eng

MDPI

Article

A Hybrid Artificial Intelligence for Fault Detection and Diagnosis of Photovoltaic Systems Using Autoencoders and Random Forests Classifiers

Katlego Ratsheola [1], Ditiro Setlhaolo [1], Akhtar Rasool [1,*], Ahmed Ali [2,*] and Nkateko Eshias Mabunda [2]

[1] Department of Electrical Engineering, University of Botswana, Gaborone UB0061, Botswana; ratsheolak@ub.ac.bw (K.R.); setlhaolo@ub.ac.bw (D.S.)
[2] Department of Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg 2092, South Africa; nkatekom@uj.ac.za
* Correspondence: drarasool83@gmail.com or rasoola@ub.ac.bw (A.R.); aali@uj.ac.za (A.A.)

Abstract

The increasing sophistication of grid-connected photovoltaic (GCPV) systems necessitates advanced fault detection and diagnosis (FDD) methods to ensure operation efficiency and security. In this paper, a novel two-stage hybrid AI architecture is analyzed that couples an autoencoder using Long Short-Term Memory (LSTM) for unsupervised anomaly detection with an RF classifier for focused fault diagnosis. The architecture is critically compared to that of a baseline-only RF baseline on a synthetic dataset. The results of this two-stage hybrid AI show a strong overall accuracy of (83.1%). The hybrid model's first stage trains only on unlabeled healthy data, reducing the reliance on extensive and often unavailable labeled fault datasets. This design has the safety-critical advantage of marking unfamiliar faults as anomalies instead of committing to a misclassification. By integrating anomaly detection with classification, the architecture enables early stage screening of faults and targeted categorization, even in data-scarce scenarios. This offers a scalable, interpretable solution suitable for deployment in real-world GCPV systems where robustness and early detection are critical. While the method exhibits reduced sensitivity to subtle or recurring faults, it demonstrates strong reliability in confidently detecting distinct and significant anomalies. Additionally, the approach improves interpretability, facilitating clearer identification of performance constraints such as the autoencoder's moderate fault sensitivity (AUC = 0.61). This study confirms the hybrid approach as a very promising FDD solution, in which the architectural advantages of safety and maintainability offer a more worthwhile proposition to real-world systems than incremental improvements in a single accuracy measure.

Keywords: photovoltaic systems; fault detection and diagnosis (FDD); artificial intelligence; LSTM; autoencoder; random forest; anomaly detection

check for updates

Academic Editor: Antonio Gil Bravo

Received: 20 June 2025
Revised: 28 July 2025
Accepted: 6 August 2025
Published: 1 October 2025

Citation: Ratsheola, K.; Setlhaolo, D.; Rasool, A.; Ali, A.; Mabunda, N.E. A Hybrid Artificial Intelligence for Fault Detection and Diagnosis of Photovoltaic Systems Using Autoencoders and Random Forests Classifiers. *Eng* **2025**, *6*, 254. https://doi.org/10.3390/eng6100254

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The global energy landscape is undergoing a transformation, driven by the dual imperatives of mitigating climate change and meeting the growing demand for cleaner energy [1]. Solar photovoltaic (PV) technology is at the forefront of this transition, aided by rising economic competitiveness, technological superiority, and inherent scalability [2]. This global dominance is quantitatively evident in the data: in 2024, solar photovoltaic energy accounted for the largest share of global renewable capacity, reaching 1865 GW, with

a record-breaking 452 GW of new capacity added in that year alone [3,4]. Grid-connected photovoltaic systems (GCPVs), in general, are the main drivers behind the phenomenon, serving as the backbone of distributed and utility-scale power generation [5].

However, this rapid deployment places the operational challenges of ensuring long-term reliability, safety, and optimal performance firmly in the spotlight [6]. GCPV systems experience a wide variety of faults due to manufacturing defects, environmental stress, and material aging, and these can range from module degradation to system-level inverter faults [7,8]. The consequences of undetected faults are severe, leading to high energy losses, expensive maintenance, and low economic returns. Moreover, certain electrical faults, i.e., arc faults, have serious safety implications [9]. Thus, effective fault detection and diagnosis (FDD) is an indispensable aspect of modern PV asset management.

While traditional FDD methods like visual inspection and thermography have been effective, they are not well suited for the scale and complexity of modern PV power plants due to their time-consuming nature, limited fault coverage, and high false alarm rates [10–12]. These limitations have necessitated the shift to data-driven FDD methods that are AI-based. However, the practical deployment of AI models is faced with a critical challenge that is pivotal: a trade-off between deploying very accurate but possibly "brittle" models that performs poorly on unknown faults, and building systems that are robust by design, safe, and viable for real-world deployment.

To fill this gap, this paper proposes and critically analyzes a two-stage hybrid FDD system that combines the integration of unsupervised learning and supervised learning. The contributions of this paper are as follows:

1. We analyses a specific **two-stage, anomaly-first framework** that involves the use of an LSTM-based autoencoder for initial anomaly filtering and a random forest classifier for accurate fault diagnosis. This framework faces the challenge of low amounts of labeled data front-on by training on healthy data.
2. We carry out an explicit evaluation of the **inherent engineering trade-offs** of this approach. Instead of optimally maximizing one accuracy metric, we demonstrate how the model sacrifices some sensitivity to very minor faults in order to achieve safety and robustness to unexpected or novel events.
3. We guarantee the **interpretability and maintainability** of the hybrid model by linking its performance bottlenecks directly with the sensitivity of the unsupervised first stage. This analysis prescribes a straightforward and actionable trajectory for future improvement.

This work argues that the architectural advantages of safety, data efficiency, and interpretability offered by this hybrid approach present a more compelling proposition for real-world FDD systems than marginal improvements in aggregate accuracy alone.

The remainder of this paper is organized as follows: Section 2 provides a review of related works in AI-based FDD. Section 3 introduces the proposed hybrid FDD framework in detail. Section 4 presents the experimental methodology, including the simulation environment and training process. Section 5 presents and analyzes the experimental results. Section 6 discusses the findings. Finally, Section 7 concludes the paper and offers suggestions for future work.

2. Literature Review

The application of artificial intelligence (AI) and machine learning (ML) techniques to the fault detection and diagnosis (FDD) of photovoltaic (PV) systems has emerged as a major area of research over the past decade, and numerous studies are available that validate the potential of data-driven methods in overcoming the limitations of traditional methods [7,13]. The literature can be categorized into three broad categories based on

supervised, unsupervised, and hybrid learning paradigms with specific strengths and practical limitations.

Supervised learning methods like support vector machines (SVMs), artificial neural networks (ANNs), and ensemble classifiers like random forest (RF) have been widely applied for fault classification [14]. These models can support high diagnostic accuracy if they are trained on comprehensive datasets. Their major drawback, however, lies in their dependence on large, well-labeled fault databases, which must have a reasonable number of instances for each potential fault under various operating conditions. In reality, tagging and compiling such enormous datasets for real PV systems is famously impractical, costly, and frequently impossible, and this presents a major obstacle to the implementation of solely supervised models [15,16].

To circumvent the challenge of labeled data insufficiency, researchers have investigated unsupervised learning methods, including techniques such as principal component analysis (PCA), clustering, and autoencoders, taking advantage of learning the typical working behavior of a system from easily accessible, unlabeled health data. Any significant departure from such learned norm is then identified as an anomaly. While this approach effectively addresses the issue of having labeled data for the first fault detection, it has one major disadvantage: it can tell whether a fault occurred. However, in the majority of instances, it often fails to classify the specific *type* of fault, thus providing incomplete data for maintenance and repair operations [16].

Noticing the complementary strengths and weaknesses of these two paradigms, the scientific community now increasingly depends on **hybrid FDD models**. These models attempt to combine the advantages of both approaches, e.g., an unsupervised method for initial anomaly detection and a supervised method for specific classification of identified anomalies. Architectures that combine autoencoders with other classifiers are particularly common, as the autoencoder can efficiently learn from time-series data [17]. The use of an LSTM autoencoder in combination with an RF classifier, used in this work, is referred to as a critically powerful association due to its ability to handle temporal dependencies and achieve reliable classification [18].

Despite the significant progress achieved in hybrid model development for fault detection and diagnosis, one major research gap remains. The majority of existing studies primarily aim to improve overall performance metrics such as overall classification accuracy or F1-score without reporting much about the inherent architectural trade-offs of such models. In particular, the impact of the unsupervised detection phase on the downstream supervised classification remains insufficiently examined. Furthermore, significant practical questions are too frequently overlooked: How does the sensitivity of the anomaly detection step influence the overall pipeline robustness? How does the model behave under unseen or emerging fault scenarios? To what extent can the model's failure modes be interpreted and trusted by domain practitioners in real-world settings? This paper addresses these questions through a critical system level analysis of a two-stage hybrid architecture. The critique goes beyond standard performance evaluation to an examination of model design feasibility in terms of safety, interpretability, and data efficiency in actual practice.

3. Proposed Hybrid FDD Framework

This study proposes a two-stage hybrid AI-based fault detection and diagnosis (FDD) framework for photovoltaic (PV) systems that integrates the strengths of both supervised and unsupervised learning. Stage one involves learning a Long Short-Term Memory (LSTM) autoencoder from normal PV system operating patterns in order to detect normality and anomalies by means of reconstruction error. Stage two employs a random forest (RF) classifier to distinguish true faults from common anomalies. A two-stage approach is

utilized to improve the automatic fault detection's specificity and reliability, particularly in noisy or complex conditions where the false alarms are costly. Figure 1 below illustrates how the model works when a new data inputs are fed into the model for evaluation.

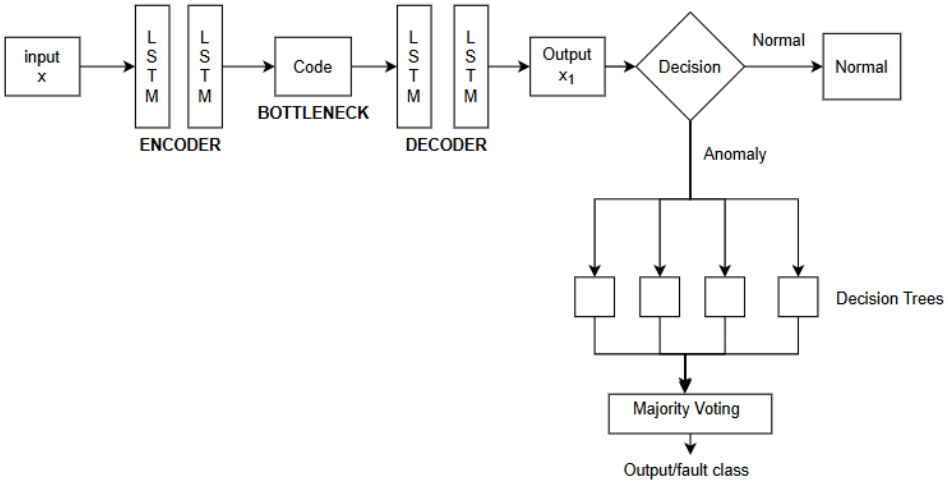


Figure 1. Model framework representation.

3.1. Stage 1: Anomaly Detection Using LSTM Autoencoders

In the initial phase of our hybrid fault detection and diagnosis (FDD) system, an unsupervised anomaly detection model based on an LSTM autoencoder is deployed. This model combines representation learning and sequential dependency modeling. The model captures the multivariate time-series data produced by grid-connected photovoltaic (GCPV) systems, i.e., voltage, current, and irradiance signals. LSTM autoencoders have been found to be superior to conventional sequence models as they capture both temporal dependencies and spatial correlations in sensor data streams [19,20].

The architecture is made up of stacked LSTM layers for both the encoder and the decoder. The encoder module compresses the high-dimensional input sequence into a low-dimensional latent vector, maintaining the salient temporal characteristics. The decoder module then reconstructs the original multivariate sequence from the latent space. Under this arrangement, the model can learn compact digital signatures of the normal system over time [19,21].

LSTM networks offer memory cells and gating mechanisms (input, forget, and output gates), which enable them to model long-term dependencies in time-series data—a crucial capability for detecting slow-developing faults in PV systems [22,23]. During training, the model is only exposed to healthy operation data. At inference time, reconstruction errors are calculated; unusually large errors indicate sequences that deviate from learned healthy patterns and may signal a fault [19,22].

A statistical threshold for anomaly detection is determined from the error distribution on a held-out validation set. Any new sequence exceeding this threshold is flagged as anomalous and forwarded to the second stage. This thresholded reconstruction-based approach is well established in the literature and avoids the need for extensive labeled fault data in the initial detection stage [19,20].

3.2. Stage 2: Supervised Fault Classification Using Random Forest

Once anomalies are flagged in Stage 1, the data are passed to a random forest (RF) classifier for precise fault diagnosis. RFs are ensemble models that create multiple decision trees using random subsets of features and training data, enabling robust non-linear classification and reducing overfitting [24,25].

For this approach, the RF classifier is trained on a labeled dataset for specific fault classes such as partial shading, degradation, line-to-line faults, and normal conditions to avoid false alarms. Either the raw multivariate sequence or calculated features like reconstruction error and latent vectors of the LSTM autoencoder are fed into the second stage as input. The RF model then generates a discrete fault label with targeted diagnostics. This two-stage pipeline leverages the power of generalization of unsupervised anomaly detection and the label specificity of supervised classification [24].

3.2.1. Classifier Rationale and Principles

The choice of the random forest classifier for this job is based on its demonstrated high accuracy and robustness in complex classification problems, making it highly suited for FDD applications in GCPV systems [18,26]. RF is an ensemble learning method, meaning its predictive power does not come from a single model, but rather from the collective intelligence of a large number of individual decision trees constructed during training [27].

The algorithm's resistance to overfitting and ability to handle high-dimensional data are attributed to two important methods de-correlating the component trees:

- **Bootstrap Aggregating (Bagging):** Rather than having all trees trained on the same exact dataset, each tree in the "forest" is trained on a unique random sample of the training data, which is selected with replacement. This guarantees diversity of the trees, as each learns from a slightly different view of the data, which improves the final model's generalization performance [28].
- **Feature Randomness:** At every split point in one decision tree, the algorithm does not look at all possible features (voltage, current, irradiance, temperature). It only looks at a random set of those features that can make few highly predictive features dominate the structure of all the trees. This makes the model learn the importance of a wide range of input variables, making it more resilient. By this, it leads to a more effective classifier [27].

3.2.2. Classification Mechanism

The final classification from the trained random forest model is determined through a democratic "majority vote" process. When a new anomalous input is presented to the model, it is passed down every single decision tree in the forest. Each tree independently assigns a class prediction, casting a vote for a specific fault type. The output of the RF model is the most voted fault class within a forest of trees [18]. Internal decisions in each tree are guided by more sophisticated splitting criteria. An effort is made by the algorithm to make splits where data separates most optimally into multiple classes, something that is really optimally tuned by minimizing something like a Gini impurity or by maximizing information gain. These design rules assure that for this problem in question of classification, every tree is performance-optimized [27].

4. Methodology

A simulation experimental procedure was used to validate the hybrid FDD model suggested. To be reproducible and consistent, input data (voltage, current, irradiance, and temperature) were normalized into a standard scaler with a mean of zero and one standard deviation. The procedure entails synthetic dataset generation, model development, training, and quantitative evaluation.

4.1. Simulation Environment and Dataset Generation

A virtual GCPV system representing a 104 kWp commercial PV array was simulated using Python 3.13.4 and the `pvlib` library. Time-series data for both healthy and faulty

conditions were synthesized by varying irradiance, temperature, and module parameters using the single-diode model.

Simulated fault classes included the following:

- Short Circuit.
- Open Circuit.
- Module Degradation.
- Partial Shading.
- Line-to-Line Fault.

The selection of these specific fault classes is intentional and highlights faults with an operationally significant impact and obvious, simulatable electrical signatures. Short-circuit, open-circuit, and line-to-line faults are abrupt, severe electrical failures that are safety-relevant. Module degradation is a widespread, long-term performance decline, while partial shading is a typical environmental factor that can cause complex signatures, simulating other faults. These five fault types thus constitute a robust and rigorous benchmark against which to measure the core skills of an FDD architecture. Other faults of note, such as DC arcs, were not considered because of the extreme complexity and randomness of their signatures, which would require complex electromagnetic transient modeling beyond the purview of this study.

This resulted in a labeled dataset of several thousand samples covering five distinct operational states. The data consisted of multivariate sensor sequences (voltage, current, irradiance, and temperature) at 5 min intervals, designed to reflect realistic PV system behavior.

4.2. Hybrid FDD Framework Architecture and Training

Two models were developed and compared:

1. **Baseline Model:** A standard RF classifier trained directly on all labeled sequences, representing a traditional one-stage supervised approach.
2. **Hybrid Model:**
    - *Stage 1—LSTM Autoencoder:* Trained on healthy data only, with 64 LSTM units, 5 time-step input sequences, and a latent space of dimension 32. The number of LSTM units and RF estimators were determined through a standard hyperparameter tuning process using a grid search methodology on the validation set to optimize the F1-score.
    - *Stage 2–Random Forest Classifier:* Trained on the full labeled dataset, consisting of 100 estimators with a maximum depth of 20. Only data flagged by Stage 1 was passed for classification.

After training the LSTM autoencoder on normal operating data, the reconstruction error for each time-series sequence was determined from the mean squared error (MSE) between the reconstructed and the original signals. To set a realistic anomaly threshold, we evaluated the reconstruction errors on the validation subset of normal data. The reconstruction errors ranged from 0.002 to 0.178 with a mean value of 0.047 and a standard deviation value of 0.031. According to standard practice, the threshold was set as the 95th percentile of the validation error distribution, which turned out to be 0.103. Any sample above this threshold was classified as anomalous and sent to the second stage for fault classification.

4.3. Dataset Partitioning and Test Set

In order to ensure a rigorous and unbiased test, the whole synthesized dataset was split into three independent, non-overlapping subsets: a **training set (70%)** for training

the models, a **validation set (15%)** for hyperparameter tuning and finding the anomaly threshold, and a **test set (15%)** for the final testing of performance. The test set was never exposed to the models during any training or tuning phase.

The detailed composition of this unseen test dataset, drawn from ground-truth labels of the data used for the confusion matrix is shown in Table 1. The dataset comprises a total of 4990 samples, divided between the normal operating state and the five synthetic fault classes. This structure ensures that the evaluation metrics detailed within Section 5 provide a balanced and complete assessment of the models' diagnostic capability for each unique condition.

Table 1. Composition of the unseen test dataset.

Class (True Label)	Number of Samples
Normal	2505
Degraded	533
Line-to-Line Fault	475
Open Circuit	503
Shaded	496
Short Circuit	488
Total	**4990**

4.4. Experimental Design and Evaluation Metrics

Performance was assessed using a dedicated test set unseen during training. Metrics included the following:

- **Accuracy, Precision, Recall, F1-Score:** Computed per fault class to evaluate classification performance.
- **AUC-ROC:** Used to evaluate the autoencoder's ability to distinguish normal from anomalous data.

These metrics collectively measure both detection effectiveness and classification accuracy, providing insights into the practical utility of the hybrid model.

5. Experimental Results

The diagnostic accuracy of the proposed hybrid model as well as the baseline RF-only model was evaluated on a given, unseen test dataset. This section presents a comparative evaluation, decomposing the outcome to reflect the architectural trade-offs and operational characteristics of the hybrid framework.

5.1. Overall Performance Comparison: An Initial Trade-Off

The initial comparison between the two models already reveals a significant performance trade-off that defines the core argument of this research (Figure 2). On initial observation, the baseline RF-only model appears superior, with a better overall accuracy (87.4%) and macro F1-score (0.803) compared to the hybrid model's 83.1% accuracy and 0.760 F1-score. While both models demonstrate extremely good resistance to false alarms on healthy data, the more aggregate scores in the simpler baseline model could be deceptive in showing unalloyed superiority.

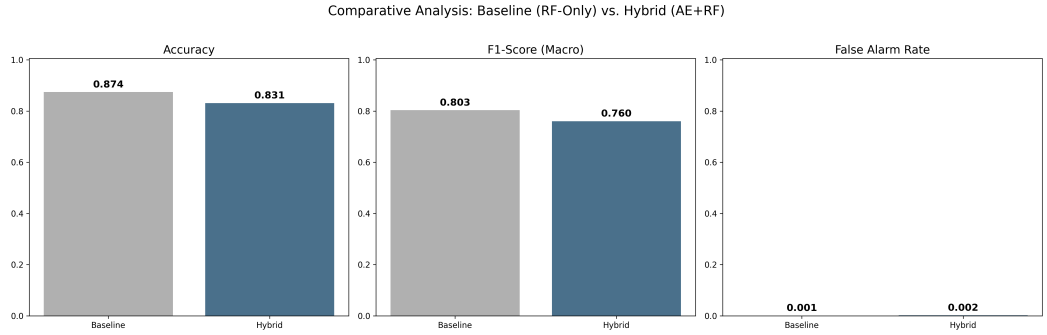


Figure 2. Overall performance between the baseline (RF-only, grey) and hybrid (AE + RF, blue) models.

However, a more experienced interpretation sees this not as a failure of the hybrid model but as the first indication of its more conservative and measured character. The RF-only model is forced to classify each and every instance, so its high precision is in fact "brittle"—it works fine on a pre-specified set of errors but has no clue how it would handle ambiguity or novel situations. The hybrid model's lower precision is the simple and expected consequence of its two-stage architecture, sacrificing diagnostic confidence for bold classification. These higher-level measures, therefore, do not tell the entire story and need closer investigation into the internal mechanics of the hybrid model to be able to utilize its true worth.

5.2. Analysis of the Anomaly Detection Stage: Identifying the Bottleneck

The behavior of the hybrid model is fundamentally governed by its first step: the LSTM autoencoder as an anomaly filter. Figure 3 shows the distribution of the reconstruction errors for normal (blue) and anomalous (red) sequences. The normal data distribution is a sharp, narrow peak at a very low error value, indicating that the model has learned a "fingerprint" of healthy operation and is reconstructing it with high fidelity. The anomalous data distribution, in contrast, is much broader and flatter, confirming that faults manifest with a wide variety of electrical signatures and, therefore, a wide variety of reconstruction errors. The significant overlap between the tail of the normal distribution and the head of the anomalous one delineates a "zone of confusion," a graphical representation of the inherent difficulty of discriminating between incipient or subtle faults and normal operating variation.

Ideally, under the best-case scenario, these two distributions would be disjoint with no overlap. This would mean that an easy threshold could perfectly separate all normal and anomalous instances with perfect anomaly detection.

This imperfection is mathematically quantified by the receiver operating characteristic (ROC) curve in Figure 4. An ideal ROC curve would proceed from the top-left corner of the plot toward a true positive rate of 1.0 at a false positive rate of 0.0, with an area under the curve (AUC) of 1.0. The resulting curve of the autoencoder itself, though, is a steep curve from this ideal corner that is needed to achieve an AUC of 0.61. This identifies the moderate sensitivity of the autoencoder as the core performance bottleneck of the overall hybrid system. The model's conservative anomaly threshold, selected to minimize false alarms under normal operation, means that faults whose signatures fall in the "zone of confusion" are deliberately and systematically omitted, which is the direct cause of the hybrid's lower overall accuracy.

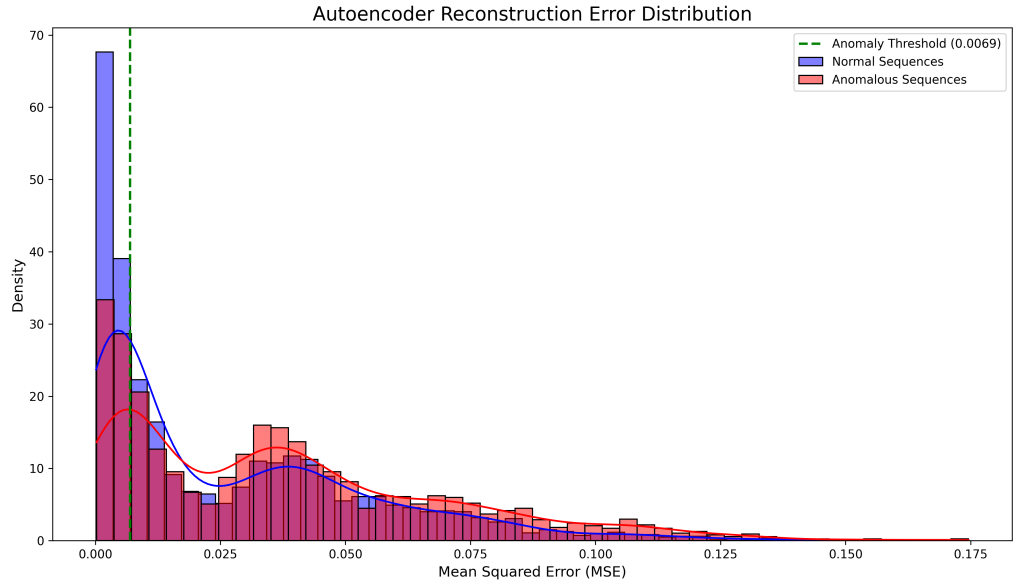


Figure 3. Reconstruction error distribution (MSE) for normal (blue) and anomalous (red) sequences. Overlapping regions appear purple. The green dashed line indicates the anomaly threshold (0.0069). The blue and red curves show the respective density estimates.

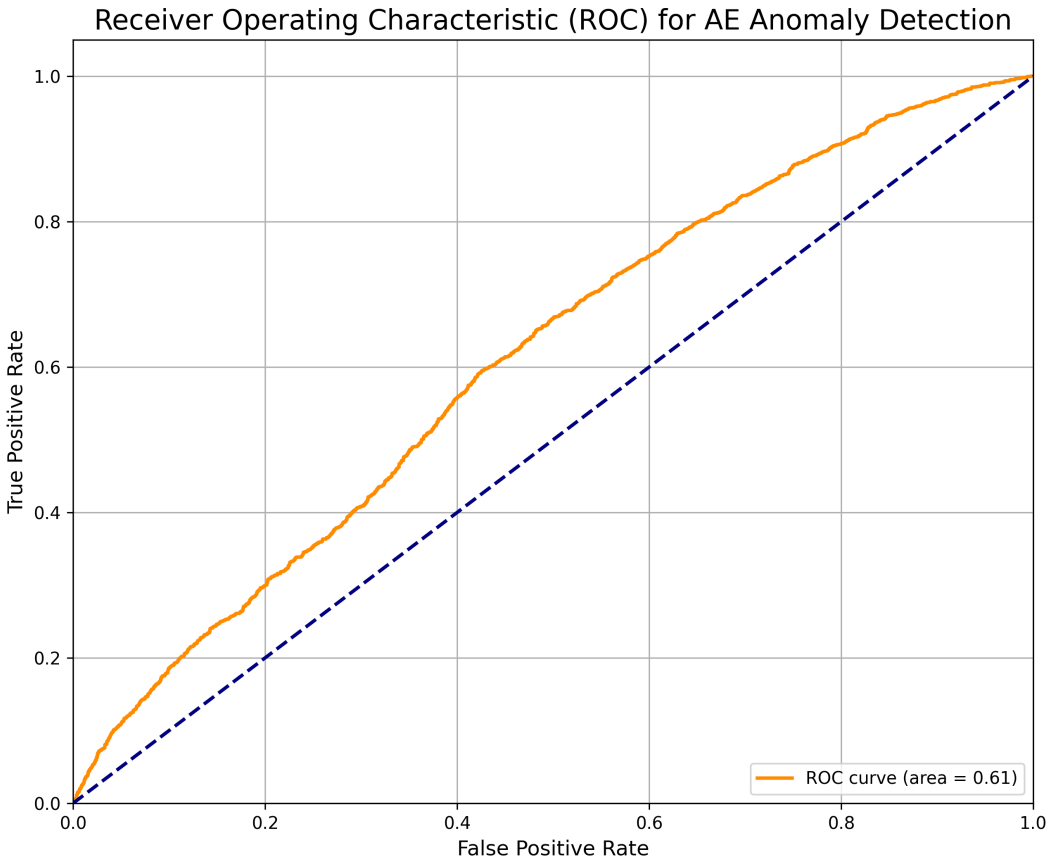


Figure 4. ROC curve for autoencoder anomaly detection in Stage 1, with an AUC of 0.61. The dashed diagonal line represents the performance of a random classifier (AUC = 0.5).

5.3. Per-Class Diagnostic Performance: A Strategic Sacrifice

The direct consequence of the imperfect filtering by the autoencoder is starkly evident in the per-class F1-score comparison (Figure 5). For the 'ShortCircuit' class—a critical and electrically well-defined fault—both models perform almost exactly and brilliantly, proving that the hybrid architecture does not compromise the detection of critical failures.

The difference in performance occurs in the more subtle classes. The baseline model appears to be better for 'Degraded', 'Normal', and 'Shaded' faults precisely because, as a single-stage classifier, it is forced to make a decision for each case. The hybrid model's poorer F1-scores for these classes are not the blame of its RF classifier; this is a direct and understandable result of the failure of the Stage 1 autoencoder to flag these events as anomalous in the first place. In an ideal hybrid system, the F1-scores would be equal to or greater than the baseline for every category, which would imply an ideal autoencoder. The outcome that is witnessed, however, highlights the essential engineering trade-off: the design consciously sacrifices sensitivity to low-priority, unclear faults in exchange for the high-confidence diagnosis and strength of its two-stage procedure.

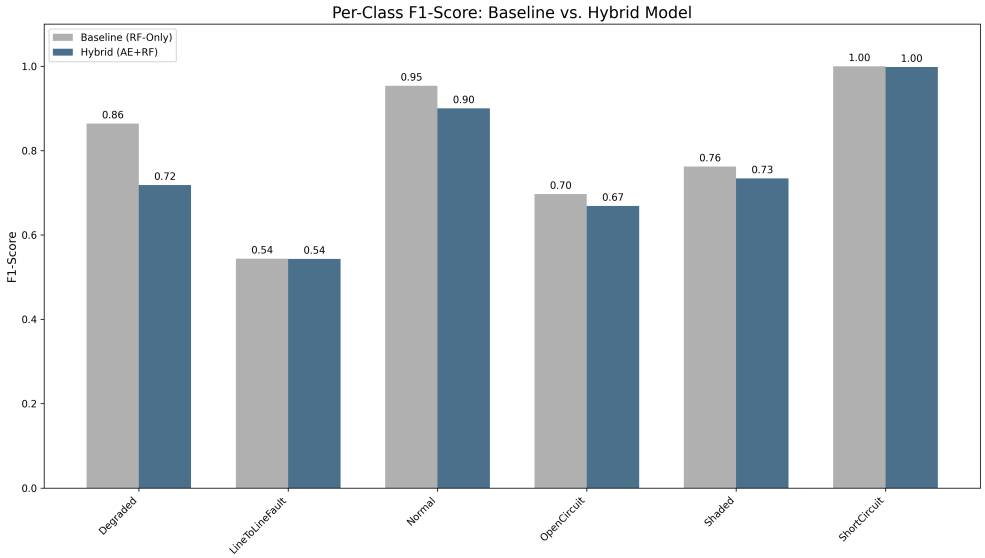


Figure 5. Per-class F1-score performance of each model on specific fault types.

5.4. Final Hybrid Model Diagnostic Accuracy: Interpretable Failure Modes

The confusion matrix in Figure 6 provides a final, qualitative summary of the end-to-end action of the hybrid system, not merely its accuracy, but also the reasoning behind its mistakes. An ideal confusion matrix would be fully diagonal with all off-diagonal elements equal to zero and all instances being well classified.

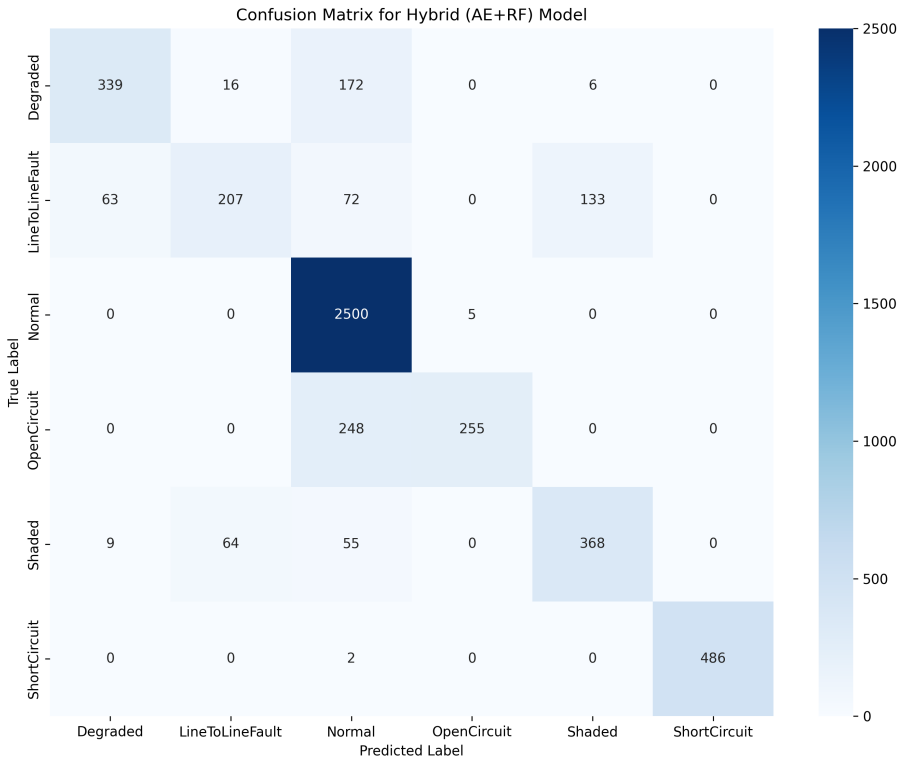


Figure 6. Confusion matrix for the end-to-end hybrid (AE+RF) model.

6. Discussion of Overall Findings

The 'near-perfect' diagonal elements in the 'Normal' (2500/2505 correct) and 'ShortCircuit' (486/488 correct) classes visually confirm the model's extremely high dependability on these dominant states. The model's failure modes are equally interpretable. The most serious errors comprise the classification of 172 'Degraded' and 248 'OpenCircuit' instances as 'Normal' classes, which are not random. These off-diagonal values in the 'Normal' column confirm that these are the extremely low-signature faults that are being filtered out by the Stage 1 autoencoder, as the analysis of reconstruction error had predicted.

Moreover, the bidirectional confusion between 'LineToLineFault' and 'Shaded' faults, seen as a little off-diagonal block, refers to the mentioned challenge in the FDD field, wherein various physical faults can cause nearly unrecognizable electrical signatures and hence are difficult to distinguish. This means that the eventual output errors are not arbitrary but are direct, traceable consequences of either the conservative Stage 1 filter or data structural ambiguities in themselves. Traceability is a valuable architectural resource because it makes model behavior predictable and its failures understandable. These off-diagonal entries in the 'Normal' column confirm that these are indeed the low-signature faults being removed by the Stage 1 autoencoder, as the analysis of reconstruction errors had predicted. This is due to the fact that electrical signatures of incipient degradation or partial open-circuit conditions can be very subtle, generating small I-V curve deviations that are difficult to distinguish from normal operational variability or sensor noise and therefore not crossing the conservative anomaly threshold.

The results of the experiments support the hybrid method strongly not only in terms of improved aggregate metrics but through its design and deployment implications. The primary contribution of this study is in the explicit trade-off between the marginally higher accuracy of the RF-only baseline and the conservative, robust design of the two-stage hybrid system. This reflects a shift in modern fault detection thinking, where safety,

interpretability, and resilience under unseen conditions are increasingly prioritized over raw performance gains.

The most perceptible feature of the proposed model is its anomaly-first modular design. Unlike monolithic classifiers that require lots of labeled data in order to generalize, this two-component system has a front-end LSTM autoencoder learned solely on unlabeled healthy training data. This design reliably addresses one of the most widely cited limitations in the literature, that of the lack of large, diverse, and well-labeled fault datasets for photovoltaic (PV) systems. By learning on a compressed signature of normal operation, the autoencoder signals major anomalies as likely faults, making it possible to operate stably even on new or inadequately monitored installations.

The moderate AUC of 0.61 of the 'autoencoder' is the program's key result, with a clear technical interpretation: 'a randomly selected 'anomaly' sample has only a 0.61 probability of being labeled a 'higher' reconstruction error than a randomly selected normal sample'. This quantifies the significant overlap in distributions (as seen in Figure 3) and recognizes the anomaly detection sensitivity as the primary bottleneck and most critical area for future work.

In addition, the modularity of the system simplifies maintenance and enables optimization of each stage independently. The analysis correctly pinpointed the primary performance bottleneck: the moderate sensitivity of the anomaly detection module. However, the design ensures that the faults that are detected are alarmed with high confidence. From a practical standpoint, especially for operations teams and asset managers, this kind of deterministic behavior might be more desirable than a model with moderately higher accuracy but vague and unpredictable failure modes.

6.1. Comparison with Existing Methods

To evaluate the effectiveness of the proposed hybrid method using autoencoders and random forest classifiers, we compare it with several commonly used techniques for fault detection and diagnosis in photovoltaic (PV) systems. Table 2 summarizes the key characteristics of each method based on the literature and practical implementation considerations.

Unsupervised models like pure autoencoders are excellent at learning normal system behavior and flagging new anomalies without having labeled fault data, but they tend to be poor at fine tuning sensitivity and handling noisy inputs [16,19]. Supervised learning algorithms such as SVMs and random forests offer better interpretability and robust classification when there are sufficient fault samples available, but their performance degrades in temporally correlated or high-dimensional settings [27,29]. Deep architectures like CNNs for spatial feature extraction, fully connected ANNs for the general pattern identification, and RNNs (like LSTM autoencoders) for sequential modeling of temporal structures achieve state-of-the-art detection performance at the cost of considerable computational burdens and reduced transparency [7]. Statistical approaches like PCA remain attractive due to their simplicity and speed but are limited by linear assumptions [7,9]. By contrasting these methods, we highlight how our two-step hybrid leverages the anomaly-screening feature of autoencoders and the decision explainability of random forests to achieve outstanding accuracy and real-time usability based on sparsely labeled data. This comparative framework not only emphasizes the advantages of our method's balanced sensitivity, strong resilience to unseen faults, and interpretability but also gives clear avenues for future optimizations, e.g., optimization of computational cost or inclusion of sequence-aware classifiers for more complex fault scenarios.

6.2. Practical Deployment Considerations

The successful transition of this FDD model as a theoretical approach into an effective field application depends on various engineering considerations. The availability of the data inputs' required voltage, current, irradiance, and temperature are standard parameters monitored by most of today's Supervisory Control and Data Acquisition (SCADA) systems employed in GCPV plants, making data availability feasible. The model can be realized as part of the software component of the plant monitoring system. Computationally, the trained LSTM-AE and random forest models are relatively lightweight in contrast to larger deep learning models, allowing for their deployment on edge devices for real-time inspection or a central cloud server for scheduled fleet-wide diagnostics without requiring massive computing resources.

Table 2. Detailed comparison of various methods for PV fault detection and diagnosis.

Method	Type	Typical Accuracy	Key Strengths	Main Weaknesses	Data Needs	Real-Time Suitability	Interpretability
Autoencoder	Unsupervised DL	Medium–High	• Learns normal behavior • Detects novel faults	• Needs lots of healthy data • Tuning sensitive	Medium (unlabeled)	High	Low
Random Forest	Supervised ML	Medium	• Robust to overfitting • Easy to tune	• Requires labeled faults • Struggles with high-dim raw signals	Low–Medium (labeled)	High	High
AE + RF (Proposed)	Hybrid (DL + ML)	High	• Unsupervised anomaly filtering • Transparent classification	• Two-stage complexity • Higher training cost	Medium	High	Moderate
SVM	Supervised ML	Medium	• Good for small datasets • Strong theoretical foundation	• Poor scalability • Kernel choice critical	Low (labeled)	Moderate	Moderate
CNN	Supervised DL	High	• Automatic feature learning • High accuracy	• Black-box • Heavy compute	High	Moderate	Low
ANN (FC)	Supervised DL	Medium–High	• Flexible architectures • Good baseline	• Prone to overfitting • Needs careful tuning	High	Moderate	Low
PCA	Unsupervised Stat	Low–Medium	• Simple anomaly detection • Fast	• Linear assumptions • Limited to global faults	Medium (unlabeled)	High	High
RNN (e.g., LSTM-AE)	DL Seq. Model	High	• Captures temporal dependencies • Good for time series	• Very heavy compute • Hard to interpret	High	Low–Moderate	Low

7. Conclusions and Future Work

This research developed and evaluated a hybrid AI GCPV fault detection and diagnosis framework, confirming that a two-stage LSTM-AE and random forest setup is an effective, interpretable diagnostic system. The primary contribution of this research is the critical consideration of this approach as a valid engineering trade-off, prioritizing diagnostic interpretability and safety over incremental enhancements of one accuracy metric. By design, the model reduces reliance on large labeled fault datasets and provides a good method of addressing novel anomalies.

Our analysis identified the limited sensitivity of the Stage 1 anomaly detector (AUC of 0.61) as the overall performance bottleneck. Future work must therefore concentrate on several key areas of enhancement and validation:

1. **Incorporating more fault signatures:** Future work should include training the model on more fault signatures like inverter faults, arc faults and line-to-ground faults).
2. **Enhancing Anomaly Detection:** The next logical step is to explore more advanced architectures for the Stage 1 detector to improve its sensitivity. This includes experimenting with alternative models such as **variational autoencoders (VAEs) or Transformer-based autoencoders**, which may better capture the complex distributions of PV system data.
3. **Broadening Benchmarks:** To better situate the model's performance, future work should include a broader comparative analysis against other hybrid FDD models (e.g., CNN-LSTM) and standalone unsupervised detectors like isolation forest.
4. **Methodological Rigor:** To ensure that the results are robust and not subject to the randomness of a single data partition, subsequent studies should employ a **K-fold cross-validation process.**
5. **Real-World Data Validation:** While the synthesized data used in this study facilitate controlled analysis, the most critical future work is to validate and refine the model using data from an operational PV plant. This is essential to test its performance against the noise, complexity, and variability of real-world conditions.

Author Contributions: Conceptualization, K.R., D.S. and A.R.; methodology, K.R. and D.S.; software, K.R.; validation, K.R., D.S., A.R. and A.A.; formal analysis, K.R., D.S., A.R. and A.A.; investigation, K.R.; writing—original draft preparation, K.R.; writing—review and editing, D.S., A.R., A.A. and N.E.M.; visualization, D.S., A.R. and A.A.; supervision, D.S. and A.R.; project administration, D.S. and A.R.; funding acquisition, A.R., A.A. and N.E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. International Renewable Energy Agency (IRENA). *Renewable Energy Capacity Statistics 2025*; Technical Report; International Renewable Energy Agency (IRENA): Bonn, Germany, 2025.
2. International Renewable Energy Agency (IRENA). Renewables in 2024: 5 Key Facts Behind a Record-Breaking Year. 2025. Available online: https://www.irena.org/News/articles/2025/Apr/Renewables-in-2024-5-Key-Facts-Behind-a-Record-Breaking-Year (accessed on 14 May 2025).
3. International Renewable Energy Agency (IRENA). Renewable Capacity Highlights 2025. 2025. Available online: https://static.poder360.com.br/2025/04/capacidade-energia-renovavel-mar2025-irena.pdf (accessed on 14 May 2025).
4. TaiyangNews. IRENA: A Record 452 GW Solar Capacity Installed in 2024. 2025. Available online: https://taiyangnews.info/business/irena-renewable-capacity-statistics-2025 (accessed on 14 May 2025).

5. International Renewable Energy Agency (IRENA). *Future of Solar Photovoltaic*; Technical Report; International Renewable Energy Agency (IRENA): Abu Dhabi, United Arab Emirates, 2019.
6. Davarifar, M.; GhaffarianHoseini, S.M.; GhaffarianHoseini, A.; Alamdari, N.M. Evaluating faults detection and their impact on photovoltaic modules. *Int. J. Renew. Energy Res.* **2013**, *3*, 834–842.
7. Mellit, A.; Tina, G.M.; Kalogirou, S.A. Fault detection and diagnosis methods for photovoltaic systems: A review. *Renew. Sustain. Energy Rev.* **2018**, *91*, 1–17. [CrossRef]
8. Madeti, S.R.; Singh, S.N. A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. *Sol. Energy* **2017**, *158*, 161–185. [CrossRef]
9. Alam, M.K.; Khan, F.; Johnson, J.; Flicker, J. A comprehensive review of catastrophic faults in PV arrays: Types, detection, and mitigation techniques. *IEEE J. Photovolt.* **2015**, *5*, 982–997. [CrossRef]
10. Yang, B.; Zheng, R.; Han, Y.; Huang, J.; Li, M.; Shu, H.; Su, S.; Guo, Z. Recent Advances in Fault Diagnosis Techniques for Photovoltaic Systems: A Critical Review. *Prot. Control Mod. Power Syst.* **2024**, *9*, 36–59. [CrossRef]
11. Venkatesh, S.N.; Sripada, D.; Sugumaran, V.; Aghaei, M. Detection of visual faults in photovoltaic modules using a stacking ensemble approach. *Heliyon* **2024**, *10*, e27894. [CrossRef] [PubMed]
12. Khan, F. *Live State of Health Monitoring of Inverter Subsystems*; Technical Report NREL/PR-5400-89801; National Renewable Energy Laboratory: Golden, CO, USA, 2024.
13. Pillai, D.S.; Blaabjerg, F.; Rajasekar, N. A Comparative Evaluation of Advanced Fault Detection Approaches for PV Systems. *IEEE J. Photovolt.* **2019**, *9*, 513–527. [CrossRef]
14. Chen, Z.; Han, F.; Wu, L.; Yu, J.; Cheng, S.; Lin, P.; Chen, H. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. *Energy Convers. Manag.* **2018**, *178*, 250–264. [CrossRef]
15. Bougoffa, M.; Benmoussa, S.; Djeziri, M.; Palais, O. Hybrid Deep Learning for Fault Diagnosis in Photovoltaic Systems. *Machines* **2025**, *13*, 378. [CrossRef]
16. Qian, J.; Song, Z.; Yao, Y.; Zhu, Z.; Zhang, X. A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes. *Chemom. Intell. Lab. Syst.* **2022**, *231*, 104711. [CrossRef]
17. Malhotra, P.; Vig, L.; Shroff, G.; Agarwal, P. Long short term memory networks for anomaly detection in time series. In Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 27–29 April 2016 .
18. Amiri, A.F.; Oudira, H.; Chouder, A.; Kichou, S. Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier. *Energy Convers. Manag.* **2024**, *300*, 118076. [CrossRef]
19. Malhotra, P.; Vig, L.; Shroff, G.; Agarwal, P. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv* **2016**. [CrossRef]
20. Wei, Y.; Jang-Jaccard, J.; Xu, W.; Sabrina, F.; Camtepe, S.; Boulic, M. LSTM-Autoencoder-Based Anomaly Detection for Indoor Air Quality Time-Series Data. *IEEE Sens. J.* **2023**, *23*, 3787–3800. [CrossRef]
21. Raihan, A.S.; Ahmed, I. A Bi-LSTM Autoencoder Framework for Anomaly Detection—A Case Study of a Wind Power Dataset. In Proceedings of the 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), Auckland, New Zealand, 26–30 August 2023; pp. 1–6. [CrossRef]
22. Lachekhab, F.; Benzaoui, M.; Tadjer, S.A.; Bensmaine, A.; Hamma, H. LSTM-Autoencoder Deep Learning Model for Anomaly Detection in Electric Motor. *Energies* **2024**, *17*, 2340. [CrossRef]
23. Lindemann, B.; Maschler, B.; Sahlab, N.; Weyrich, M. A survey on anomaly detection for technical systems using LSTM networks. *Comput. Ind.* **2021**, *131*, 103498. [CrossRef]
24. Baradieh, K.; Zainuri, M.; Kamari, M.; Yusof, Y.; Abdullah, H.; Zaman, M.; Zulkifley, M. RSO based Optimization of Random Forest Classifier for Fault Detection and Classification in Photovoltaic Arrays. *Int. Arab. J. Inf. Technol.* **2024**, *21*, 636–660 . [CrossRef]
25. Gaaloul, Y.; Kechiche, O.B.H.B.; Oudira, H.; Chouder, A.; Hamouda, M.; Silvestre, S.; Kichou, S. Faults Detection and Diagnosis of a Large-Scale PV System by Analyzing Power Losses and Electric Indicators Computed Using Random Forest and KNN-Based Prediction Models. *Energies* **2025**, *18*, 2482. [CrossRef]
26. Do, T.N.; Lenca, P.; Lallich, S.; Pham, N.K. Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees. In *Advances in Knowledge Discovery and Management*; Guillet, F., Ritschard, G., Zighed, D.A., Briand, H., Eds.; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2010; Volume 292, pp. 33–50. [CrossRef]
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

28. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009. [CrossRef]
29. Harrou, F.; Sun, Y.; Taghezouit, B.; Saidi, A.; Hamlati, M.E. Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches. *Renew. Energy* **2018**, *116*, 22–37. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.