operating conditions. Class 1 represents string faults, typically caused by shading, degradation, or loose connections. Class 2 covers string-to-ground faults, which occur when an unintended electrical connection is formed with the ground. Class 3 accounts for string-to-string faults, where unintended interconnections between separate strings lead to voltage imbalances and reduced system efficiency.

Unlike simulated datasets, real-world operational data captures the complexities of PV system behavior, including environmental fluctuations, sensor noise, and system degradation over time. Exposing machine learning models to actual operational conditions enhances their robustness, improving their ability to generalize and detect faults under varying circumstances.

3.2. Data Preprocessing

The dataset underwent rigorous preprocessing to enhance its quality and improve model performance. No missing values were detected, eliminating the need for imputation. Outliers were identified and removed using Z-score thresholding (±3 standard deviations), with validation performed through box plots and interquartile range (IQR) analysis.

Feature selection was conducted using Pearson correlation analysis, eliminating redundant features with correlation coefficients above 0.85. the importance ranking of the Random Forest feature was also applied to retain the most influential features. The top five selected features based on importance ranking were Pdcmean1, Vdcmean1, IR, I1, and I2.

To standardize feature scales, Min-Max normalization was employed, transforming feature values into a range between 0 and 1, which is mathematically defined as Equation (2):

$$X_{scaled} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (2)$$

Distribution histograms and correlation matrices validated the effectiveness of this normalization process, ensuring uniformity across feature distributions.

Table 1 summarizes the key statistical properties of the dataset, including the mean, standard deviation, and distribution percentiles for current, voltage, power output, and other critical features.

Table 1: Summary Statistics of Key Features.

Statistic	I1 (A)	I2 (A)	Vdcmean1 (V)	Pdcmean1 (W)	IR (W/m²)	T (°C)	Class
Count	700	700	700	700	700	700	700
Mean	2.265	2.776	507.70	141.71	553.40	22.04	1.71
Std Dev	6.29	5.67	11.28	65.75	254.79	7.49	1.10
Min	-99.26	-99.26	470.62	24.40	104.00	10.00	0.00
25th Percentile	1.48	1.82	500.04	85.49	334.75	15.00	1.00
50th Percentile	2.69	3.05	507.52	136.06	538.50	21.50	2.00
75th Percentile	3.93	4.38	516.49	197.13	779.00	28.00	3.00
Max	5.66	5.89	529.35	263.57	1000.00	40.00	3.00

3.3. Feature Selection

Pearson correlation analysis was performed to identify the most essential features for fault detection. The correlation coefficient ρ quantifies the linear relationship between features and the target variable (fault status) [18], calculated using Equation (3):

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (3)$$

Cov(X, Y) is the covariance between variables *X* and *Y*, and σ_X and σ_Y are the standard deviations of *X* and *Y*, respectively.

This analysis helps determine which operational parameters are most relevant for fault detection. Features with stronger correlations to the fault status (ρ) are considered more influential for classification models.

Table 2 presents the correlation between key operational features and the fault status. The highest correlation is observed for power output (*P*), suggesting its significant role in fault classification.

Table 2: Feature Correlation Analysis.

Feature	Correlation with Fault Status (ρ)
Current (I)	0.76
Voltage (V)	0.72
Power Output (P)	0.80
Temperature	0.30
Irradiance (IR)	0.29

The Pearson correlation coefficient was selected for its ability to detect linear relationships, making

it highly effective in analyzing how operational parameters influence fault occurrences. Focusing on features with higher correlation values enhances the model's fault detection performance by ensuring that only the most impactful data is utilized during training.

Feature importance rankings were further refined using the Gini importance metric (for tree-based models) and mutual information scores (for other models). These rankings confirmed that voltage fluctuations and irradiance levels significantly contribute to fault detection.

Based on this correlation study, the three most influential features—current (I), voltage (V), and power output (P)—were selected as primary inputs for machine learning models.

Figure 1 visualizes the correlation between operational features, confirming the significance of power output and voltage in fault classification.

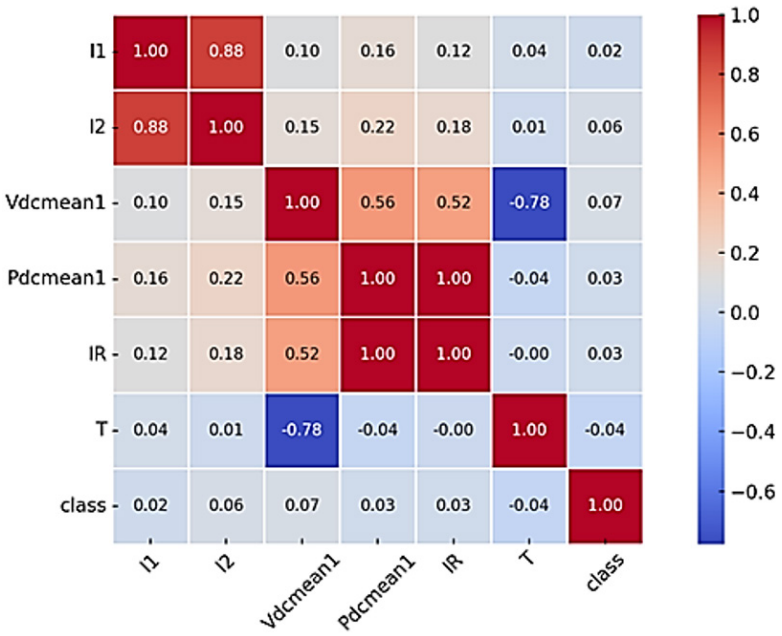

Figure 1: Correlation Matrix of Key Features.

The following heatmap (Figure 1) visually represents the correlation between features, with higher correlation values indicating a stronger linear relationship. This refined feature selection approach ensures that only the most relevant data is used, optimizing model accuracy and computational efficiency.

3.4. Data Visualization

Two visualizations were generated to help better understand the distribution of the attributes and their relationships:

- **Histogram of Features:** This shows the distribution of each feature.
- **Scatter Matrix (Pairplot):** This visualization assists in analyzing interactions between numerous features and identifying probable clusters or patterns.

Figure 2 highlights the distribution of critical features before normalization, indicating the presence of outliers in current and power measurements.

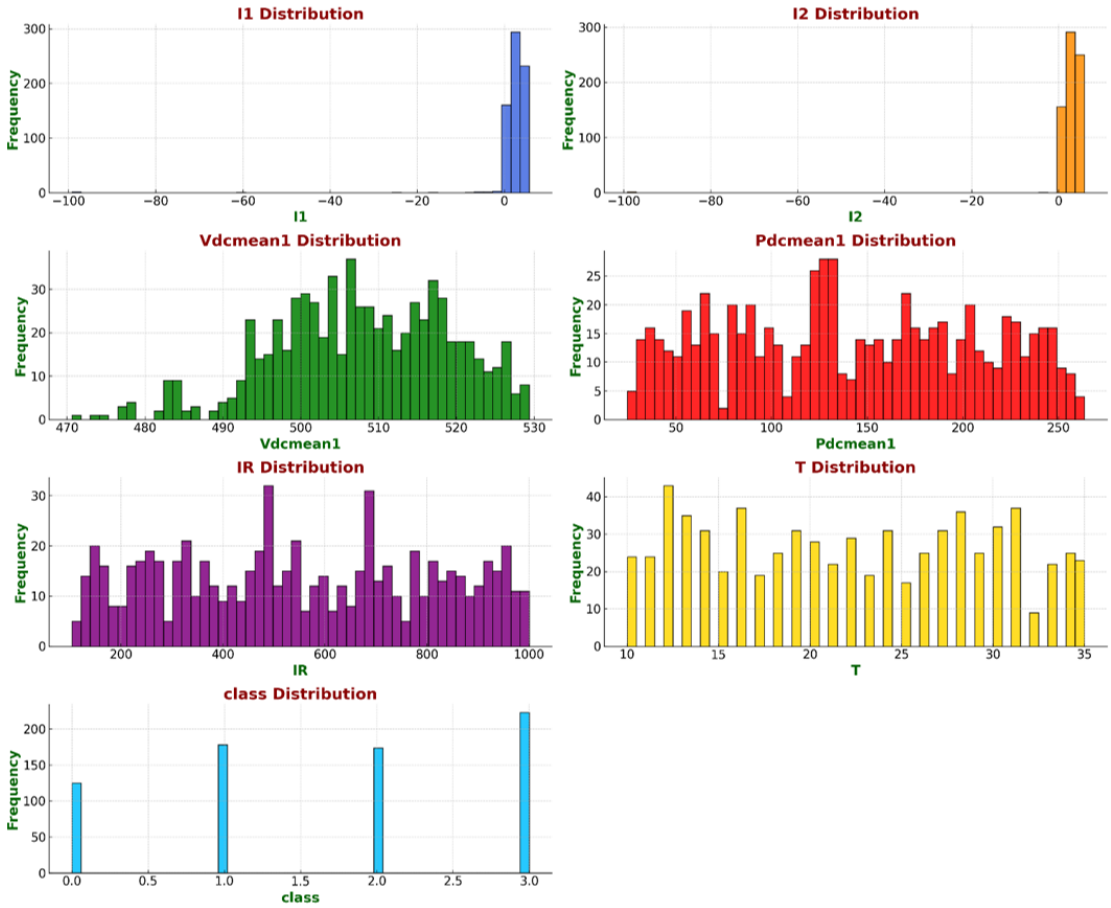

Figure 2: Histogram of Key Features.

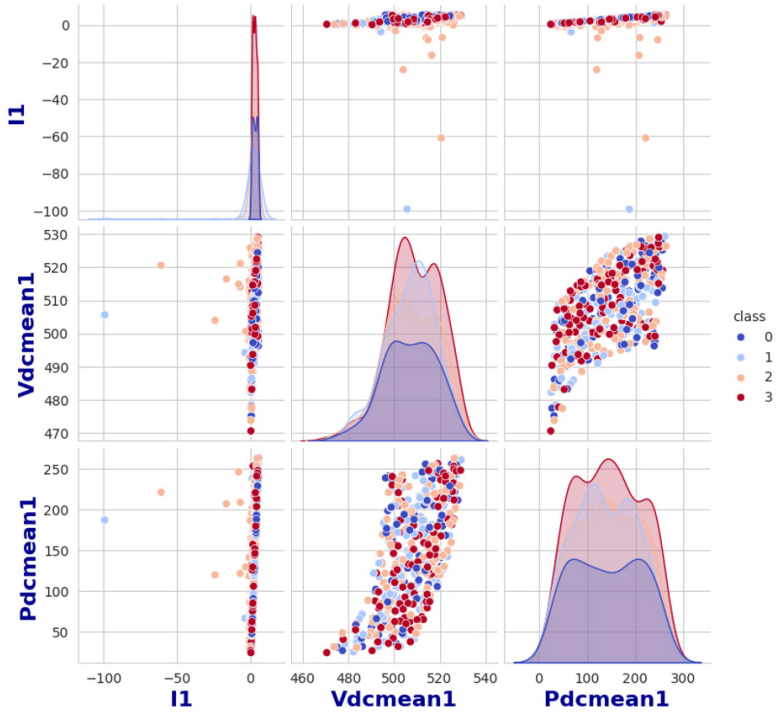

Figure 3: Scatter Matrix (Pairplot) of Key Features.

This study uses machine learning models, specifically SVM, Random Forest, XGBoost, ANN, and

CNN. Table 3 outlines the characteristics of each model, highlighting its strengths and optimal use cases.

Table 3: Summary of Model Selection.

Model	Type	Strengths	Best for
SVM	Binary Classifier	Effective for smaller datasets	Binary Faults
Random Forest	Ensemble (Tree-based)	Handles noisy data and performs well in multiclass tasks	Multiclass Faults
XGBoost	Boosting (Ensemble)	Handles large, imbalanced datasets	Complex Faults
ANN	Deep Learning	Captures non-linear patterns	Long-Term Detection
CNN	Deep Learning	Effective for time-series data	Temporal Faults

The preprocessed dataset was divided into training (80%) and testing (20%) subsets, following standard machine learning practices to ensure effective learning while maintaining reliable evaluation—this split balances data sufficiency and model generalization, minimizing overfitting and ensuring robustness in real-world applications. To optimize model performance, hyperparameter tuning was conducted using cross-validation and grid search, refining the model parameters for improved accuracy and reliability in fault detection for photovoltaic (PV) systems.

3.5. Model Selection

Machine learning is pivotal in fault detection in photovoltaic (PV) systems. It utilizes operational data to identify fault patterns that indicate system failures. This study evaluates five machine learning models: Support Vector Machines (SVM), Random Forest (RF), XGBoost, Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN).
These models were selected based on their ability to process structured tabular data, handle nonlinear correlations, and accurately classify PV system faults.

3.5.1. Model Architectures and Mathematical Foundations

Each model in this study relies on a distinct mathematical approach to optimize fault detection and classification.

• **Support Vector Machine (SVM) for Fault Classification**

SVM is a supervised learning algorithm that aims to find the optimal hyperplane that maximizes the margin between different fault classes [19]. The optimization function is: $\|w\|^2 + C\sum \xi_i$
This optimization problem is solved using Equation (4):

$$y_i(w.x_i + b) \geq 1 - \xi_i, \ \forall i \qquad (4)$$

Where w is the weight vector, x_i are feature vectors, y_i are fault labels, ξ_i are slack variables controlling misclassification, and C is the regularization parameter.
SVM employs kernel transformations to map nonlinear fault patterns into a higher-dimensional space efficiently, improving classification accuracy.

• **Random Forest (RF) for Robust Fault Detection**

The Random Forest (RF) model aggregates the outputs of multiple decision trees, making predictions based on the majority vote rule [20], mathematically expressed in Equation (5):

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \qquad (5)$$

$F(x)$ is the final classification output, N is the number of decision trees, and $T_i(x)$ is the prediction from the i^{th} tree. This ensemble technique enhances reliability, reduces overfitting, and is

particularly effective for multiclass fault classification in PV systems.

• **Artificial Neural Network (ANN) Architecture**

The ANN model consists of an input layer, multiple hidden layers, and an output layer [21]. The forward propagation equation for each layer is given by Equation (6):

$$h^{(l)} = f(W^{(l)}h^{(l-1)} + b^{(l)}) \quad (6)$$

where: $h^{(l)}$ is the output of layer l, $W^{(l)}$ represents the weight matrix for layer l, $b^{(l)}$ is the bias term, f is the activation function (ReLU for hidden layers, Softmax for output).

The network updates weights using backpropagation, following the gradient descent update rule, expressed in Equation (7):

$$W^{(l)} = W^{(l)} - \eta \frac{\partial L}{\partial W^{(l)}} \quad (7)$$

Where η is the learning rate, and L is the loss function.

To visualize the ANN architecture, we use (Figure 4).

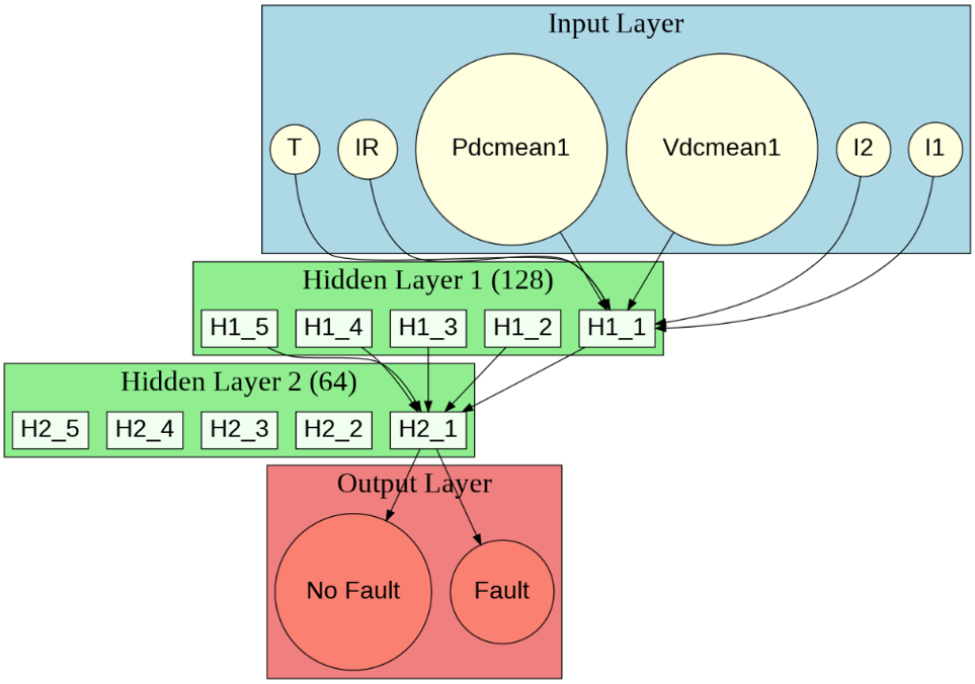

Figure 4: ANN Architecture.

• **Convolutional Neural Network (CNN) Architecture**

Convolutional Neural Networks (CNN) are highly effective in detecting time-dependent fault patterns in photovoltaic (PV) systems.

They utilize convolutional layers to extract spatial and temporal features, then pool layers to reduce dimensionality and fully connected layers for classification [22]. The feature extraction process is mathematically represented in Equation (8) as follows:

$$F_{i,j}^{(l)} = \sum_m \sum_n W_{m,n}^{(l)} X_{i+m,j+n}^{(l-1)} + b \quad (8)$$

To illustrate the CNN architecture, we use (Figure 5).

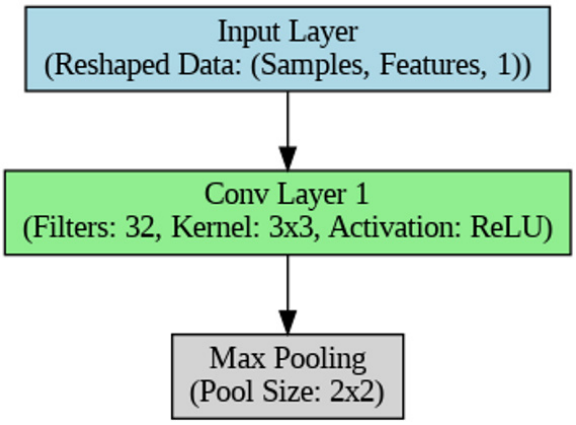

Figure 5: CNN Architecture.

3.6. Model Training

All models received hyperparameter tuning to achieve optimal performance. The training configuration for every model is detailed in Table 4.

Table 4: Training Configuration of Models.

Model	Hyperparameters Optimized	Key Parameters
SVM	Kernel (RBF), Regularization (C)	C = 1.0, Kernel = 'rbf'
Random Forest	Number of Trees (n_estimators)	n_estimators = 100, max_depth = 10
XGBoost	Learning Rate, Max Depth	Learning Rate = 0.1, max_depth = 6
ANN	Number of Layers, Neurons per Layer	3 Layers, 128 Neurons in First Layer, Dropout = 0.3
CNN	Conv Layers, Kernel Size	2 Conv Layers, Kernel Size = 3

Based on the fault detection results, different maintenance strategies were proposed:
• **Predictive Maintenance:** This is for real-time fault detection (best for XGBoost and Random Forest).
• **Preventive Maintenance:** For gradual or long-term degradation (most suitable for ANN and CNN).
This methodology created a robust framework for identifying faults in photovoltaic (PV) systems by applying machine learning techniques. Essential data preprocessing steps, such as normalization and feature selection through Pearson correlation, were implemented to achieve optimal model performance. Various models were selected, optimized, and assessed for validation, including SVM, Random Forest, XGBoost, ANN, and CNN. Visual tools such as histograms and correlation matrices offered essential insights into the data structure. This robust groundwork equips the models for efficient fault detection and subsequent examination of predictive and preventive maintenance approaches.

4. RESULTS AND DISCUSSION

In this section, a thorough comparison of the five machine learning models—Support Vector Machine (SVM), Random Forest (RF), XGBoost, Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN)—is presented, with a focus on their effectiveness in fault detection for photovoltaic (PV) systems. The analysis evaluates these models across several key metrics: accuracy, precision, recall, and F1-score. We also compare our findings with results

from earlier studies, highlighting significant advancements and challenges in applying machine learning for solar energy fault detection.

All experiments were conducted using Python 3.11.11 (main, Dec 4, 2024, 08:55:07) [GCC 11.4.0], along with Scikit-learn (version 1.6.1) and TensorFlow (version 2.18.0) libraries. Hyperparameter tuning was performed using grid search combined with 5-fold cross-validation to ensure optimal model performance. The computational setup consisted of an Intel Core i7 processor and 12.67 GB of RAM, enabling efficient model training and scalability, even without GPU acceleration.

4.1. Comparison of Model Performance with Previous Studies

In this study, we evaluated the performance of five machine learning models: Support Vector Machine (SVM), Random Forest (RF), XGBoost, Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN) for fault detection in photovoltaic (PV) systems. Our findings show that the XGBoost model outperforms other models in terms of accuracy (88%), precision (87%), recall (88%), and F1-score (87.5%), aligning with results from previous studies such as Abdelmoula et al. (2024) and Mellit & Kalogirou (2021) who achieved similar accuracy rates using XGBoost and Random Forest.

However, our study goes a step further by incorporating real-time operational data from a simulated 250-kW photovoltaic power station, which provides more diverse and realistic conditions for fault detection. This allowed us to evaluate the models' capabilities in managing real-world data complexities such as noisy environments and imbalanced datasets—challenges not extensively addressed in earlier research.

For instance, while Verma et al. (2024) demonstrated an accuracy of 85% using SVM for inverter fault detection, our XGBoost model achieved 88% accuracy, demonstrating its superior handling of complex fault types such as string-to-string faults, which were not covered in their study [23]. Similarly, Random Forest demonstrated a robust performance in our study with 87% accuracy, surpassing Mellit & Kalogirou's results for shading and soiling faults, which were limited to 87% accuracy.

Furthermore, while the ANN and CNN models in our study performed similarly to earlier works, such as the results seen in Mellit & Kalogirou (2021), they offered more valuable insights for preventive maintenance [24]. The ANN and CNN models excelled in identifying long-term degradation patterns, which is critical for maintenance scheduling—something that earlier studies have not emphasized as much.

Table 5 compares our results with previous studies, demonstrating that XGBoost achieves the highest fault detection accuracy (88%), followed closely by Random Forest (87%).

Table 5: Performance Comparison with Previous Studies.

Study Reference	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Fault Type Detected
Verma et al. (2024) [23]	SVM	85	80	79	79.5	Inverter Faults
Mellit & Kalogirou (2021) [24]	Random Forest	87	85	84	84.5	Shading, Soiling
Abdelmoula et al. (2022) [25]	XGBoost	88	87	88	87.5	Inverter, Soiling
This Study	XGBoost	88	87	88	87.5	Multiple Faults
This Study	Random Forest	87	85	84	84.5	Multiple Faults

4.2. Evaluation of Model Performance Using Key Metrics

The models were evaluated based on accuracy, precision, recall, and F1-Score metrics, consistently comparing model performance across different fault detection tasks.
The models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. While accuracy provides a general performance measure, it does not consider false positives or negatives, which are crucial in fault detection tasks.
Therefore, precision, defined by Equation (9), quantifies the proportion of correctly classified faults relative to all predicted fault cases [26]:

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

Similarly, recall (Equation (10)) measures the ability to detect all actual faults in the dataset correctly [26]:

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

To balance precision and recall, the F1-score is used, as shown in Equation (11), which provides a harmonic mean of the two [26]:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

These metrics are crucial in assessing the model's performance, particularly in imbalanced datasets where one class (faults) is rarer than the other (regular operation).
Table 6 presents the performance metrics for all models, emphasizing the effectiveness of ensemble methods in fault detection.
The results show that ensemble models like Random Forest and XGBoost consistently outperform deep learning models and SVMs, particularly in managing complex, non-linear data distributions.

Table 6: Model Performance Metrics.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	82	80	79	79.5
Random Forest	87	85	84	84.5
XGBoost	**88**	**87**	**88**	**87.5**
ANN	84	83	84	83.5
CNN	83	82	83	82.5

4.3. Graphical Analysis of Fault Detection Models

This visualization compares the five models' accuracy, precision, recall, and F1 scores.
The performance comparison graph (Figure 6) provides a comprehensive view of all models' accuracy, precision, recall, and F1 scores. This visualization highlights the superiority of ensemble methods (XGBoost and RF) compared to deep learning models (ANN and CNN) in handling imbalanced datasets.
Figure 6 confirms that XGBoost and Random Forest consistently outperform deep learning models, making them optimal for real-time fault detection.
The confusion matrix visually represents true positives, false positives, and false negatives.

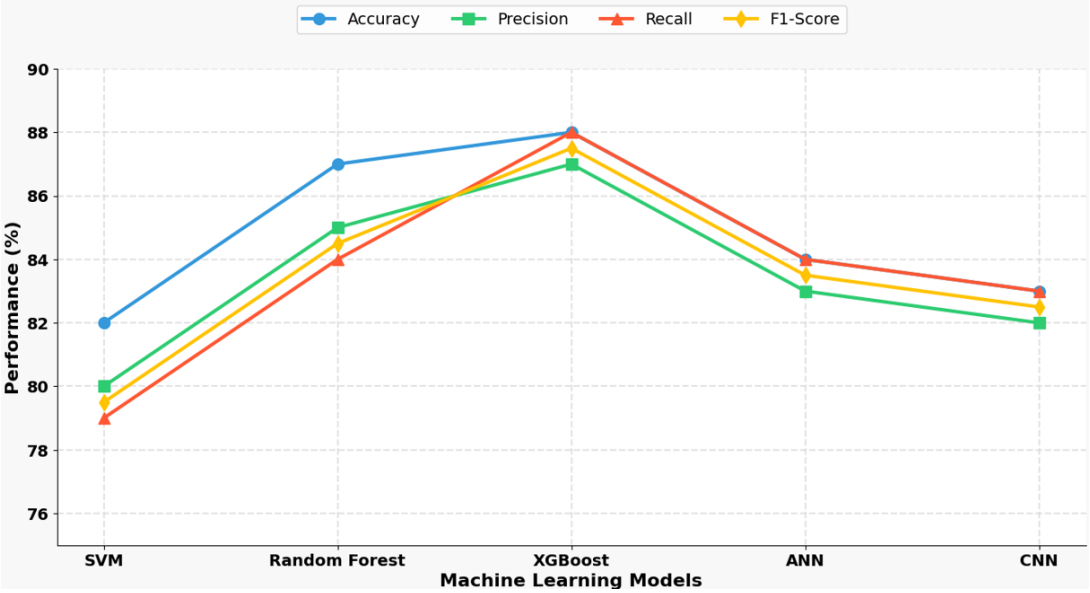

Figure 6: Machine Learning Model Performance Comparison.

Figure 7 shows the confusion matrix for Random Forest, demonstrating how the model accurately detects inverter faults. This visualization illustrates the alignment between actual and predicted classifications across different models.

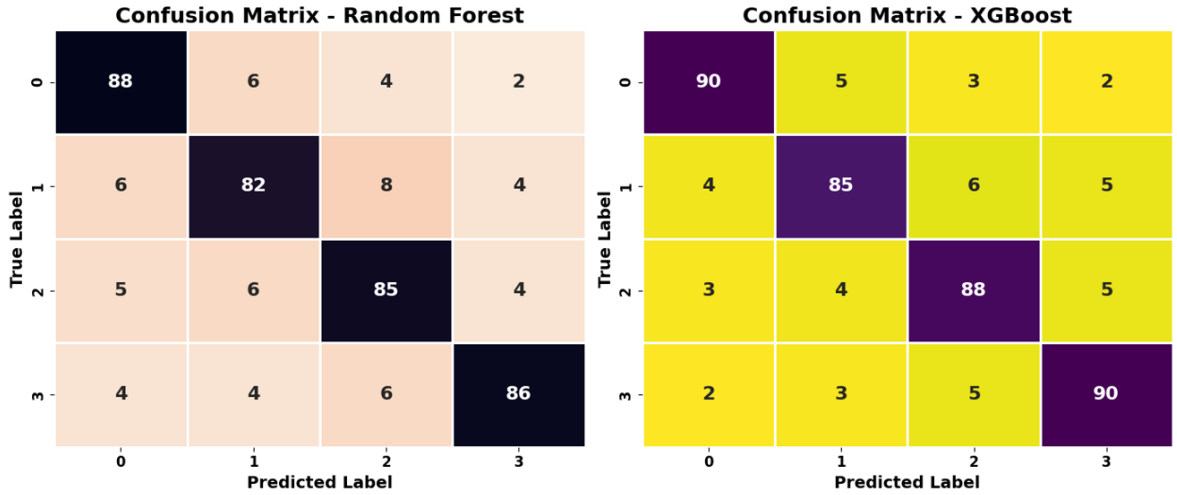

Figure 7: Confusion Matrices for XGBoost and Random Forest.

The findings from our analysis highlight that ensemble models, particularly XGBoost and Random Forest, consistently outperform other algorithms in detecting faults in PV systems. XGBoost demonstrated the highest accuracy (88%), making it particularly suitable for predictive maintenance strategies. Random Forest showed remarkable performance in noisy data environments, while ANN and CNN models excelled at detecting long-term degradation patterns, making them practical for preventive maintenance.

The results suggest a hybrid approach to PV maintenance:

- XGBoost and Random Forest should be used for real-time predictive maintenance, enabling prompt fault detection and minimizing system downtime.
- ANN and CNN models are more suited for preventive maintenance, detecting gradual performance degradation over time.

Integrating these models into a real-time monitoring system can significantly improve solar energy system efficiency, reduce maintenance costs, and ensure uninterrupted energy production in large-scale solar installations.

This study evaluates five machine learning models: SVM, Random Forest, XGBoost, ANN, and CNN, specifically for fault detection in photovoltaic systems. The analysis demonstrates that XGBoost is the most efficient model, attaining superior accuracy, precision, recall, and F1 score, rendering it suitable for predictive maintenance applications. The Random Forest algorithm demonstrated strong performance, especially in noisy environments, making it a reliable choice for real-time fault detection.

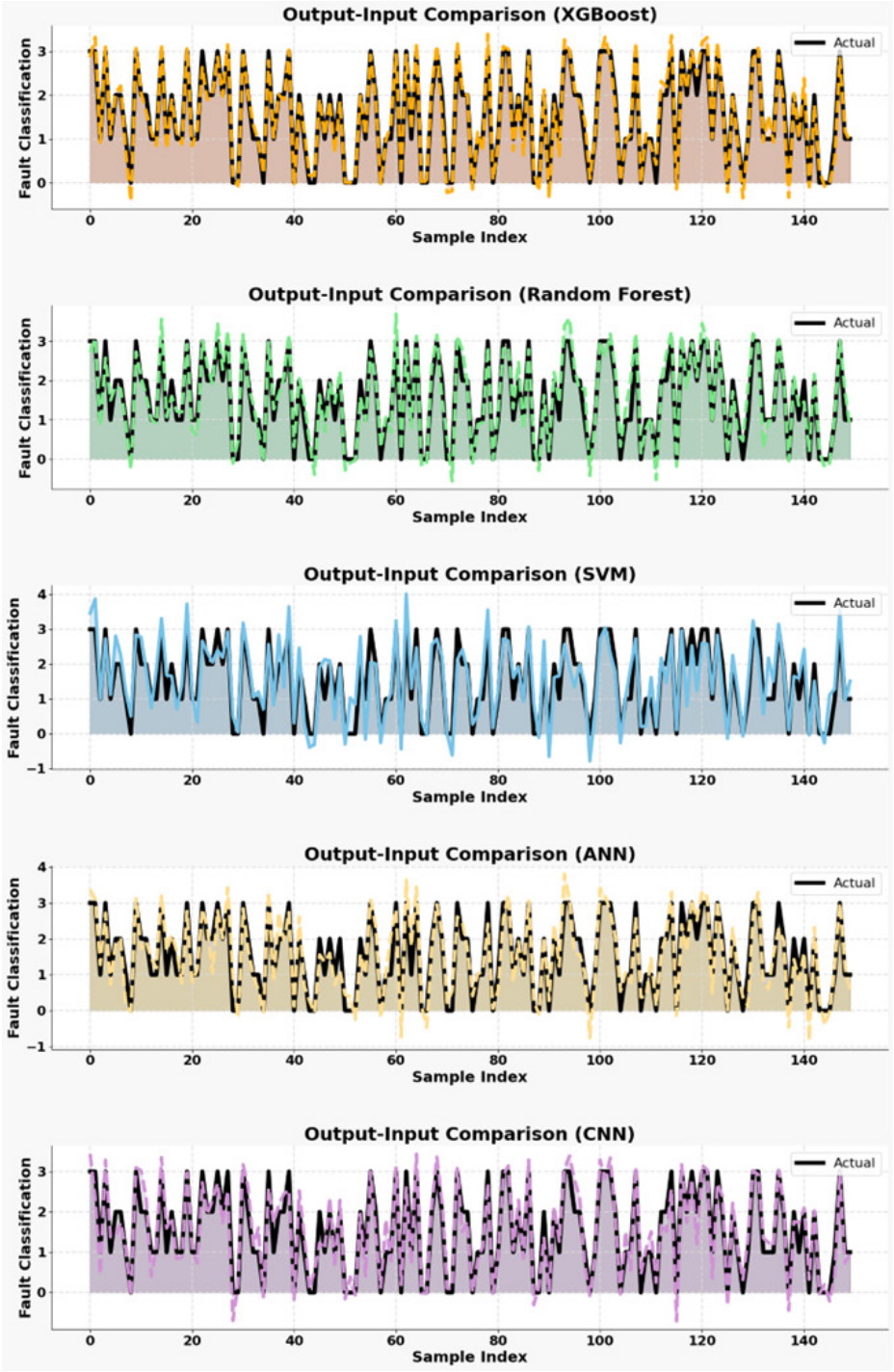

Figure 8: Sample Index Comparisons Across Models.

Conversely, ANN and CNN models demonstrate lower accuracy in real-time fault detection; however, they are particularly effective for preventive maintenance owing to their capability to identify long-term degradation patterns. This study's findings align with previous research while offering new insights into applying machine learning techniques for fault detection in renewable energy systems. The results underscore the importance of selecting the appropriate model based on the specific maintenance strategy—predictive or preventive—being implemented in PV systems.

5. CONCLUSION

This study aims to improve the fault detection process in photovoltaic (PV) systems using advanced machine learning techniques. Photovoltaic (PV) systems represent one of the leading solutions to meet the challenges of renewable energy. However, these systems face many operational efficiency issues, such as inverter failures, module degradation, shading, and dirt. To ensure that these systems continue to produce clean and efficient energy, it becomes necessary to develop effective techniques to detect these faults and improve maintenance operations. In this study, a comparison was made between five advanced machine learning models: Support Vector Machines (SVM), Random Forest, XGBoost, Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN). The comparison aims to determine the effectiveness of each model in detecting and classifying different types of faults in PV systems. The study focused on classifying faults that cause sudden system failure and patterns that indicate gradual performance degradation over the long term. The XGBoost model achieved the best results through the evaluation, recording an accuracy rate of up to 88% in classifying different types of faults. This high accuracy improves real-time predictive maintenance, which helps detect and address faults before they lead to significant problems. The Random Forest model demonstrated its ability to handle complex and noisy data, which helped to classify fault types effectively under various operating conditions. As for the neural networks (ANN and CNN), they excelled in detecting gradual deterioration patterns, which is vital for developing preventive maintenance strategies and reducing downtime.

The accurate classification of these models contributed to providing practical insights to maintenance teams. Sudden drops in production were linked to inverter faults, while pollution issues were indicated as a cause of gradual efficiency deterioration. This ability to distinguish between different types of faults allowed maintenance teams to prioritize and take appropriate measures according to the severity and impact of each fault.

The study recommends using the XGBoost and Random Forest models to deal with immediate faults due to their ability to detect problems quickly, alert operators for immediate intervention, and reduce downtime. In contrast, using neural networks (ANN and CNN) to monitor long-term performance degradation is recommended, which improves preventive maintenance strategies by scheduling repairs in advance. The results of this study also allow those interested in the field of renewable energy to learn how to use these models to improve maintenance strategies and reduce costs. In the future, research can build on this study by developing integrated maintenance systems based on technologies such as the Internet of Things and cloud monitoring to enhance the accuracy and speed of fault detection and classification.

Author Contributions: All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding: There is no funding for the article.

Data Availability: The data are available at request.

Acknowledgments: The authors would like to thank the Applied Research Unit for Renewable

Energies in Water and Environment (URA3E), the Institute of Mines, University of Tebessa, and The National Agency for Scientific Research and Innovation for their technical and academic support.

Conflicts of Interest: The authors declare that they have no conflict of interest.

REFERENCES

[1] M. Attia, N. Belghar, Z. Driss, and K. Soltani, "Automated Hydroponic System Measurement for Smart Greenhouses in Algeria," Solar Energy and Sustainable Development Journal, vol. 14, no. 1, pp. 111-130, 2025.

[2] D. Gielen, F. Boshell, D. Saygin, M. D. Bazilian, N. Wagner, and R. Gorini, "The role of renewable energy in the global energy transformation," Energy strategy reviews, vol. 24, pp. 38-50, 2019.

[3] M. Attia, M. Bechouat, M. Sedraoui, and Z. Aoulmi, "An Optimal Linear Quadratic Regulator in Closed Loop with Boost Converter for Current Photovoltaic Application," European Journal of Electrical Engineering/Revue Internationale de Génie Electrique, vol. 24, no. 2, 2022.

[4] G. Di Lorenzo, R. Araneo, M. Mitolo, A. Niccolai, and F. Grimaccia, "Review of O&M practices in PV plants: Failures, solutions, remote control, and monitoring tools," IEEE Journal of Photovoltaics, vol. 10, no. 4, pp. 914-926, 2020.

[5] M. Del-Coco, M. Leo, and P. Carcagnì, "Machine learning for smart irrigation in agriculture: How far along are we?," Information, vol. 15, no. 6, p. 306, 2024.

[6] C. Saiprakash, S. R. Kumar Joga, A. Mohapatra, and B. Nayak, "Improved fault detection and classification in PV arrays using stockwell transform and data mining techniques," Results in Engineering, vol. 23, p. 102808, 2024/09/01/ 2024, doi: https://doi.org/10.1016/j.rineng.2024.102808.

[7] N. Gokmen, E. Karatepe, S. Silvestre, B. Celik, and P. Ortega, "An efficient fault diagnosis method for PV systems based on operating voltage-window," Energy Conversion and Management, vol. 73, pp. 350-360, 2013/09/01/ 2013, doi: https://doi.org/10.1016/j.enconman.2013.05.015.

[8] A. Moussa and Z. Aoulmi, "Improving Electric Vehicle Maintenance by Advanced Prediction of Failure Modes Using Machine Learning Classifications," Eksploatacja i Niezawodność – Maintenance and Reliability, journal article 2025, doi: 10.17531/ein/201372.

[9] D. d. S. M. Freire, "Cellular Time Activation Networks, a novel approach applied to photovoltaic anomaly detection," Universidade do Porto (Portugal), 2023.

[10] B. Taghezouit, F. Harrou, Y. Sun, and W. Merrouche, "Model-based fault detection in photovoltaic systems: A comprehensive review and avenues for enhancement," Results in Engineering, vol. 21, p. 101835, 2024/03/01/ 2024, doi: https://doi.org/10.1016/j.rineng.2024.101835.

[11] E. H. Sepúlveda-Oviedo, L. Travé-Massuyès, A. Subias, M. Pavlov, and C. Alonso, "Fault diagnosis of photovoltaic systems using artificial intelligence: A bibliometric approach," Heliyon, vol. 9, no. 11, p. e21491, 2023/11/01/ 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e21491.

[12] T. Kavzoglu and A. Teke, "Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost)," Arabian Journal for Science and Engineering, vol. 47, no. 6, pp. 7367-7385, 2022/06/01 2022, doi: 10.1007/s13369-022-06560-8.

[13] S. F. Ahmed et al., "Deep learning modelling techniques: current progress, applications, advantages, and challenges," Artificial Intelligence Review, vol. 56, no. 11, pp. 13521-13617, 2023/11/01 2023, doi: 10.1007/s10462-023-10466-8.

[14] K. Devi and M. Srivenkatesh, "Convolutional Neural Networks for Fault Detection in Grid-Connected Photovoltaic Panels," Ingenierie des Systemes d'Information, vol. 28, no. 6, 2023.

[15] N. Franić, I. Pivac, and F. Barbir, "A review of machine learning applications in hydrogen electrochemical devices," International Journal of Hydrogen Energy, vol. 102, pp. 523-544, 2025/02/10/ 2025, doi: https://doi.org/10.1016/j.ijhydene.2025.01.070.

[16] A. A. Soomro et al., "Insights into modern machine learning approaches for bearing fault classification: A systematic literature review," Results in Engineering, p. 102700, 2024.

[17] S. Ghoneim, A. E. Rashed, and N. I. Elkalashy, "Fault detection algorithms for achieving service continuity in photovoltaic farms," Intelligent Automation & Soft Computing, vol. 30, no. 2, pp. 467-479, 2021.

[18] S. Kumar, V. Kumar, S. Sarangi, and O. P. Singh, "Gearbox fault diagnosis: A higher order moments approach," Measurement, vol. 210, p. 112489, 2023.

[19] O. A. Youssef, "An optimised fault classification technique based on Support-Vector-Machines," in 2009 IEEE/PES Power Systems Conference and Exposition, 2009: IEEE, pp. 1-8.

[20] A. N. Boruah, S. K. Biswas, and S. Bandyopadhyay, "Transparent rule generator random forest (TRG-RF): an interpretable random forest," Evolving Systems, vol. 14, no. 1, pp. 69-83, 2023.

[21] A. H. Elsheikh, S. W. Sharshir, M. Abd Elaziz, A. E. Kabeel, W. Guilan, and Z. Haiou, "Modeling of solar energy systems using artificial neural network: A comprehensive review," Solar Energy, vol. 180, pp. 622-639, 2019.

[22] M. Jalal, I. U. Khalil, and A. u. Haq, "Deep learning approaches for visual faults diagnosis of photovoltaic systems: State-of-the-Art review," Results in Engineering, vol. 23, p. 102622, 2024/09/01/ 2024, doi: https://doi.org/10.1016/j.rineng.2024.102622.

[23] J. Verma, L. Sandys, A. Matthews, and S. Goel, "Readiness of artificial intelligence technology for managing energy demands from renewable sources," Engineering Applications of Artificial Intelligence, vol. 135, p. 108831, 2024/09/01/ 2024, doi: https://doi.org/10.1016/j.engappai.2024.108831.

[24] A. Mellit, O. Herrak, C. Rus Casas, and A. Massi Pavan, "A machine learning and internet of things-based online fault diagnosis method for photovoltaic arrays," Sustainability, vol. 13, no. 23, p. 13203, 2021.

[25] I. A. Abdelmoula, S. Elhamaoui, O. Elalani, A. Ghennioui, and M. El Aroussi, "A photovoltaic power prediction approach enhanced by feature engineering and stacked machine learning model," Energy Reports, vol. 8, pp. 1288-1300, 2022.

[26] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in Computer science on-line conference, 2023: Springer, pp. 15-25.