Detection and Diagnosis in Photovoltaic Panels Using Artificial Intelligence Techniques Based on Random Forest and LSTM Neural Networks

G.Olteanu, I.Fagarasan and M.Dobrea, *Member, IEEE*

Abstract—This paper investigates the use of artificial intelligence techniques for monitoring photovoltaic systems and for fault detection and diagnosis. The proposed methodology uses approximately 1.37 million records collected over 16 days, including electrical variables, such as the voltage and DC current of the two photovoltaic strings, as well as meteorological parameters, such as solar irradiation and module temperature. The analysis process is structured in two stages: a binary detection stage, which differentiates normal from faulty operation, and a multiclass diagnostic stage, which identifies the type of fault: short circuit, degradation, open circuit or shading. Random Forest, LSTM and CNN-LSTM models are implemented and compared, to capitalize on both the nonlinear relationships between variables and the temporal dependencies in sequential data. Random Forest provides interpretability by analyzing the importance of features, LSTM captures the temporal evolution of signals, and the CNN-LSTM hybrid architecture combines automatic local feature extraction with temporal modeling. Evaluation based on accuracy, F1 score, and confusion matrices demonstrates high classification performance, with the CNN-LSTM model standing out for its increased robustness in identifying complex and transient photovoltaic defects.

I. INTRODUCTION

Traditional fault detection methods based on manual inspections or fixed-rule monitoring systems have limitations in identifying complex and progressive faults [4]–[5]. In contrast, artificial intelligence and machine learning techniques have shown strong capabilities in analyzing photovoltaic operational data and improving fault diagnosis performance [2], [7], [11]. In this context, the proposed methodology integrates fault detection and multiclass fault diagnosis using Random Forest (RF), LSTM, and hybrid CNN-LSTM models, combining non-linear relationship modeling [10] with temporal dependency analysis [8]–[9] to support predictive maintenance and early anomaly detection [14].

To preserve the chronological consistency of the time-series data and avoid temporal information leakage, the preprocessing and evaluation stages were carefully designed. RF models were used to capture non-linear relationships, while LSTM networks modeled temporal dependencies within sequential data. In addition, based on recent studies [8], [16], a hybrid CNN-LSTM architecture was also investigated to improve the robustness and classification accuracy of photovoltaic fault detection and diagnosis.

The paper presents the main photovoltaic faults, the proposed preprocessing methodology, and the implementation of the Random Forest (RF) , LSTM, and CNN-LSTM models, followed by a comparative evaluation using accuracy and F1-score metrics. Finally, the main conclusions and future directions for intelligent photovoltaic monitoring applications are discussed.

II. THEORETICAL FOUNDATIONS OF FAULTS IN PHOTOVOLTAIC SYSTEMS

Photovoltaic system faults can occur due to electrical issues, environmental factors, or material degradation, making predictive maintenance essential for the rapid identification and diagnosis of anomalies [10]. In the literature, these faults are generally classified into electrical faults, structural module faults, and environmentally induced faults, with their analysis based on criteria such as defect nature, occurrence frequency, impact on system performance, and detectable electrical signatures within the monitored data [10], [11].

2.1 .Classification of Faults in Photovoltaic Systems

The following table summarizes the main types of faults identified in photovoltaic systems, along with the way they affect system operation and the indicators through which they can be detected.

Technical faults in photovoltaic systems can compromise the performance of the entire installation, not only by reducing power generation and increasing energy losses but also by causing accelerated module degradation and increasing the risk of irreparable component damage.

In the case of faults such as shading or module degradation, the impact on energy production can be gradual and difficult to identify using traditional monitoring methods; thus, in recent years, methods based on artificial intelligence have been used to analyze data collected by photovoltaic systems and to automatically detect abnormal behaviors.

III. PROPOSED METHODOLOGY AND IMPLEMENTATION OF THE PHOTOVOLTAIC FAULT DETECTION AND DIAGNOSIS SYSTEM

The proposed approach combines classical machine learning methods and deep learning methods, used for the analysis of electrical and meteorological data originating from a monitored photovoltaic system; the primary goal is to develop an automated system capable of identifying deviations from the installation's normal operation and determining the specific type of fault affecting the system [7,12-13].

The general flow of the proposed method is presented in Figure 1, where the main stages of the data analysis and fault classification process are illustrated, highlighting how raw data collected from the photovoltaic system are gradually transformed into relevant information for fault detection and diagnosis through artificial intelligence models.

TABLE I. Classification of Faults in Photovoltaic Systems

Fault Type	Fault Category	Impact on the System	Detection Methods
Short Circuit	Electrical	Direct connection of two circuit points causes current to increase and voltage to decrease, potentially leading to component overheating.	I-V analysis, electrical monitoring, ML
Open Circuit	Electrical	Interruption of the electrical connection between modules or strings leads to a drastic decrease in current.	Electrical monitoring, time series analysis
Degrad ation	Electrical / Structural	Progressive reduction of module efficiency due to material aging and environmental conditions.	Performance analysis, machine learning
Shadow ing	Environm ental	Obstruction of solar radiation by external objects causes a reduction in current and power.	Production monitoring, ML
Hot Spot	Electrical / Structural	Localized cell overheating due to internal defects or shading.	Thermography
Micro-cracks	Structural	Microscopic cracks in cells that reduce electrical conductivity.	Electrolumine scence
Delamin ation	Structural	Separation of the PV module layers, affecting light transmission.	Visual inspection, thermography
Inverter Faults	System	Malfunctioning of the inverter affects the DC-AC conversion.	System monitoring

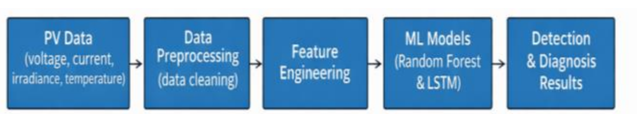

Figure 1. General flowchart of the methodology for fault detection and diagnosis in photovoltaic systems using Random Forest and LSTM models.

The proposed methodology for intelligent photovoltaic system monitoring is based on a continuous analysis flow that begins with the acquisition of electrical and meteorological parameters, followed by data preprocessing through abnormal value filtering, missing value removal, and temporal synchronization to ensure a reliable dataset. Subsequently, the dataset undergoes a feature engineering stage, during which additional attributes such as string power and the power-to-irradiance ratio are generated to better highlight patterns associated with abnormal operation. These processed data are then used to train complementary artificial intelligence models, including the Random Forest algorithm for statistical relationship analysis and LSTM neural networks for time-series processing [10], [18]. The trained models perform fault detection through binary classification and fault diagnosis through multiclass classification, enabling the identification of anomalies such as short circuits, degradation, open circuits, or shading phenomena, while contributing to improved operational efficiency and reduced maintenance time [8], [9].

3.1 The Dataset Used

The analysis was carried out using an open-source dataset collected from a photovoltaic system with two string arrays connected to the DC side of an inverter, over a period of 16 days, containing approximately 1.37 million observations. The electrical measurements and environmental parameters, such as solar irradiance and module temperature , were used to develop and validate an artificial intelligence-based method for the detection and classification of photovoltaic faults.

The analyzed variables include both electrical and meteorological parameters that describe the system's operating conditions. Within the dataset used, these variables are explicitly defined by the following names:

- vdc1 and vdc2 – the DC voltage of the two photovoltaic strings;
- idc1 and idc2 – the DC current generated by each photovoltaic string;
- irr – the level of solar irradiance incident upon the panels;
- pvt – the temperature of the photovoltaic module.

These variables represent the primary electrical and meteorological measurements used to analyze the photovoltaic system behavior and operating conditions. Each observation in the dataset is associated with the label *f_nv*, which defines the operating state or fault type and serves as the target variable for the machine learning classification process, while the corresponding fault classes and numerical codes are presented in Table 2.

Before training, the dataset was chronologically divided into 80% training and 20% testing subsets to preserve temporal continuity and prevent information leakage. Using a sliding-window strategy, the data were transformed into multivariate temporal sequences of 60 consecutive observations with a step size of 10 samples between windows, enabling the models to capture both local variations and temporal dependencies related to photovoltaic fault evolution.

TABLE II. Coding of Operating States and Fault Types for the Photovoltaic System.

f_nv Code	**Fault Type**
0	Normal Operation
1	Short circuit
2	Degradation
3	Open circuit
4	Shadowing

3.2. Data Preprocessing, Feature Extraction, and Implementation of Machine Learning Models

Before the actual training of the artificial intelligence models, the dataset obtained from the photovoltaic systems underwent a preprocessing stage to ensure data consistency. This process began with the removal of missing values caused by transmission errors and continued with the filtering of abnormal values, excluding physically implausible observations such as negative irradiance or unusual module temperatures.

To enable the mathematical operations required for modeling, all variables were converted into a compatible numerical format, followed by the temporal synchronization of information from the electrical and meteorological sensors, ensuring that each observation accurately reflected the system state at a given moment [8], [12].

Based on this stabilized dataset, the methodology continued with the feature engineering stage, during which derived variables were generated to better describe the electrical behavior of the system. Attributes such as string power, voltage and current differences between strings, average electrical quantities, and the ratio between total generated power and solar irradiance were calculated. The generation of these features was motivated by the fact that photovoltaic faults frequently appear as deviations from the normal correlation between generated power and incident irradiance.

The extracted features were subsequently normalized using standard scaling techniques to ensure numerical stability during neural network optimization and to avoid bias caused by differences in variable magnitudes. Under nominal operating conditions, power increases proportionally with irradiance, while faults or shading phenomena produce significant deviations from this relationship, as illustrated in Figure 2.

To obtain this graph, the raw data were subjected to rigorous preprocessing and feature extraction stages, a process during which the total power of the photovoltaic system was calculated by summing the products of the voltages and currents of the two strings $(vdc_1, vdc_2, idc_1, idc_2)$, according to the relation:

$$P_{tot}=(vdc_1 * idc_1) + (vdc_2 * idc_2) \qquad (1)$$

Subsequently, the value P_{tot} was plotted against the solar irradiance i_{rr}.To facilitate visual classification and the identification of the footprint of each operating state, the points were distinctly colored based on the fault label f_nv, thus highlighting the specific deviations of each type of anomaly compared to the normal operating regime.

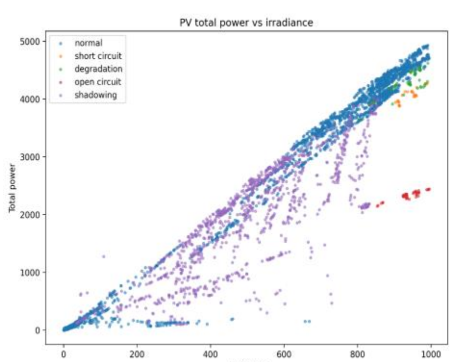

Figure 2. The relationship between total generated power and solar irradiance for various states of the photovoltaic system.

In the final phase, the extracted features were used as input variables for the models implemented in the Python environment using specialized libraries. The analysis process was structured in two sequential stages: initially, the Random Forest algorithm performed fault detection through binary classification to identify deviations from normal operating conditions, while observations classified as abnormal were subsequently analyzed using Long Short-Term Memory (LSTM) neural networks for multiclass fault diagnosis. This combined approach enabled the complementary advantages of both models to be exploited, with RF efficiently modeling non-linear relationships between variables and LSTM effectively capturing temporal dependencies within the data series, resulting in a robust and flexible framework for photovoltaic system monitoring.

3.3. The Random Forest model

For the fault classification stage, the Random Forest algorithm was selected, a robust ensemble learning method based on multiple decision trees trained on random data subsets and combined through majority voting [5], [10]. Its structure reduces the risk of overfitting while effectively modeling complex non-linear relationships between electrical and meteorological parameters. Using derived features such as string power and voltage/current differences, the model also provides variable importance ranking, offering a transparent interpretation of each feature's contribution to photovoltaic fault diagnosis, as illustrated in Figure 3 [18].

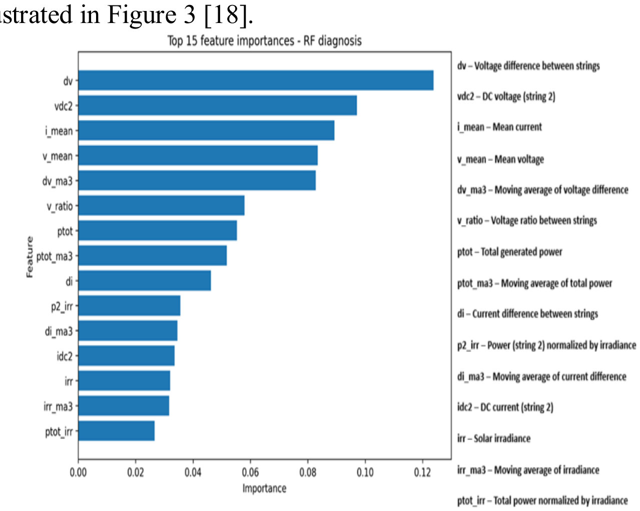

Figure 3. Feature Importance in the RF Model for PV Fault Diagnosis

The relevance analysis of the 15 critical variables used by the Random Forest model was performed based on importance scores calculated through the algorithm's internal mechanism, utilizing Gini impurity reduction and aggregating the contribution of each variable across the ensemble of decision trees.

The results highlight that the dv indicator exerts the greatest influence on the classification process, with a weight of approximately 0.125, thus confirming that the voltage differential between strings constitutes the primary predictor of anomalies. This is followed in the hierarchy by the parameters vdc2 (~0.10), i_mean (~0.09), and v_mean (~0.085), whose average values reflect the fundamental electrical state of the system.

In addition, derived variables such as dy_ma3 (~0.085) and v_ratio (~0.060) emphasize the essential role of temporal dynamics and voltage ratios.

This process is supported by power, current, and irradiance metrics which, although presenting lower weights, contribute to the informational granularity necessary for differentiating complex fault typologies and demonstrate that the model bases its decisions on the direct correlation between the physical behavior of the installation and the variations in electrical parameters.

3.4 The LSTM Model

For sequential photovoltaic data analysis, both standalone LSTM and hybrid CNN-LSTM architectures were implemented and comparatively evaluated [8–9], [16]. The LSTM model captured temporal dependencies and progressive behavioral changes in electrical and meteorological signals, while the CNN-LSTM architecture combined one-dimensional convolutional layers for automatic extraction of local electrical patterns with LSTM layers for temporal sequence modeling, improving the classification of complex and transient photovoltaic faults. The implemented CNN-LSTM model included two 1D convolutional layers with 64 and 128 filters, followed by batch normalization, max-pooling, dropout regularization, and an LSTM layer with 64 hidden units, trained using the Adam optimizer and categorical cross-entropy loss function over 30 epochs with a batch size of 256 observations [13].

To reduce overfitting and improve convergence stability, EarlyStopping and ReduceLROnPlateau callbacks were implemented during training. These mechanisms dynamically adjusted the learning process according to the validation loss evolution.

The efficiency and stability of this learning process were closely monitored by tracking accuracy over successive epochs, with the performance dynamics illustrated in detail in Figure 4, which clearly presents the evolution of the LSTM model's accuracy during the training process.

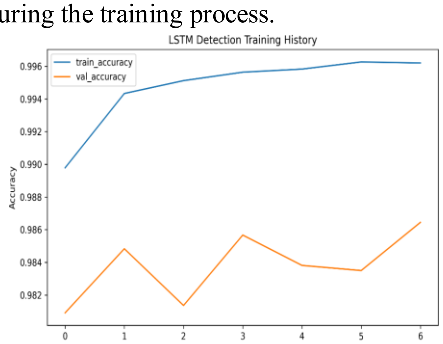

Figure 4. The evolution of LSTM model accuracy during the training proces

The LSTM model showed progressive performance improvement over the 7 training epochs, with train accuracy increasing from 99.0% to 99.7% and validation accuracy from 98.1% to 98.6%. The small fluctuations observed during training were part of the normal convergence process, while the narrow gap between training and validation results confirmed good generalization capability and the absence of overfitting.

IV. FAULT DETECTION AND DIAGNOSIS

The general process for analyzing anomalies in the photovoltaic system was strategically structured into two complementary stages, beginning with fault detection,a stage formulated strictly as a binary classification problem. Its objective is the rapid identification of any abnormal behavior by dividing observations into two main classes: normal operation and the presence of a fault.

This approach constitutes the first decision-making level of the intelligent monitoring process, the efficiency of which is validated by the performance of the Random Forest (RF) algorithm. The practical results of this first phase are illustrated in detail in Figure 5, which presents the model's confusion matrix. The analysis of this matrix clearly indicates that the system succeeds in correctly identifying the vast majority of instances, thus demonstrating a high and robust capacity to distinguish between the optimal state of the installation and failure situations [7,12].

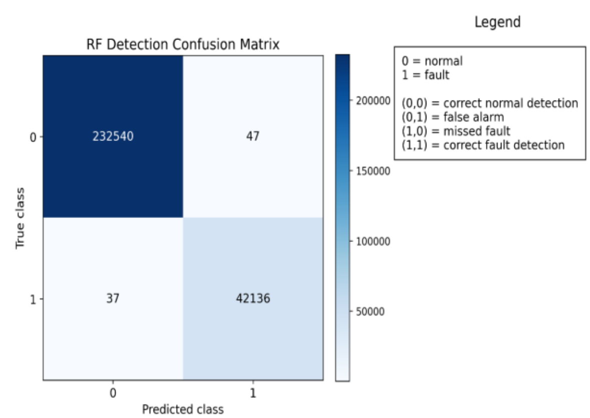

Figure 5. Confusion matrix for fault detection using the RF model.

The matrix indicates that the model correctly classified 232,540 observations corresponding to normal operation and 42,136 observations containing faults, while 47 normal instances were erroneously classified as faults and 37 fault cases were interpreted as being normal.

To compare classical machine learning and deep learning approaches, photovoltaic fault detection was evaluated using standalone LSTM and hybrid CNN-LSTM models. The confusion matrices in Figures 6 and 9 demonstrate their capability to identify abnormal operating conditions, although shadowing and open-circuit faults remained more difficult to distinguish due to similar electrical signatures. Therefore, model performance was additionally assessed using precision, recall, and F1-score metrics to provide a more rigorous evaluation of diagnostic capability.

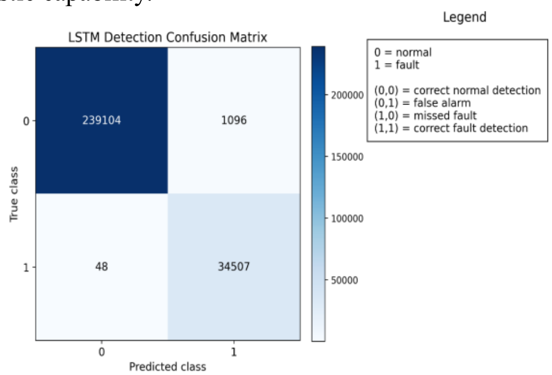

Figure 6. Confusion matrix for fault detection using the LSTM model.

Once the primary detection stage was completed, the observations identified as abnormal were further analyzed through a multiclass classification process to determine the exact fault category while the confusion matrix presented in Figure 7 demonstrates the high diagnostic accuracy of the Random Forest model, with a high rate of correct predictions and minimal confusion between classes, thereby validating the effectiveness of the proposed automated photovoltaic monitoring system.

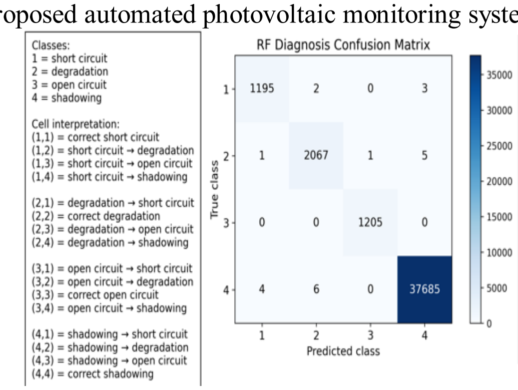

Figure 7. Confusion matrix for PV systems fault diagnosis with RF

For the diagnosis stage, the confusion matrix shows that the Random Forest model correctly identified 1,195 cases of short circuit, 2,067 cases of degradation, 1,205 cases of open circuit, and 37,685 cases of shading, with only a few erroneous classifications between classes, such as 2 cases of short circuit classified as degradation or 6 cases of shading classified as degradation.

For a comparative evaluation of model performance, the same diagnostic stage was also performed using the LSTM neural network, whose results are detailed in Figure 8, illustrating the precise manner in which the model classifies the different types of faults. This allows for a comparative analysis of the confusion matrices, from which it emerges that both Random Forest and LSTM are capable of identifying the analyzed typologies with high accuracy.

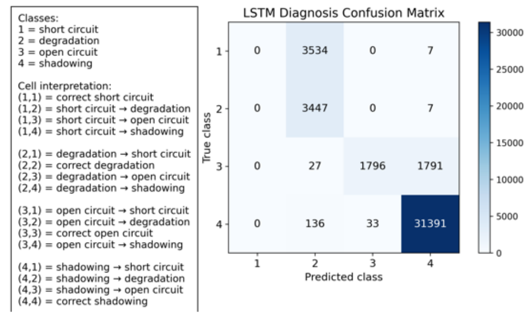

Figure 8. Confusion matrix for fault diagnosis in pv systems for LSTM

To further improve the capability of the system to identify complex fault signatures and reduce classification ambiguities between similar operating conditions, the multiclass diagnostic stage was additionally evaluated using a hybrid CNN-LSTM architecture. The corresponding confusion matrix, presented in Figure 9, demonstrates that the proposed hybrid model achieves a significantly improved classification performance by combining convolutional feature extraction with temporal dependency learning.

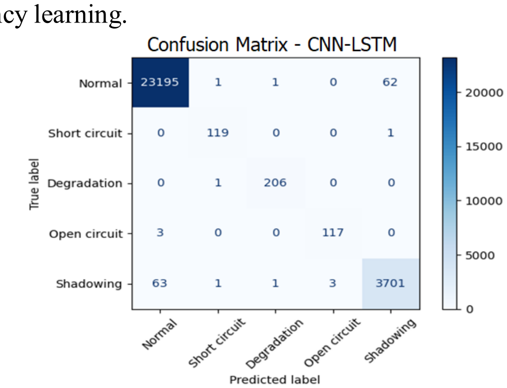

Figure 9. Confusion matrix for photovoltaic fault diagnosis using the proposed CNN-LSTM architecture.

4.1.Model Comparison

The performance of the Random Forest, LSTM, and CNN-LSTM models was evaluated using standard metrics commonly employed in classification problems, specifically accuracy and F1-score. These metrics were calculated for both the fault detection stage and the fault type diagnosis stage, with the comparative results summarized in Figure 10 [14], [17].

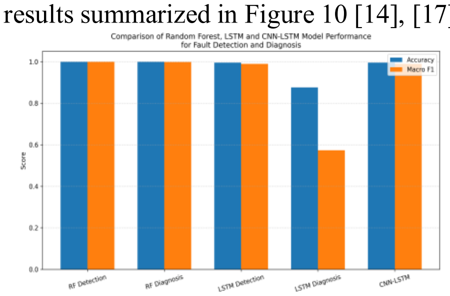

Figure 10. Comparison of RF, LSTM and CNN-LSTM model performance for fault detection and diagnosis.

The analysis of the results highlights that both models achieve high values for accuracy and the F1-score, indicating a strong capacity to identify faulty states and correctly classify the analyzed fault types. The F1-score represents the harmonic mean between precision and recall, defined as:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \qquad (2)$$

This metric provides a balanced evaluation of model performance, especially for imbalanced datasets, while the high F1-score values obtained experimentally indicate an effective compromise between accurate fault identification and the reduction of misclassifications [17–20]. The comparative analysis highlights significant differences between the evaluated models: Random Forest efficiently captures non-linear relationships between electrical and meteorological variables based on engineered features, the standalone LSTM model analyzes the temporal evolution of signals and sequential dependencies between observations, while the hybrid CNN-LSTM architecture combines convolutional extraction of local electrical signatures with LSTM-based modeling of long-term temporal dependencies. Consequently, the CNN-LSTM model demonstrated superior robustness and classification performance, particularly for complex and transient photovoltaic faults.

V. Conclusion

The present work validated the efficiency of an intelligent methodology based on Random Forest, LSTM, and CNN-LSTM algorithms for the detection and diagnosis of faults in photovoltaic systems, demonstrating high classification performance and robust fault identification capabilities across multiple photovoltaic operating conditions. By combining the interpretability of the Random Forest model with the temporal sequence learning capability of LSTM networks and the automatic feature extraction provided by convolutional neural networks, a robust diagnostic framework was obtained, capable of accurately identifying fault conditions such as short circuit, open circuit, degradation, and shading.

These artificial intelligence approaches improve predictive maintenance by reducing energy losses, optimizing fault localization, and decreasing intervention times in photovoltaic systems. The results confirm that Machine Learning and Deep Learning methods increase the reliability and operational efficiency of modern green energy infrastructures. Furthermore, the comparative analysis showed that the hybrid CNN-LSTM architecture achieved the best classification performance by combining convolutional feature extraction with temporal dependency learning, although further validation on larger and more diverse datasets is still required.

As future research directions, the proposed methodology can be extended through the integration of larger heterogeneous datasets and the implementation of the developed models into real-time monitoring and AIoT-based predictive maintenance systems, facilitating the transition toward fully automated intelligent photovoltaic infrastructures.

References

[1] G. R. Venkatakrishnan, R. Rengaraj, S. Tamilselvi, et al., “Detection, location, and diagnosis of different faults in large solar PV system—A review,” *International Journal of Low-Carbon Technologies*, vol. 18, 2023, pp. 659–674.

[2] M. A. Green, “Solar cell efficiency tables (Version 62),” *Progress in Photovoltaics: Research and Applications*, vol. 31, 2023, pp. 651–66

[3] F. Blaabjerg, Y. Yang, D. Yang, and X. Wang, “Distributed power-generation systems and protection,” *Proceedings of the IEEE*, vol. 105, no. 7, 2017, pp. 1311–1331.

[4] G. M. El-Banby, A. M. El-Sayed, and M. A. Mosaad, “Photovoltaic system fault detection techniques: A review,” *Neural Computing and Applications*, vol. 35, 2023, pp. 20603–20630.

[5] B. Taghezouit, F. Harrou, and Y. Sun, “Model-based fault detection in photovoltaic systems: A comprehensive review and avenues for enhancement,” *Results in Engineering*, vol. 21, 2024, p. 101835.

[6] A. Ndiaye, C. M. Faye, and M. El Hajjaji, “Fault detection methods for photovoltaic systems: A review,” *Renewable Energy*, vol. 214, 2023, pp. 118–132.

[7] A. Mellit et al., “Deep learning-based approaches for fault detection in photovoltaic systems,” Energy Reports, vol. 8, 2022, pp. 123–135.

[8] W. Gao and R. J. Wai, “A novel fault identification method for photovoltaic array via convolutional neural network and residual gated recurrent unit,” *IEEE Access*, vol. 8, 2020, pp. 159493–159510.

[9] K. Yang, J. Zhang, and H. Wang, “Fault diagnosis of photovoltaic modules using convolutional neural networks and infrared images,” *Applied Energy*, vol. 268, 2020, Art. no. 114914.

[10] D. H. Katea and I. J. Awda, “Artificial intelligence and machine learning techniques for fault detection and predictive maintenance in photovoltaic systems: A comprehensive review,” *Al-Rafidain Journal of Engineering Sciences*, 2026.

[11] A. R. Jordehi, “Machine learning in power systems fault detection: A review,” IEEE Systems Journal, vol. 17, no. 2, 2023, pp. 1500–1510.

[12] S. Deitsch, F. Christlein, C. Berger, et al., “Automatic classification of defective photovoltaic module cells in electroluminescence images,” *IEEE Journal of Photovoltaics*, vol. 9, no. 2, 2019, pp. 443–452.

[13] A. F. Amiri, M. Ghofrani, and S. Mekhilef, “Fault detection and diagnosis of a photovoltaic system using deep learning,” *Sustainability*, vol. 16, no. 3, 2024, pp. 1012–1025.

[14] M. Jalal et al., “Deep learning approaches for visual faults diagnosis of photovoltaic systems,” *Results in Engineering*, 2024, Art. no. 102622.

[15] N. Salazar-Pena et al., “AI-powered dynamic fault detection and performance assessment in photovoltaic systems,” arXiv preprint, 2024.

[16] M. Bougoffa et al., “Hybrid deep learning for fault diagnosis in photovoltaic systems,” Machines, vol. 13, no. 5, 2025, pp. 378.

[17] M. Parvin et al., “Photovoltaic fault detection using ensemble learning and deep neural networks,” Results in Engineering, 2025.

[18] Vidhya, S., Balaji, M., Raj, E. F. I., & Kamaraj, V. (2023). Differential evolution algorithm supported random forest classifier for effective feature selection and classification of power quality disturbances. UPB Sci. Bull. Ser. C Electr. Eng. Comput. Sci., 85(1), 131-142.

[19] Q. Li, Y. Zhang, and X. Wang, “Incipient fault detection using LSTM-based deep learning approach,” *IEEE Transactions on Industrial Informatics*, vol. 19, 2023, pp. 200–210.

[20] M. M. Hasan, M. A. Islam, and M. R. Islam, “Fault detection in photovoltaic systems using LSTM networks,” *IEEE Access*, vol. 11, 2023, pp. 45000–45010.