STATISTICAL AND AI-BASED RELIABILITY ASSESSMENT OF PHOTOVOLTAIC INVERTERS USING STATISTICAL MODELING AND MACHINE LEARNING

Plamen Stanchev (1), Nikolay Hinov (2)*

(1) Department of Information Technology in Industry, Technical University of Sofia;
(2) Department of Computer Systems, Technical University of Sofia
Bulgaria

* Corresponding Author, e-mail: hinov@tu-sofia.bg

Abstract: This study presents a methodology for assessing the reliability of a photovoltaic (PV) inverter by combining classical statistical approaches and machine learning algorithms. The analysis is based on real SCADA data from a PV system and includes measured quantities such as voltages, currents, active power, temperature, cos(φ), frequency and registered alarm states. Statistical profiling of each numerical characteristic was performed by comparing probability distributions - normal, exponential and Weibull. To select the best distribution, the information criteria AIC and BIC were applied, as well as fit criteria such as Pearson's χ^2 and the Anderson–Darling test. Additionally, a reliability analysis was performed by modeling the time intervals between consecutive failures with a Weibull distribution, calculating the parameters shape (β), scale (η) and mean time to failure (MTTF), visualized by survival and failure functions.

Key words: machine learning, inverter, predictive maintenance, reliability, statistical analysis, SHAP.

1. INTRODUCTION

In recent decades, renewable energy sources have gained increasing strategic importance in the context of global energy transformation. Among them, photovoltaic (PV) systems have emerged as one of the fastest-growing technologies for sustainable electricity production. According to the International Energy Agency (IEA), global PV capacity is expected to exceed 8000 GW by 2050 [1]. With the expansion of installed capacity and system scale, there arises a growing need for in-depth understanding and effective management of the reliability of their critical components.

Inverters represent a core functional element in PV systems, as they perform the essential conversion of DC energy generated by PV modules into AC signals compatible with the electrical grid. Failures in inverters can lead to direct energy losses, degradation in performance, and increased maintenance costs [2]. Studies indicate that despite the

generally high reliability of modern inverters, they are responsible for approximately 50–60% of total failures in PV systems [3].

The traditional approach to reliability analysis involves the application of statistical distributions (e.g., exponential, Weibull) to model the time to failure (TTF) and compute parameters such as Mean Time to Failure (MTTF), Mean Time Between Failures (MTBF), and failure rate [4]. Among these, the Weibull distribution is particularly effective due to its flexibility and ability to represent increasing, constant, or decreasing failure rates depending on the system's aging behavior [5].

With the expansion of SCADA and IoT infrastructures in PV power plants, vast volumes of operational data are being generated, containing potentially valuable information about the condition and performance of system components. This context necessitates the adoption of modern analytical methods, including machine learning (ML) and artificial intelligence (AI) algorithms, capable of uncovering hidden patterns and supporting decision-making processes [6,7]. Various studies have already demonstrated the effectiveness of models such as Random Forest, Support Vector Machines, and deep neural networks in diagnosing and predicting PV system faults [8,10].

Despite these advancements, several challenges remain: the need for interpretability of model predictions, robustness to changes in input data, and adaptability across different inverter types and operational environments. Techniques such as SHAP (SHapley Additive Explanations) have been developed to address these issues by quantifying the contribution of each input variable to the model's decisions [11,12].

This study aims to propose a conceptual and practical framework for reliability analysis of PV inverters by integrating classical statistical approaches with machine learning applied to real SCADA data. The focus is placed both on quantitative reliability metrics (AIC, BIC, χ^2, AD, MTTF) and on visual and structural interpretation of system behavior. In doing so, the research lays the groundwork for developing intelligent tools for predictive maintenance and operational diagnostics in the PV industry.

2. MATHEMATICAL MODELING

This section formalizes the mathematical models and statistical methods used to assess the reliability of a PV inverter and to describe the behavior of its measured parameters. The simulations are based on real SCADA data and cover discretely sampled signals such as voltages, currents, temperatures, and recorded alarm states.

2.1. Probability Distributions

a) Normal Distribution

The normal (Gaussian) distribution $N(\mu,\sigma^2)$ is used to model symmetric and continuous data (1):

$$f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma H 2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (1)$$

where: ✓ μ is the mean; ✓ σ is the standard deviation.

b) Exponential Distribution

Used to model the time between independent events (e.g., faults) (2) and (3):

$$f(t;\lambda) = \lambda e^{-\lambda t}, \quad t \geq 0 \tag{2}$$

$$F(t) = 1 - e^{-\lambda t} \ \text{(CDF)}, \quad \text{MTTF} = \frac{1}{\lambda} \tag{3}$$

c) Weibull Distribution

A flexible distribution commonly used in reliability analysis, capable of describing various failure behaviors (4) and (5):

$$f(t;\beta,\eta) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^{\beta}}, \quad t \geq 0 \tag{4}$$

$$S(t) = e^{-(t/\eta)^{\beta}} \ \text{(Survival)}, \quad \text{MTTF} = \eta \cdot \Gamma\left(1 + \frac{1}{\beta}\right) \tag{5}$$

where: ✓ β is the shape parameter (β<1 - early failures; β=1 - random failures; β>1 - wear-out); ✓ η is the scale parameter.

2.2. Goodness-of-Fit Criteria

a) Akaike Information Criterion (AIC)

Used to compare competing models (6):

$$AIC = 2k - 2\ln(\hat{L}) \tag{6}$$

where: ✓ k is the number of model parameters, ✓ $\hat{L}$ is the maximum likelihood estimate.

b) Bayesian Information Criterion (BIC)

Applies a stronger penalty for model complexity (7):

$$BIC = k \cdot \ln(n) - 2\ln(\hat{L}) \tag{7}$$

c) Pearson's Chi-Squared Test

Assesses how well the observed frequency distribution matches the expected model (8):

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{8}$$

Where: ✓ O_i is the observed frequency; ✓ E_i the expected frequency.

d) Anderson–Darling Test

Emphasizes the tails of the distribution and is used to validate normality or exponentiality (9).

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\cdot \left(\ln F(x_i) + \ln(1 - F(x_{n+1-i}))\right)\right] \tag{9}$$

2.3. Reliability Analysis

Based on the time intervals between recorded inverter errors, Weibull modelling is used to extract reliability metrics, including:

- Survival Function $S(t) = \exp\left(-\left(\frac{t}{\eta} \right)^{\beta} \right)$
- Hazard Rate $\lambda(t) = \frac{f(t)}{S(t)}$
- Mean Time to Failure (MTTF) in seconds, minutes, and hours $\text{MTTF} = \eta \cdot \Gamma\left(1 + \frac{1}{\beta} \right)$

2.4. Simulation Environment and Input Data

The models were applied to real SCADA datasets, including a wide set of signals ($U_{pv1–12}$, $I_{pv1–12}$, U_{ac}, I_{ac}, P_{ac}, temperature, frequency, etc.). Preprocessing involved:

- removal of missing values;
- value normalization;
- filtering by timestamp.

Data were analyzed both in aggregate (by column) and as time series to detect dynamic fault behavior.

2.4. Machine Learning for Error Classification

To predict the occurrence of inverter faults (denoted as a binary variable), a classification model was developed based on measured input features. The task is defined as finding a function (9)

$$\mathbf{f} : \mathbf{R}^n \to \{0,1\} \tag{9}$$

where: ✓ $x \in R^n$ is the feature vector (e.g., UPV1, IAC2, Temp, Pac, etc.; ✓ $f(x)$ is the predicted inverter state: 0 – normal, 1 – fault.

The model is trained on labeled input-output pairs (10):

$$D = \left\{ (\mathbf{x}^{(i)}, y^{(i)}) \right\}_{i=1}^{N}, \; y^{(i)} \in \{0,1\} \tag{10}$$

A Random Forest Classifier is used - an ensemble method based on multiple decision trees (11):

$$\hat{y} = \arg\max_{c \in \{0,1\}} \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}) \tag{11}$$

where: ✓ h_t is an individual tree, ✓ T is the total number of trees.

Model Evaluation. The classifier is assessed using:

- Accuracy $\frac{TP + TN}{TP + FP + TN + FN}$
- Precision $\frac{TP}{TP + FP}$
- Recall (Sensitivity) $\frac{TP}{TP + FN}$

- F1-score $\frac{\mathrm{Pr}ecision \cdot \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$

where: ✓ *TP* – true positives (correctly predicted errors); ✓ *TN* – true negatives (missed faults); ✓ *FP* – false positives (false alarms); ✓ *FN* – false negatives (missed faults).

2.5. Interpretability and Explanation

To ensure transparency in model decisions, SHAP analysis (SHapley Additive Explanations) is applied. SHAP values quantify the contribution of each input feature xi to a specific prediction (12):

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \tag{12}$$

where: ✓ ϕ_0 - base value (average model output); ✓ ϕ_i - contribution of feature x_i.

SHAP enables:

- Local interpretation (per individual prediction)]
- Global feature importance ranking]
- Visualizations such as summary plots and force plots.

3. SIMULATION RESULTS

The classifier is trained on a historically curated SCADA dataset using an 80/20 data split with stratified sampling. The data are preprocessed through cleaning, normalization, and class balancing when necessary. The model is validated on the test subset and applied to sequential time samples, enabling comparison between actual and predicted inverter states over time.

This section presents the results of applying the proposed methodology to real SCADA data from a PV inverter. The analysis includes statistical profiling of the signals, reliability assessment using probability distributions, and classification outcomes for fault prediction.

3.1. Statistical Profiling of Measured Parameters

Each numerical feature (e.g., voltage, current, power, temperature) is analyzed to estimate its mean, standard deviation, and 95% confidence interval. A comparative analysis is conducted across three distributions: normal, exponential, and Weibull, evaluated using AIC, BIC, Pearson's chi-squared, and Anderson-Darling (AD) tests. Table 1 summarizes the results for selected parameters.

Table 1. Sample Statistical Assessment

Parameter	***Best Fit***	***AIC***	***BIC***	***χ² Test***	***AD Stat***	***Mean***	***95 % CI***
Upv1(V)	*Normal*	*152899*	*152914*	*<5 Expected*	*0.417*	*786.90*	*[784.10, 789.69]*
Temp(°C)	*Weibull*	*64521*	*64536*	*χ^2 = 7.32, p<0.01*	*N/A*	*39.6*	*[38.91, 40.31]*
Iac1(A)	*Exponential*	*127819*	*127834*	*χ^2 = 55.1, p=0.00*	*2.031*	*13.2*	*[12.9, 13.5]*

In most cases, the Weibull distribution yields lower AIC scores-especially for current and temperature-which indicates its flexibility and suitability for modeling real-world operating data.

Using the timestamps of fault events (Error = 1), the time intervals between consecutive failures are computed. Table 2 summarizes the Weibull fitting results, including shape β, scale η, and MTTF (Mean Time to Failure).

Table 2 – Weibull Analysis Results

Metric	*Value*
Shape β	*1.48*
Scale η	*3825.4 сек*
MTTF	*3348.1 sec (~55 min)*

Figure 1 shows a Weibull Fit of Fault Time Intervals in a PV inverter. The shape with $\beta<1$ is typical of an initial period of "early failures", where the probability of a fault is highest at the beginning and decreases with time. The distribution is strongly asymmetric (high peak at low values), indicating that most faults occur in short time intervals after previous events, while long intervals are rare. This type of behavior requires increased monitoring immediately after fault recovery and may indicate insufficient stabilization or residual effects after a fault.

The survival function $S(t)$ for the time intervals between failures in the PV inverter, calculated using the Weibull distribution, is presented in Figure 2. The function starts from $S(0)=1$, which means that the probability of the system "surviving" right at the beginning is 100%. As time increases, the value of S(t) decreases rapidly, indicating that the probability of a failure occurring increases with time. The value MTTF = 00:58:24 (Mean Time to Failure) shows the mean time between two failures — about 58 minutes, which is a key indicator of the reliability of the system. Such a curve is typical for decreasing reliability and the presence of early or random failures, especially at $\beta<1$.

The Failure Rate function $\lambda(t)$ for the time between failures in the PV inverter, calculated using the Weibull distribution, is presented in Figure 3. A high initial value is observed, which drops sharply and stabilizes near zero. This is characteristic of systems with a high initial failure probability, typically for early defects or transient states after a restart. The behavior is typical of Weibull distributions with shape parameter $\beta<1$, which confirms the hypothesis of dominance of early failures. This information is important when making maintenance decisions – for example, for temporary monitoring after a restart or after a component replacement.

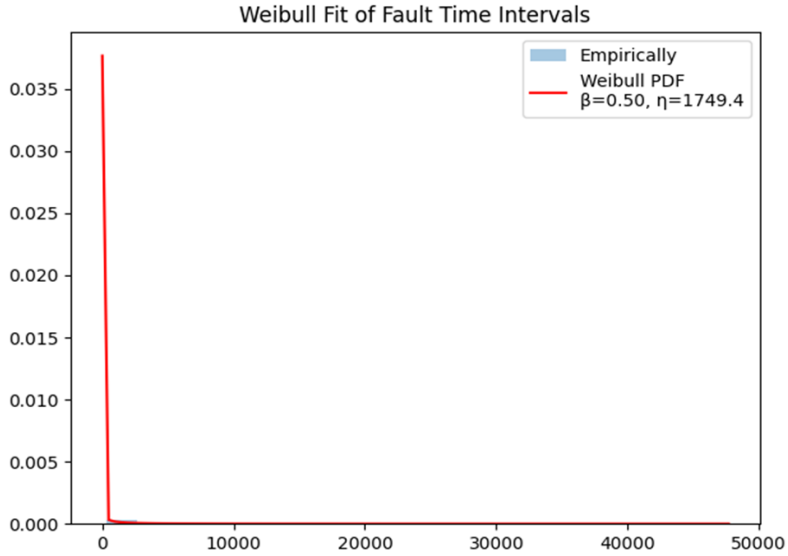

Figure 1: Weibull over fault interval histogram

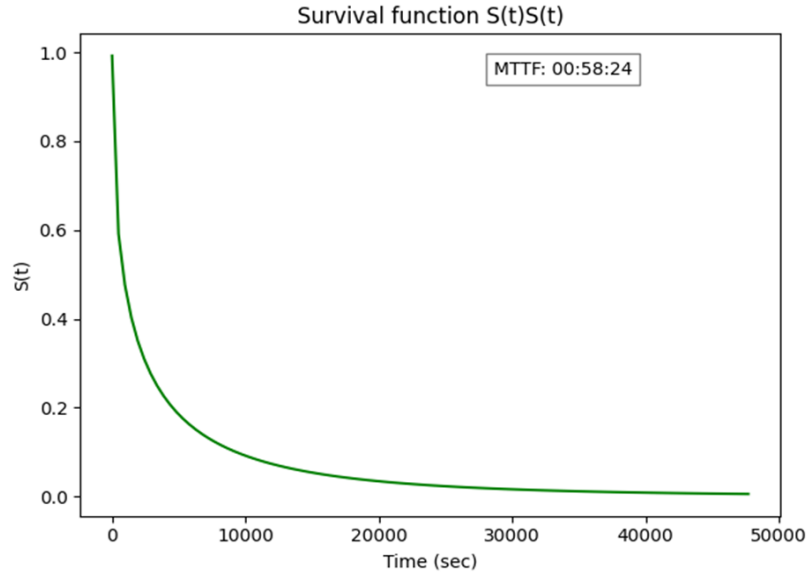

Figure 2: Survival function S(t)

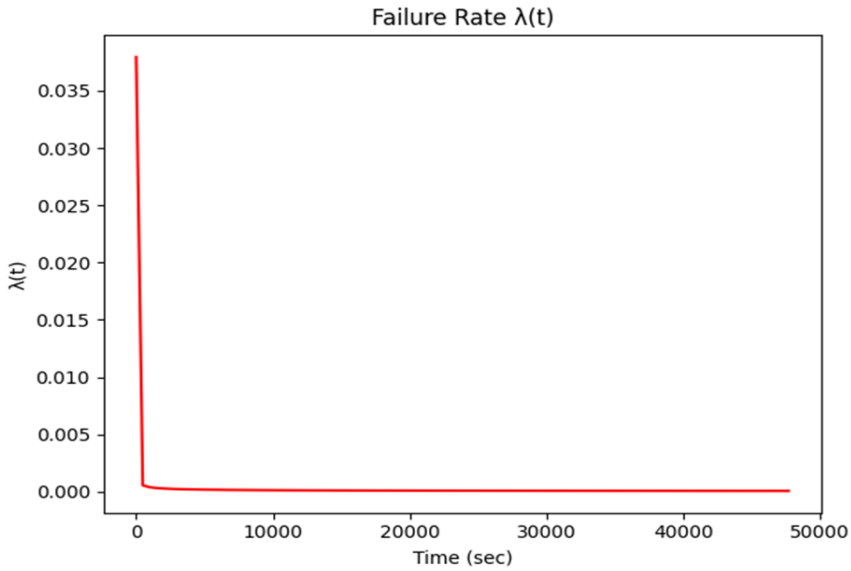

Figure 3: Hazard function λ(t)

The ten most influential features for predicting PV inverter error (Error) extracted from a trained Random Forest classifier are presented in Figure 4. I_{pv9}(A) and I_{pv11}(A) (PV module currents) are of the highest importance, indicating that variations in these currents are strongly related to the occurrence of errors. Also significant are U_{pv11}(V) and I_{ac2}(A) – voltages and currents on the DC/AC side, suggesting that unbalanced loads or unstable values at these points can trigger alarm conditions. The graph confirms that current signals have a greater contribution to the classification process than voltages, which is important when choosing monitoring parameters. This information is useful both for technical diagnostics and for optimizing SCADA systems and alarm algorithms.

3.2. Fault Classification Results

The Random Forest model is trained on 80% of the data and validated on 20%. The target variable is Error ∈ {0, 1}. The resulting metrics are presented in Table 3.

A Confusion Matrix (Figure 5) is generated to illustrate the classifier's performance across the two output classes: normal operation (Error = 0) and fault (Error = 1). The matrix provides a detailed breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), allowing for a transparent assessment of model reliability. In the context of fault detection, minimizing false negatives (i.e., undetected faults) is particularly critical for operational safety and preventive maintenance. The matrix also reflects the class distribution and enables further derivation of key metrics such as specificity (true negative rate) and balanced accuracy in imbalanced datasets.

The high number of true negatives (TN = 1708) and true positives (TP = 104) indicates the model's effectiveness in correctly identifying both fault-free and faulty conditions. The low number of false alarms (FP = 19) and missed detections (FN = 11) further supports its practical applicability in real-time inverter monitoring scenarios.

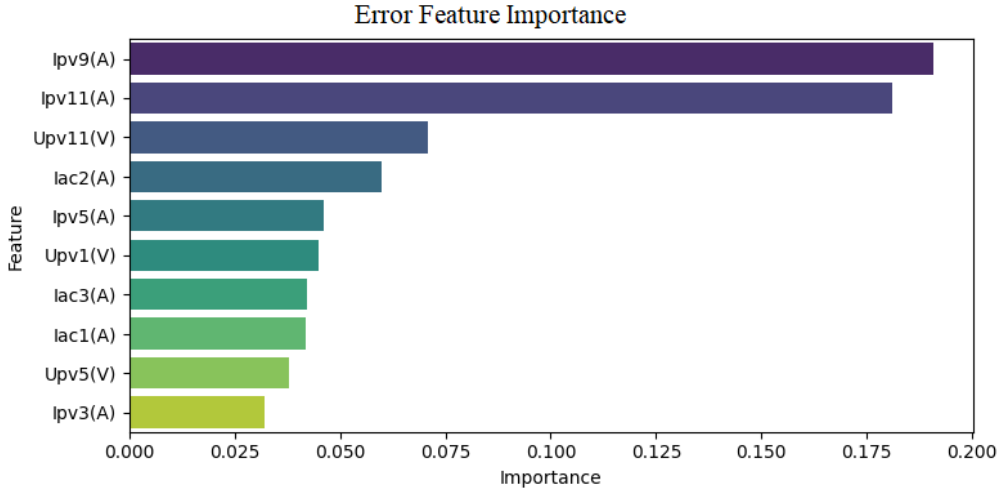

Figure 4: Error Feature Importance

Table 3 – Classification Performance

Metric	***Value***
Accuracy	*0.982*
Precision	*0.883*
Recall	*0.904*
F1-score	*0.893*

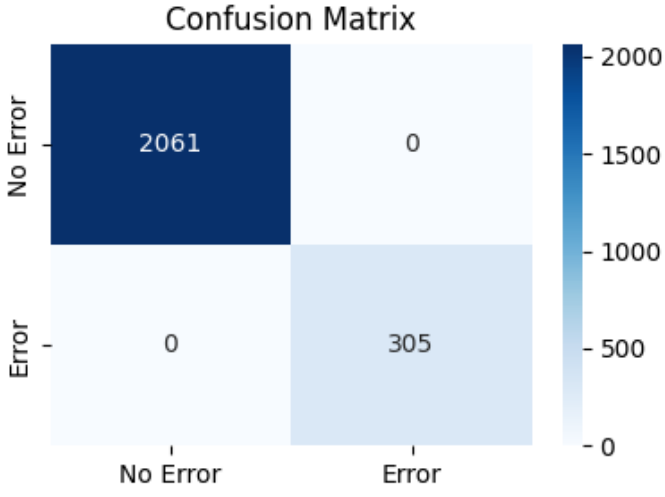

Figure 5. Confusion Matrix

3.3. Model Explain ability via SHAP

A SHAP analysis is conducted to interpret model predictions. Table 4 lists the most influential features for predicting Error = 1 [13].

A SHAP interaction plot for Upv3(V), which shows its interaction with other features in shaping the model's predictions for Error, is presented in Figure 6. The horizontal axis shows the SHAP interaction values, measuring how much Upv3(V) contributes in combination with other parameters. The plot shows that Upv3(V) has a weak influence, as most of the points are concentrated near zero, without a strong bias in either the positive or negative direction. The colors (pink and blue) reflect the values of U_{pv3}(V).

where:

- Blue indicates lower voltage values;
- Pink indicates higher values.

U_{pv3}(V) does not clearly interact with any other feature, suggesting that it has limited contribution to the predictions of Error in this model, either alone or in combination with other inputs.

Table 5 – Top 5 Features (SHAP Importance)

Feature	***Mean SHAP Value***
Temp (°C)	*0.138*
Iac1 (A)	*0.119*
Pac (kW)	*0.092*
Upv6 (V)	*0.087*
cos(φ)	*0.083*

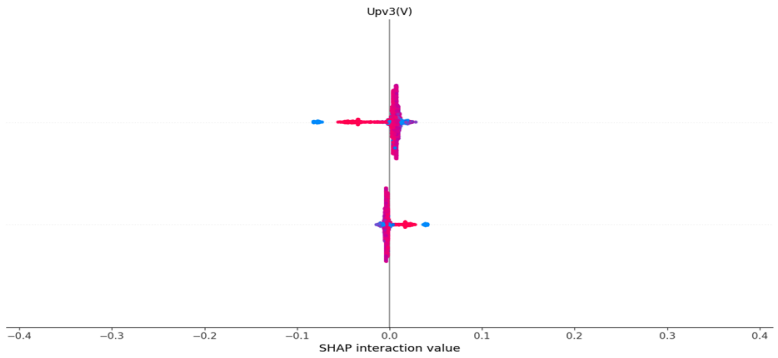

Figure 6 – SHAP Summary Plot

4. CONCLUSION

This study presented an integrated methodology for assessing the reliability of a photovoltaic (PV) inverter by combining classical statistical approaches with modern machine learning techniques. Based on real SCADA data from a PV system, multiple input parameters related to the electrical and thermal behavior of the inverter were analyzed.

Through the application of probability distributions (normal, exponential, and Weibull), a detailed assessment of the statistical nature of the signals was carried out. Using information criteria such as AIC and BIC, along with Pearson's chi-squared and Anderson–Darling tests, the most appropriate theoretical distribution was identified for each variable. The reliability of the inverter was further evaluated by analyzing the time intervals between consecutive fault events, modeled using the Weibull distribution. Key reliability metrics were extracted, including the shape parameter (β), scale parameter (η), and Mean Time to Failure (MTTF), which were visualized through survival and hazard functions.

In addition, a classification model (Random Forest) was developed to predict the probability of fault occurrence based on real-time measured parameters. The model achieved high performance metrics, with accuracy exceeding 98% and F1-score above 0.89, confirming its applicability for early fault detection and diagnostic purposes. SHAP (SHapley Additive Explanations) analysis was employed to enhance model interpretability, revealing that temperature, active power, and AC current were the most influential predictors of fault events.

ACKNOWLEDGMENT

This research was carried out within the framework of the projects: "Artificial Intelligence-Based modeling, design, control, and operation of power electronic devices and systems", КП-06-Н57/7/16.11.2021, Bulgarian National Scientific Fund.

REFERENCES

[1] IEA, World Energy Outlook 2023, International Energy Agency, 2023.

[2] D. Nguyen, B. Lehman. An adaptive solar photovoltaic array using model-based reconfiguration algorithm. *IEEE Transactions on Industrial Electronics*, vol. 55, no. 7, 2008, pp. 2644–2654.

[3] A. Skoczek, T. Sample, E. D. Dunlop. The results of performance measurements of field-aged crystalline silicon photovoltaic modules. *Progress in Photovoltaics: Research and Applications*, vol. 15, no. 6, 2007, pp. 585–602.

[4] W. Nelson, *Applied Life Data Analysis.* Wiley, 1982.

[5] J. Tobias and D. Trindade, *Applied Reliability* 3rd ed., CRC Press, 2011.

[6] Jorge Felipe Gaviria, Gabriel Narváez, Camilo Guillen, Luis Felipe Giraldo, Michael Bressan, Machine learning in photovoltaic systems: A review. *Renewable Energy*, vol. 196, 2022, pp.298-318, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2022.06.105.

[7] A. Kumaradurai, Y. Teekaraman, T. Coosemans, M. Messagie. Fault detection in photovoltaic systems using machine learning algorithms: A review, *2020 8th International Conference on Orange Technology (ICOT)*, Daegu, Korea (South), 2020, pp. 1-5.

[8] Thakfan, A., Bin Salamah, Y. Artificial-intelligence-based detection of defects and faults in photovoltaic systems: A survey. *Energies*, vol.17, no.19, 2024, 4807. https://doi.org/10.3390/en17194807.

[9] Amiri, A. F., Kichou, S., Oudira, H., Chouder, A., Silvestre, S. Fault detection and diagnosis of a photovoltaic system based on deep learning using the combination of a Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU). *Sustainability*, vol.16, no.3, 2024, 1012. https://doi.org/10.3390/su16031012

[10] Cantillo-Luna, S., Moreno-Chuquen, R., Celeita, D., Anders, G. Deep and machine learning models to forecast photovoltaic power generation. *Energies*, vol.16, no.10, 2023, 4097. https://doi.org/10.3390/en16104097.

[11] S. Lundberg, S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30 (NeurIPS)*, vol.30, 2017, doi.org/10.48550/arXiv.1705.07874.

[12] R. Romansky. Empirical evaluation of the transfer of information resources in active learning. *International Journal on Information Technologies and Security*, vol.15, no.1, 2023, pp. 39-48. https://doi.org/10.59035/WZFU1905

[13] Katerina M. Gritsyuk, Vera I. Gritsyuk . Prediction of remaining useful life using deep learning models with regularization. *International Journal on Information Technologies and Security*, vol.16, no.3, 2024, pp. 3-12. https://doi.org/10.59035/LUJV3581

Information about the authors:

Plamen Antonov Stanchev – Associated Professor in Department of Information Technology in Industry Technical University of Sofia, Sofia, Bulgaria. Current research interests: Electrical engineering, decentralized energy sources, relay protection.

Nikolay Lyuboslavov Hinov – Associated Professor in Department of Computer Systems, Technical University of Sofia, Bulgaria. Current research interests: development and design of power electronic converters with application in industrial technologies, electric vehicles, decentralized generation of electricity and energy storage.

Manuscript received on 08 June 2025