Benchmarking Time-Series Anomaly Detection Algorithms for Photovoltaic Plants

Hannes Bradl[1], Katharina Hofer-Schmitz[1], Pasquale Grippa[2] and Gernot Hofer[3]

[1] *JOANNEUM RESEARCH, Graz, Styria, 8010, Austria*
hannes.bradl@joanneum.at
katharina.hofer-schmitz@joanneum.at

[2] *Lakeside Labs, Klagenfurt, Carinthia, 9020, Austria*
grippa@lakeside-labs.com

[3] *CAMPUS 02 Fachhochschule der Wirtschaft, Graz, Styria, 8010, Austria*
gernot.hofer@campus02.at

ABSTRACT

Photovoltaic (PV) systems are increasingly important for renewable energy generation, but faults can reduce yield if they remain undetected. This paper presents a benchmark of multivariate time-series anomaly detection (TSAD) methods for PV monitoring using two complementary datasets: a newly collected real-world twin-plant PV dataset (FHC), in which controlled faults were physically introduced during operation and additional anomalies were injected after data collection, and the publicly available PV fault dataset (UTFPR). Using the TimeEval framework, we evaluate a broad range of unsupervised and semi-supervised TSAD algorithms on the FHC dataset, and only unsupervised algorithms on the UTFPR dataset, as the latter does not contain fault-free days required for semi-supervised training. On the FHC dataset, the highest performance is achieved by semi-supervised reconstruction-based methods. On the UTFPR dataset, fully unsupervised distance- and density-based methods perform best. The ablation study further shows that environmental features do not necessarily improve detection performance and may introduce confounding operating-regime structure for some algorithms. These findings highlight the importance of dataset characteristics and feature selection when applying TSAD methods to PV monitoring.

1. INTRODUCTION

In 2025, the electricity generated by photovoltaic (PV) systems reached 13.2% (369 TWh) of total electricity in Europe, representing a 20.1% increase compared to the previous year (Ember, 2026). Given this substantial and rapidly growing share, optimizing the operation of PV systems has become increasingly important. However, PV systems may operate under faulty or sub-optimal conditions that remain undetected for extended periods, thereby reducing the effective energy yield. Causes of these operating conditions include electrical malfunctions and partial panel coverage, for example due to leaves or debris.

Operational time-series data, such as string-level voltage and current measurements, are available in many residential and commercial PV installations and provide a valuable basis for automated fault detection. Continuous monitoring of these data streams enables early identification of anomalous operating conditions, allowing maintenance actions to restore optimal performance and recover otherwise lost energy. Accurate and timely anomaly detection can therefore improve annual energy yield, reduce CO_2 emissions, and shorten the economic payback period of PV systems.

In this paper, we introduce a real-world PV dataset and benchmark various time-series anomaly detection (TSAD) algorithms that can be executed on the historical time-series whenever new data becomes available, thereby enabling prompt anomaly detection, as motivated above. The evaluation is conducted on the electrical quantities typically available in standard inverters, and additional meteorological measurements such as solar irradiance, ambient temperature, wind speed, and wind direction. This allows us to assess performance under both commercial deployment conditions, where only inverter data are available, and experimental installations equipped with additional sensors.

The main contributions of this work are as follows. First,

we introduce a new real-world PV anomaly detection dataset (FHC) collected from a twin-plant setup, containing both controlled physical faults and artificially injected sensor or data-quality anomalies. Second, we perform a comprehensive benchmark of a wide range of unsupervised and semi-supervised TSAD methods within a unified and reproducible evaluation framework. Finally, we analyze the impact of environmental features by evaluating models with and without weather-related variables, providing insights into how feature selection influences anomaly detection performance in PV monitoring systems.

The remainder of this paper is structured as follows. Section 2 reviews related work on anomaly detection in PV systems and TSAD methods. Section 3 describes the datasets used in this study. Section 4 outlines the benchmark methodology and evaluation setup. Section 5 presents the experimental results. Section 6 discusses the findings and their implications for PV monitoring. Finally, Section 7 concludes the paper and outlines directions for future work.

2. RELATED WORK

Numerous studies have explored machine-learning-based fault detection and diagnosis in PV systems. Much of this literature frames PV fault diagnosis as a supervised classification problem, using labeled examples of fault conditions such as open-circuit faults, short-circuit faults, mismatch, degradation, and shading. However, semi-supervised and unsupervised anomaly-detection methods are increasingly relevant for real-world PV monitoring, where fault labels are sparse, incomplete, or unavailable.

Several reviews address fault detection in PV systems. B. Li, Delpha, Diallo, and Migan-Dubois (2021) survey artificial neural networks for PV fault detection, covering architectures, input features, fault types, and data sources. Gaviria, Narváez, Guillen, Giraldo, and Bressan (2022) review deep learning methods across PV tasks like fault diagnosis, forecasting, and energy management, noting the dominance of supervised methods. Berghout et al. (2021) categorize PV failures (e.g., shading, open-circuit, degradation) and provide an overview of conventional machine learning, deep learning, and knowledge-guided methods. Similarly, Abuhashish, Refaat, Kalas, Hamad, and Elfar (2025) provide an overview of common PV array faults and review recent machine learning and deep learning approaches for their diagnosis. They note that supervised learning methods remain the predominant approach for fault diagnosis in PV systems. Alhazmi, Kholoud, and Eke (2025) review deep learning approaches and datasets for photovoltaic applications, highlighting inadequate data quality and limited data availability as major challenges in the field. Y. Liu et al. (2025) observe that most existing research is limited to simulated or laboratory datasets, which fail to capture the noise and complexity inherent in real-world time-series data. They identify a critical need for publicly accessible benchmarks and data-efficient techniques that incorporate meteorological and operational parameters. In summary it can be stated that supervised learning methods dominate the field for fault detection. Further challenges lie in the lack of publicly available data and standardized benchmarks to compare different methods.

Supervised PV fault diagnosis has been studied extensively for fault types such as open-circuit faults, short-circuit faults, ground faults, mismatch, and partial shading, using methods ranging from classical machine learning and ensemble classifiers to CNN-, GRU-, and time–frequency-based deep learning models (Amiri, Kichou, Oudira, Chouder, & Silvestre, 2024; Khandeparkar, Shreshtha, & Ramu, 2025; Seghiour et al., 2026; Toche Tchio, Kenfack, Kassegne, Menga, & Ouro-Djobo, 2024). However, these approaches require representative labeled examples of the relevant fault classes, which limits their applicability in real-world monitoring scenarios where labels are sparse, incomplete, or unavailable.

Because diverse labeled datasets are costly and time-consuming to produce, TSAD methods provide an attractive alternative. In particular, semi-supervised approaches can learn normal operating behavior from fault-free historical data, while unsupervised methods can be applied directly to unlabeled operational data. Schmidl, Wenig, and Papenbrock (2022) present one of the most comprehensive benchmarks of TSAD algorithms, evaluating 71 methods across 976 datasets from diverse application domains. Their study highlights the broad methodological diversity of TSAD techniques. Complementary to this work, Zamanzadeh Darban, Webb, Pan, Aggarwal, and Salehi (2024) provide a recent survey of deep learning methods for TSAD, summarizing architectures and detection principles relevant for multivariate sensor data.

Within the PV domain, several studies explore the application of anomaly detection techniques to identify abnormal system behavior. Branco, Gonçalves, and Costa (2020) developed tailored anomaly detection algorithms to identify production anomalies in PV plants by analyzing deviations in electrical measurements. Miraftabzadeh, Longo, Leva, and Matera (2025) propose a deep learning approach based on autoencoders for detecting anomalies in PV power time series, demonstrating its effectiveness on operational data from a PV installation in Genoa, Italy.

In summary, while much of the PV fault-detection literature focuses on supervised fault classification, these methods require representative labeled fault examples and therefore suffer from costly labeling and limited applicability in real-world monitoring scenarios. Recent reviews and benchmarks highlight the shift toward semi-supervised and unsupervised anomaly detection methods for handling unlabeled or partially labeled time-series data. However, existing PV fault-detection studies often focus either on supervised fault classification or

on individual anomaly-detection methods. What is still missing is a reproducible benchmark of heterogeneous TSAD algorithms on realistic PV time series, especially one that compares semi-supervised settings with fault-free reference data against fully unsupervised settings on contaminated operational data.

3. DATA

We used two datasets: a novel dataset introduced in this paper (Section 3.1), and a publicly available dataset, detailed in Section 3.2.

3.1. FHC Twin-Plant PV Dataset

The primary dataset used in this study is the FHC twin-plant PV anomaly dataset (Bradl, Hofer-Schmitz, Grippa, & Hofer, 2026).[1] The dataset originates from a multi-month measurement campaign involving two physically identical PV plants installed at the FH Campus 02 (FHC) in Graz, Austria, and operating under identical environmental conditions. An overview of the plants is shown in Figure 1.

Each plant is equipped with one inverter (Hoymiles HM-1500) and four strings. Each string consists of a single PV module (Risen Energy Titan S RSM40-8-400MB). Electrical measurements are recorded at a temporal resolution of 30 s and include string-level DC voltage and current for all four strings. In addition, environmental variables are measured, namely solar irradiance, ambient temperature, wind speed, and wind direction.

One of the two PV plants is operated under normal conditions and is used to construct a fault-free reference dataset for training semi-supervised TSAD algorithms. The second plant is deliberately modified to simulate realistic PV faults. For the present study, we selected 25 days between 17th June 2025 and 16th July 2025, focusing on two fault scenarios: partial shading and induced mismatch. Partial shading was created using sheets of paper (sizes DIN A5 and A4) to simulate the presence of objects such as leaves and debris on the panel. The mismatch was achieved by altering the tilt angle of selected modules, which reduces incident solar irradiance on the modules.

In addition to these physical fault scenarios, artificial anomalies were injected into the data to simulate common data-quality and sensor issues. Artificial anomalies were injected only into the electrical measurements of the modified plant, i.e., the four string-level voltage and current channels. Injection was performed at the segment level using a fixed random seed to ensure reproducibility. The target artificial-anomaly ratio was set to 2% of the dataset length. Segment lengths were sampled uniformly between 10 and 180 samples, corresponding to 5–90 minutes at the 30 s sampling interval, while spike anomalies were injected as single-sample events. For each injected segment, one to three electrical channels were selected randomly. The following anomaly types were considered: signal dropouts, implemented by setting the selected values to zero; scaling anomalies, implemented by multiplying the selected values by a factor sampled uniformly from 0.5 to 0.8; spikes, implemented by adding or subtracting six times the empirical standard deviation of the affected channel; stuck-sensor anomalies, implemented by repeating the previous valid value; and additive noise, implemented by adding Gaussian noise with a standard deviation equal to 0.5 times the empirical standard deviation of the affected channel. Artificial anomalies were injected only into samples not already labeled as anomalous, so that they do not overwrite the manually annotated physical fault intervals.

Figure 1. Photograph of the FHC twin-plant PV installation used for data collection.

Table 1 summarizes the segment-level class distribution of the FHC dataset. After anomaly injection, the dataset contains 31 labeled anomaly intervals, comprising both single-point anomalies and long contiguous segments. The minimum anomaly length is 1 sample (30 s), while the longest anomaly spans 813 samples (406.5 min). The resulting dataset contamination is 5.33 %. This mixture of short-duration and extended anomalies reflects realistic PV monitoring conditions and poses a challenging benchmark for TSAD methods.

3.2. UTFPR PV Fault Dataset

To assess the generalizability of the evaluated methods, we additionally consider a publicly available PV fault dataset, referred to in this paper as the UTFPR PV fault dataset (UTFPR), named after the institutional affiliation *Universidade Tecnológica Federal do Paraná* of the dataset's authors. The dataset is openly available in the authors' repository and is described in detail in the corresponding publication (Lazzaretti et al., 2020).[2]

[1] The dataset is publicly available on Zenodo: https://doi.org/10.5281/zenodo.18979877

[2] UTFPR PV fault dataset repository: https://github.com/clayton-h-costa/pv_fault_dataset

Table 1. Segment-level class statistics of the FHC twin-plant dataset (modified plant).

Class	# Segments	Total dur.	Min. length	Median length	Max. length	Dataset share
Normal operation	33	34 079.5 min	12.0 min	624.5 min	3726.5 min	94.665 %
Shadowing (A4 sheet)	1	282.5 min	282.5 min	282.5 min	282.5 min	0.785 %
Shadowing (A5 sheet)	2	516.5 min	110.0 min	258.2 min	406.5 min	1.435 %
Tilt increase	1	329.5 min	329.5 min	329.5 min	329.5 min	0.915 %
Current dips	4	9.0 min	0.5 min	1.5 min	5.5 min	0.025 %
Spikes	8	4.0 min	0.5 min	0.5 min	0.5 min	0.011 %
Dropouts	4	176.0 min	33.5 min	38.8 min	65.0 min	0.489 %
Scaling	5	283.5 min	23.5 min	72.0 min	76.5 min	0.788 %
Noise	5	265.5 min	36.5 min	56.5 min	68.0 min	0.738 %
Stuck sensor	1	54.0 min	54.0 min	54.0 min	54.0 min	0.150 %

The UTFPR dataset contains 16 days of electrical and environmental measurements from a PV system operating under normal and faulty conditions, with labeled fault events. The PV plant used to record this dataset consists of two strings with 8 modules (CS6U-330P) each. Both strings are connected to a 5 kW grid-tie power inverter (NHS Solar 5K-GDM1).

The recorded variables include DC voltage and current measurements at string level, solar irradiance, and module temperature. Fault conditions are intentionally introduced and include partial shading, short-circuit faults, open-circuit faults, and degradation-related effects. The original sampling interval of the dataset is 1 s; for consistency with the FHC dataset and to reduce computational cost, the data are downsampled to a sampling interval of 30 s in this study.

Since the UTFPR dataset differs in system configuration, recorded variables, and fault characteristics compared to the FHC dataset, it provides a complementary benchmark for evaluating TSAD algorithms.

Table 2 summarizes the segment-level class distribution of the UTFPR PV fault dataset, including the number of segments, duration statistics, and the relative share of each class. The dataset is dominated by normal operation and shadowing events, while short-circuit, degradation, and open-circuit faults occur less frequently and form comparatively short intervals.

Table 3 compares the anomaly characteristics of the FHC and UTFPR datasets. Compared to FHC, UTFPR exhibits a higher anomaly contamination and a larger number of shorter anomaly intervals.

4. EXPERIMENTAL SETUP

4.1. Benchmark Framework

All experiments are conducted using the *TimeEval* (Wenig, Schmidl, & Papenbrock, 2022) framework,[3] a modular benchmarking environment that enables a fair and reproducible comparison of TSAD algorithms. It provides a unified execution interface for a broad range of TSAD methods. Using a single benchmarking framework across all experiments ensures that performance differences observed in this work are attributable to the characteristics of the datasets and algorithms rather than to differences in experimental setup.

Algorithms are evaluated under standardized experimental conditions, including consistent data handling, preprocessing, training, execution, and evaluation. Hyperparameters for all algorithms were optimized using grid search over predefined parameter ranges. Each configuration was evaluated independently, and the best-performing configuration for each method was selected based on the area under the precision-recall curve (PR-AUC).

4.2. Training and Test Splits

For the FHC dataset, we exploit the twin-plant setup to define a semi-supervised evaluation scenario. The unmodified plant is assumed to represent fault-free operation and is therefore used as the training dataset for semi-supervised algorithms. The modified plant contains the deliberately introduced physical faults as well as the synthetically injected anomalies and is used as the test dataset for all evaluated algorithms. Thus, semi-supervised methods learn normal operating behavior exclusively from the unmodified reference plant and are evaluated on the modified plant.

Unsupervised algorithms do not require a separate training dataset. For these methods, the modified FHC plant is used both for model fitting and for testing. However, the labels of the modified plant are used only for evaluation and hyperparameter selection, not during model fitting. This setup reflects the practical case in which anomaly detection methods are applied directly to operational data that may already contain faults.

For the UTFPR dataset, no fault-free training subset is available, since each day contains at least one labeled fault event. Consequently, semi-supervised algorithms cannot be trained

[3] TimeEval GitHub repository: https://github.com/TimeEval/TimeEval

Table 2. Segment-level class statistics of the UTFPR PV fault dataset.

Class	# Segments	Total duration	Min. length	Median length	Max. length	Dataset share
Normal operation	157	19 242.0 min	0.5 min	23.5 min	945.0 min	84.037 %
Short-circuit	12	96.0 min	0.5 min	9.25 min	10.5 min	0.419 %
Degradation	12	170.5 min	8.0 min	10.0 min	45.0 min	0.745 %
Open circuit	10	99.5 min	9.5 min	10.0 min	10.5 min	0.435 %
Shadowing	167	3289.0 min	0.5 min	1.5 min	113.5 min	14.364 %

Table 3. Comparison of the FHC twin-plant PV dataset and the UTFPR PV fault dataset.

Statistic	FHC	UTFPR
Total length	600.0 h	381.6 h
Contamination	5.33 %	15.96 %
Number of anomaly segments	31	156
Minimum anomaly length	30 s	30 s
Median anomaly length	36.5 min	10.5 min
Maximum anomaly length	406.5 min	113.5 min

in a comparable way on this dataset. We therefore evaluate only fully unsupervised algorithms on UTFPR, using the entire dataset for fitting and testing. As for the FHC dataset, labels are used only for evaluation and hyperparameter selection.

4.3. Algorithms

We evaluate a diverse set of unsupervised and semi-supervised multivariate TSAD methods. In this context, semi-supervised refers to algorithms trained exclusively on fault-free data to characterize normal system behavior; anomalies are subsequently detected as deviations from this learned baseline. In contrast, unsupervised methods are fitted directly on the contaminated test data without a separate training phase.

All algorithms and their reference implementations were obtained from the TimeEval framework. Detailed methodological descriptions can be found in (Schmidl et al., 2022). Each method outputs a continuous anomaly score between 0 and 1 for each timestamp, where higher scores indicate more anomalous behavior. For subsequence-based methods, TimeEval first computes anomaly scores for sliding windows and then maps these window-level scores back to individual timestamps.

In addition to the following algorithms, a random baseline is included as a reference, assigning an independent random anomaly score drawn from a uniform distribution to each timestamp, without using any information from the data. This baseline provides a lower bound for detection performance and helps contextualize metric values.

Reconstruction- and prediction-based deep models. These methods learn a representation of normal behavior—usually in a semi-supervised way—and use reconstruction or prediction errors as anomaly scores.

- **Autoencoder (AE)** and **denoising autoencoder (DAE)** (Sakurada & Yairi, 2014): feed-forward autoencoders trained to reconstruct the input. DAE adds input corruption during training for robustness.
- **EncDec-AD** (Malhotra et al., 2016): sequence-to-sequence (encoder–decoder) recurrent model; anomalies are indicated by reconstruction deviations over windows.
- **DeepAnT** (Munir, Siddiqui, Dengel, & Ahmed, 2018): convolutional predictor trained on normal data; anomalies correspond to high prediction errors.
- **LSTM-AD** (Malhotra, Vig, Shroff, Agarwal, et al., 2015): forecasting/reconstruction approach based on long short-term memory (LSTM); prediction errors define anomaly scores.
- **OmniAnomaly** (Su et al., 2019): stochastic recurrent model that detects anomalies based on reconstruction errors.
- **MTAD-GAT** (Zhao et al., 2020): leverages parallel graph attention networks (GAT) to explicitly map dependencies between different sensors and across time, identifying anomalies through a combination of forecasting and reconstruction tasks.
- **TAnoGAN** (Bashar & Nayak, 2020): based on generative adversarial networks (GAN), it learns normal behavior using adversarial training.
- **Telemanom** (Hundman, Constantinou, Laporte, Colwell, & Soderstrom, 2018): uses LSTM networks to identify anomalies by applying a nonparametric, dynamic thresholding technique to the resulting prediction errors.
- **Torsk** (Heim & Avery, 2019): spatially aware echo state network (ESN); anomalies are derived from prediction residuals.
- **HealthESN** (Chen, Zhang, Huang, He, & Song, 2020): optimizes an ESN using a modified objective function that penalizes the misclassification of minority-class anomalies more heavily than normal data points.

Distance- and density-based detectors. These methods identify observations that lie far from typical data regions. With the exception of Hybrid KNN, all of those algorithms are unsupervised.

- **K-nearest neighbors (KNN)** (Ramaswamy, Rastogi, & Shim, 2000): uses distance to the k nearest neighbors as outlier score.
- **Subsequence KNN (Sub-KNN)**: applies KNN on sliding windows to detect anomalous subsequences.
- **Local outlier factor (LOF)** (Breunig, Kriegel, Ng, & Sander, 2000): is based on local density deviations.
- **Connectivity-based outlier factor (COF)** (Tang, Chen, Fu, & Cheung, 2002): replaces density-based metrics of LOF with a connectivity-based approach to better isolate anomalies located in non-spherical or linear structural patterns.
- **Cluster-based local outlier factor (CBLOF)** (He, Xu, & Deng, 2003): anomaly score is measured by both the size of the cluster the object belongs to and the distance between the object and its closest "large" cluster.
- **Hybrid KNN** (Song, Jiang, Men, & Yang, 2017): solves the curse of dimensionality by using a deep autoencoder to map high-dimensional data into a compact subspace before applying an ensemble of KNN graphs to calculate anomaly scores.
- **k-Means** (Yairi, Kato, & Hori, 2001): distance to nearest centroid is used as anomaly score.

Isolation- and ensemble-based detectors. These methods isolate anomalous points using randomized partitions or ensembles.

- **Isolation Forest (iForest)** (F. T. Liu, Ting, & Zhou, 2008): detects anomalies by randomly partitioning the data space until a point is isolated, assigning higher outlier scores to points that require the fewest number of splits to be separated.
- **Extended Isolation Forest (EIF)** (Hariri, Kind, & Brunner, 2021): extension of iForest using random hyperplanes for improved performance in higher dimensions.
- **IF-LOF** (Cheng, Zou, & Dong, 2019): two-stage approach that combines isolation-based candidate selection with density-based refinement.
- **Random Black Forest (RBF)** (Ziegelmeir, 2019): predicts the next multivariate point using an ensemble of random-forest regressors and flags large prediction errors as anomalies.

Probabilistic and pattern-based methods. These methods rely on robust covariance estimation, projection, distributional assumptions, or symbolic/motif representations.

- **Histogram-based outlier score (HBOS)** (Goldstein & Dengel, 2012): fast and simple algorithm assuming feature-wise independence.
- **Fast-MCD** (Rousseeuw & Driessen, 1999): allows for the calculation of robust Mahalanobis distances, making it easier to flag outliers that would otherwise hide in a standard analysis.
- **Robust principal component analysis (RobustPCA)** (Paffenroth, Kay, & Servi, 2018): decomposes signals into low-rank (normal) and sparse (anomalous) components.
- **Copula-based outlier detection (COPOD)** (Z. Li, Zhao, Botta, Ionescu, & Hu, 2020): copula-based probabilistic outlier detection using empirical tail probabilities.
- **Principal component classifier (PCC)** (Shyu, Chen, Sarinnapakorn, & Chang, 2003): correlation-based detector that flags deviations from typical correlation structure.
- **Discord aware matrix profile (DAMP)** (Lu, Wu, Mueen, Zuluaga, & Keogh, 2022): focuses only on potential discords and skips the redundant calculations of normal patterns.
- **mSTAMP** (Yeh et al., 2016): matrix-profile-based subsequence anomaly detection (discord discovery) on sliding windows.
- **GrammarViz** (Senin et al., 2015): grammar-based symbolic representation; anomalies correspond to rare/discord subsequences in the symbolic domain.
- **Subsequence LOF (Sub-LOF)** (Breunig et al., 2000): LOF applied to sliding windows to capture interval anomalies.

Overall, this algorithm set covers complementary detection principles, enabling a systematic comparison under identical preprocessing and evaluation settings as recommended by Schmidl et al. (2022).

4.4. Evaluation Metrics

To rigorously assess the performance of the various TSAD algorithms, we employ two primary, well-established metrics: the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC). While ROC-AUC is widely used and facilitates comparison with prior benchmarking studies, it can present an overly optimistic view in highly imbalanced settings, as it does not fully reflect the impact of false positives when the number of normal samples strongly exceeds the number of anomalies. In contrast, PR-AUC focuses on precision and recall and is therefore more sensitive to class imbalance. Since anomaly detection is inherently characterized by low anomaly contamination rates, PR-AUC is generally considered more informative and better suited for evaluating detection performance in this context. For completeness and comparability, we report both metrics throughout this study. In the result tables, the best hyperparameter configuration per algorithm is selected based on PR-AUC, reflecting its higher relevance for anomaly detection scenarios.

4.5. Experimental Procedure

First, both datasets were standardized using a z-score normalization, applied independently to each input variable (i.e., feature). For the FHC dataset, scaling parameters were computed using the fault-free training data only and subsequently applied to the corresponding test set. For the UTFPR dataset, scaling parameters were computed on the entire dataset due to the absence of a fault-free training subset.

Second, depending on the algorithm category, models are either trained on fault-free data (semi-supervised setting – FHC only) or fitted directly on the test data (unsupervised setting). Note that all algorithms operate directly on the standardized raw sensor measurements—hereafter referred to as features. No additional feature engineering was applied.

Third, each algorithm is executed with multiple predefined hyperparameter configurations. We evaluate all configurations independently and report the best-performing run per algorithm based on PR-AUC. Note that we use the test set for hyperparameter selection, since no separate labeled validation set is available. This protocol can lead to optimistic metric values because the same labeled anomalies are used both to select the best hyperparameter configuration and to report final performance. However, since this procedure is applied consistently across all evaluated algorithms, the results remain comparable and the benchmark focuses on relative performance rather than absolute metric values which can be expected in real-world deployments. Explored hyperparameters vary by method; examples include window sizes, number of nearest neighbors, latent dimensionality, and network architectures.

Finally, continuous anomaly scores are generated for each timestamp. For window-based methods, window-level scores are mapped back to time points using the aggregation strategy implemented in TimeEval. Performance is then evaluated using ROC-AUC and PR-AUC, and runtime (training plus inference) is recorded to assess computational efficiency.

4.6. Ablation Study

To analyze the impact of feature selection on anomaly detection performance, we conduct a systematic ablation study on both datasets. Starting from the full multivariate feature set (electrical and environmental features), we progressively remove selected environmental measurements and re-run the complete benchmarking pipeline.

For the FHC dataset, three configurations are considered:

- all available features,
- without wind speed, wind direction and ambient temperature,
- without wind speed, wind direction, ambient temperature, and irradiance.

For the UTFPR dataset, we evaluate:

- all available features,
- without module temperature,
- without module temperature and irradiance.

Note that the UTFPR dataset contains *module* temperature, which is physically influenced by ambient temperature as well as wind speed and wind direction through convective cooling effects. Although ambient temperature and wind measurements are not explicitly available in the UTFPR dataset, module temperature implicitly captures part of this environmental information. Consequently, removing module temperature from the feature set corresponds to eliminating a proxy feature that encodes weather-dependent operating conditions.

5. RESULTS

5.1. Results on FHC Dataset

Table 4 reports the best-performing hyperparameter configuration of each algorithm selected by PR-AUC. Overall, reconstruction-based deep models achieve the strongest detection performance on this dataset. The AE yields the highest PR-AUC when all features are used (PR-AUC 0.868; ROC-AUC 0.966), closely followed by the DAE (PR-AUC 0.839; ROC-AUC 0.956). When excluding wind and temperature measurements, AE and DAE remain the two top-performing methods (PR-AUC 0.860 and 0.853, respectively). In the most reduced feature set (excluding wind, temperature, and irradiance), DAE achieves the best PR-AUC of 0.863 (ROC-AUC 0.965), while AE remains competitive (PR-AUC 0.839).

Among non-deep methods, Fast-MCD performs consistently well across configurations and notably improves when irradiance is removed (PR-AUC increases from 0.777 with all features to 0.829 without wind, temperature, and irradiance). HealthESN shows a strong positive response to feature ablation, improving from PR-AUC 0.564 to 0.737 in the reduced feature set. Distance- and density-based methods (e.g., KNN, LOF, CBLOF) show the largest gains from removing environmental features, reaching PR-AUC values in the range 0.55–0.67 in the reduced configuration, whereas they perform substantially worse when all environmental measurements are included.

Figure 2 (left) compares runtime and performance of the best-performing run for each algorithm. AE and DAE achieve the strongest performance at moderate runtime compared to more complex deep architectures. Fast-MCD provides an attractive trade-off between runtime and detection quality, offering high PR-AUC at low execution time.

5.2. Results on UTFPR Dataset

Table 5 shows the best-performing run for each algorithm on the UTFPR dataset. In contrast to the FHC dataset, the best-

Table 4. Best-performing run for each algorithm (selected by PR-AUC) on the FHC dataset under different feature configurations. The class column denotes the method family: reconstruction- and prediction-based (R), distance- and density-based (D), isolation- and ensemble-based (I) or probabilistic and pattern-based methods (P). The learning type is indicated with ss for semi-supervised and u for unsupervised.

			all features		w/o wind & temp		w/o wind & temp & irrad	
Algorithm	type	class	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC
AE	ss	R	**0.868**	**0.966**	**0.860**	0.968	0.839	0.930
DAE	ss	R	0.839	0.956	0.853	**0.969**	**0.863**	**0.965**
Fast-MCD	ss	P	0.777	0.951	0.748	0.945	0.829	0.963
HealthESN	ss	R	0.564	0.876	0.694	0.890	0.737	0.923
Telemanom	ss	R	0.364	0.816	0.358	0.808	0.362	0.817
GrammarViz	u	P	0.349	0.557	0.165	0.824	0.170	0.812
KNN	u	D	0.298	0.854	0.481	0.875	0.673	0.918
EIF	u	I	0.246	0.812	0.358	0.831	0.396	0.840
OmniAnomaly	ss	R	0.234	0.799	0.213	0.778	0.218	0.805
IF-LOF	u	I	0.222	0.807	0.487	0.875	0.555	0.893
LOF	u	D	0.221	0.806	0.488	0.876	0.556	0.893
k-Means	u	D	0.211	0.767	0.242	0.817	0.252	0.805
LSTM-AD	ss	R	0.210	0.708	0.149	0.609	0.158	0.637
Sub-KNN	u	D	0.205	0.781	0.252	0.824	0.269	0.836
CBLOF	u	D	0.199	0.807	0.439	0.880	0.617	0.848
EncDec-AD	ss	R	0.198	0.656	0.174	0.687	0.233	0.690
iForest	u	I	0.194	0.796	0.202	0.776	0.193	0.688
RBF	ss	I	0.154	0.755	0.318	0.839	0.349	0.854
mSTAMP	u	P	0.147	0.772	0.174	0.781	0.196	0.792
HBOS	u	P	0.144	0.701	0.124	0.640	0.131	0.641
COF	u	D	0.128	0.701	0.281	0.844	0.315	0.863
COPOD	u	P	0.115	0.711	0.100	0.663	0.089	0.634
RobustPCA	ss	P	0.114	0.658	0.121	0.617	0.150	0.617
DeepAnT	ss	R	0.111	0.567	0.153	0.683	0.231	0.743
MTAD-GAT	ss	R	0.111	0.739	0.118	0.716	0.097	0.711
Sub-LOF	u	P	0.106	0.604	0.159	0.731	0.169	0.734
Torsk	u	R	0.106	0.661	0.138	0.708	0.117	0.683
PCC	u	P	0.099	0.641	0.102	0.610	0.095	0.583
Hybrid KNN	ss	D	0.084	0.635	0.149	0.737	0.109	0.678
DAMP	u	P	0.078	0.680	0.097	0.637	0.084	0.648
Random	-	-	0.054	0.502	0.054	0.502	0.054	0.502
TAnoGan	ss	R	0.035	0.310	0.035	0.313	0.036	0.333

performing methods are unsupervised distance- and density-based detectors. Using all available features of the dataset, KNN (PR-AUC 0.740; ROC-AUC 0.953) and CBLOF (PR-AUC 0.734; ROC-AUC 0.932) achieve the highest PR-AUC, followed by HBOS (PR-AUC 0.692; ROC-AUC 0.942).

Removing module temperature consistently improves performance for most methods. KNN reaches PR-AUC 0.815 (ROC-AUC 0.964) without temperature, while CBLOF improves to PR-AUC 0.798 (ROC-AUC 0.941). When additionally removing irradiance, CBLOF attains the best PR-AUC of 0.810 (ROC-AUC 0.946), and KNN remains essentially tied (PR-AUC 0.809; ROC-AUC 0.965).

Figure 2 (right) illustrates runtime vs. performance on the UTFPR dataset. The top-performing methods (KNN, CBLOF, HBOS) provide favorable runtime–accuracy trade-offs, whereas other approaches are substantially more expensive without delivering higher performance.

6. DISCUSSION

The experiments reveal two main insights: (i) the best-performing detection principle depends strongly on whether fault-free reference data is available, and (ii) environmental measurements can either support or hinder detection depending on the method class and dataset characteristics.

Semi-supervised advantage on the FHC dataset. The FHC setting provides an uncommon but practically relevant scenario: a physically identical reference system supplies fault-free data for training, while faults occur only in the test plant. In this scenario, semi-supervised reconstruction-based methods (AE/DAE) perform best. This suggests that learning a compact representation of "normal" PV behavior enables robust detection of both physically induced faults and injected data-quality anomalies. A practical nuance is that, despite the twin-plant design, the two plants still exhibit small sys-

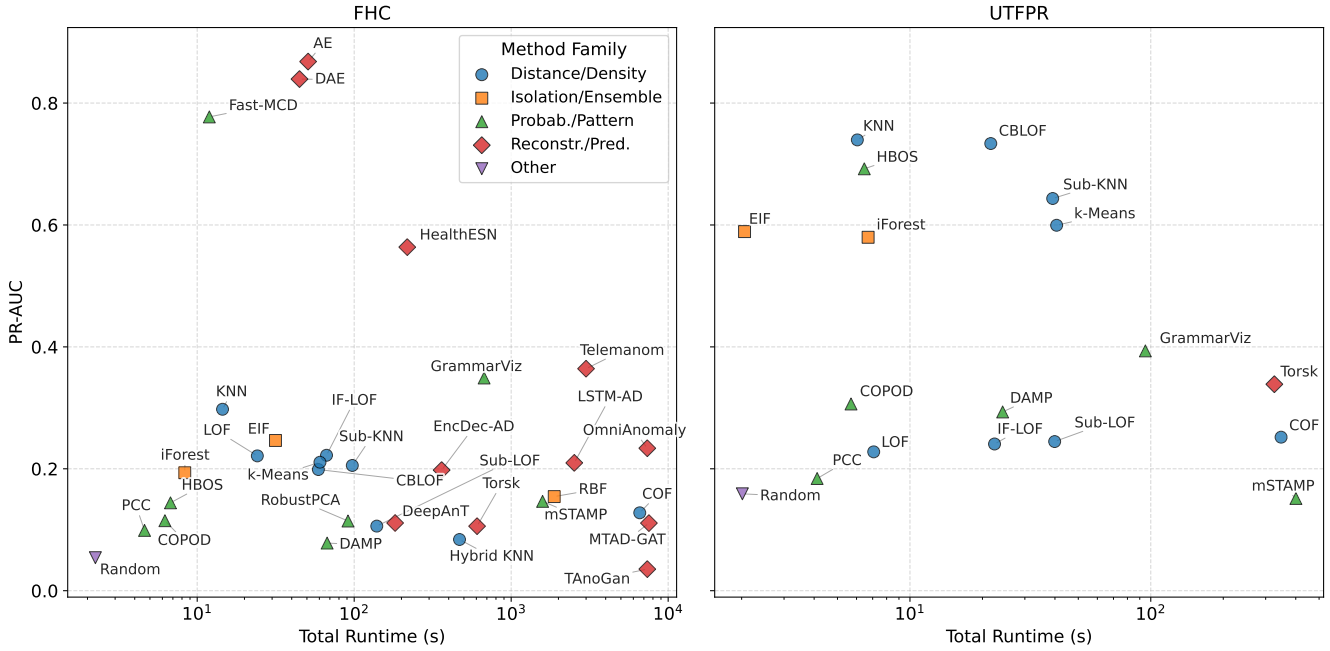

Figure 2. Runtime–performance comparison of anomaly detection methods on the FHC (left) and UTFPR (right) datasets using all available features. Each point shows PR-AUC versus total runtime (log scale). Different symbols indicate method families.

Table 5. Best-performing run for each algorithm (selected by PR-AUC) on the UTFPR dataset under different feature configurations. The class column denotes the method family: reconstruction- and prediction-based (R), distance- and density-based (D), isolation- and ensemble-based (I) or probabilistic and pattern-based methods (P). All methods use unsupervised learning.

		all features		w/o temp		w/o temp & irrad	
Algorithm	class	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC
KNN	D	**0.740**	**0.953**	**0.815**	**0.964**	0.809	**0.965**
CBLOF	D	0.734	0.932	0.798	0.941	**0.810**	0.946
HBOS	P	0.692	0.942	0.728	0.953	0.731	0.948
Sub-KNN	D	0.643	0.941	0.657	0.945	0.647	0.943
k-Means	D	0.599	0.924	0.659	0.939	0.644	0.932
EIF	I	0.589	0.898	0.654	0.927	0.684	0.927
iForest	I	0.580	0.900	0.593	0.910	0.512	0.890
GrammarViz	P	0.393	0.836	0.460	0.860	0.488	0.864
Torsk	R	0.339	0.774	0.305	0.762	0.321	0.785
COPOD	P	0.306	0.807	0.320	0.812	0.349	0.822
DAMP	P	0.293	0.712	0.254	0.715	0.337	0.775
COF	D	0.252	0.739	0.342	0.791	0.334	0.790
Sub-LOF	P	0.245	0.664	0.166	0.572	0.169	0.584
IF-LOF	I	0.241	0.642	0.323	0.733	0.339	0.750
LOF	D	0.228	0.626	0.319	0.708	0.336	0.726
PCC	P	0.184	0.561	0.185	0.537	0.219	0.639
Random	-	0.159	0.497	0.159	0.497	0.159	0.497
mSTAMP	P	0.151	0.450	0.156	0.486	0.146	0.462

tematic differences (e.g., slightly different shading patterns in the morning and evening). Such differences can appear "anomalous" to models trained purely on the reference plant. In the evaluation, these plant-to-plant differences are not labeled as anomalies. Detections triggered by systematic differences between the reference and modified plant are counted as false positives. Consequently, the FHC results evaluate not only fault sensitivity, but also the robustness of each method to moderate reference-to-target domain shift.

Why most algorithms improve after removing environmental features. Several distance- and density-based detectors perform substantially better when environmental features are excluded. A plausible explanation is that irradiance, wind, and temperature introduce dominant global structure corresponding to operating regimes. Methods that rely on global or local distances can then separate data primarily by regimes rather than by fault-related electrical deviations. Removing environmental features reduces this confounding structure and makes anomaly-relevant electrical deviations more salient. Importantly, this does not imply that environmental features are generally unhelpful: rather, their utility depends on dataset coverage. In longer-term datasets spanning multiple seasons and weather patterns, meteorological measurements may become essential for disambiguating benign variability from genuine faults.

The strong performance without irradiance can also be explained by the structural properties of our datasets. In both the FHC dataset and the UTFPR dataset, all strings are electrically identical and operate under the same environmental conditions. Under normal operation, this enforces a strong cross-string consistency constraint. Many evaluated methods might implicitly learn this multivariate consistency structure and can therefore detect faults as deviations between strings, even without explicit irradiance information. However, this property does not necessarily generalize to heterogeneous PV systems. In installations with differing string orientations, module types, or anomalies that affect all modules simultaneously (e.g., soiling, gradual degradation, etc.), irradiance information becomes important for reliable anomaly detection.

Why the best methods on UTFPR do not generalize to FHC. A notable observation is that algorithms that dominate the UTFPR dataset (KNN, CBLOF, HBOS) do not attain the same top performance on FHC. This discrepancy can be explained by structural dataset differences. UTFPR contains no completely fault-free days, and labeled faults occur almost exclusively during daytime operation. In this situation, algorithms that partition the data primarily by global structure can achieve decent scores by separating day and night regimes. In contrast, the FHC test set includes many fault-free days and anomalies that occur during both daytime and nighttime operation. Consequently, trivial regime separation does not yield strong performance. Detectors must instead capture subtle deviations within each regime. This difference highlights that "benchmark winners" can reflect dataset-specific separability rather than universally superior fault sensitivity.

Runtime and deployment implications. From a deployment perspective, methods such as AE/DAE/Fast-MCD (on FHC) and KNN/CBLOF/HBOS (on UTFPR) offer strong performance with manageable runtime, whereas several complex deep architectures incur high computational cost without commensurate gains. This indicates that for inverter- or edge-adjacent monitoring, lighter models may often be preferable, especially when frequent scoring is required.

A limitation of the benchmark is that identical hyperparameter search budgets could not be applied to all algorithms. The evaluated methods differ substantially in computational complexity, training time, and number of relevant hyperparameters. Applying the same exhaustive grid size to every method would have made the benchmark computationally impractical and would have allocated disproportionate resources to algorithms with very long runtimes. We therefore used method-specific tuning grids that cover the most relevant hyperparameters of each algorithm while keeping the total benchmark runtime feasible. As a consequence, the reported results should be interpreted as performance under a practical benchmarking budget rather than as the theoretical optimum achievable by each method. This particularly affects computationally expensive deep learning models, whose performance might improve under more extensive architecture and training-parameter optimization. Nevertheless, because all methods were evaluated using the same preprocessing, datasets, metrics, and selection criterion, the results remain informative for comparing algorithm families and identifying practically useful runtime-performance trade-offs.

A second limitation is that, as discussed in Section 4.5, the best hyperparameter configuration for each algorithm was selected based on PR-AUC on the labeled test set, since no separate labeled validation set was available. This can lead to optimistic metric values and means that the reported PR-AUC values should not be interpreted as expected deployment performance. However, the same selection procedure was applied consistently across all algorithms. The benchmark therefore focuses on relative comparisons between methods rather than on absolute performance estimates for real-world deployment.

7. CONCLUSION

This paper benchmarked multivariate time-series anomaly detection algorithms on two photovoltaic datasets: the newly introduced FHC twin-plant PV dataset and the publicly avail-

able UTFPR PV fault dataset. The results show that the most suitable detection approach depends strongly on the available data scenario. When fault-free reference data are available, as in the FHC twin-plant setup, semi-supervised reconstruction-based models achieve the strongest performance. In contrast, on the fully unsupervised UTFPR setting, classical distance- and density-based methods perform best.

The ablation study further shows that adding environmental measurements does not automatically improve anomaly detection performance. In the evaluated datasets, removing weather-related variables often improved PR-AUC, suggesting that such features can introduce confounding operating-regime structure for some algorithms. However, this finding should not be generalized without caution, since environmental information may become essential in longer-term datasets, heterogeneous PV systems, or fault scenarios affecting all strings simultaneously.

Overall, the study highlights the need to evaluate PV anomaly detection methods under realistic data conditions and with attention to feature selection, available reference data, and dataset-specific fault characteristics. Future work will extend the FHC measurement campaign to additional seasons and fault types, and investigate physics-informed features such as irradiance-normalized power to improve robustness across operating regimes and installation types.

ACKNOWLEDGMENT

This research was funded by the Austrian Research Promotion Agency (FFG) within the Digital Technologies 2022 program under the project PVSwarm (project number 902657).

During the preparation of this work, the authors used large language models (LLMs) in order to improve language and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

Abuhashish, M. N., Refaat, A., Kalas, A., Hamad, M. S., & Elfar, M. H. (2025). Towards accurate and reliable fault diagnosis in PV systems: Techniques, challenges, and future directions. *Process Safety and Environmental Protection*, *198*, 107217.

Alhazmi, A., Kholoud, M., & Eke, C. I. (2025). A systematic review of advances in deep learning architectures for efficient and sustainable photovoltaic solar tracking: Research challenges and future directions. *Sustainability*, *17*(21), 9625.

Amiri, A. F., Kichou, S., Oudira, H., Chouder, A., & Silvestre, S. (2024). 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*, *16*(3), 1012.

Bashar, M. A., & Nayak, R. (2020). TAnoGAN: Time series anomaly detection with generative adversarial networks. In *2020 IEEE symposium series on computational intelligence (SSCI)* (pp. 1778–1785).

Berghout, T., Benbouzid, M., Bentrcia, T., Ma, X., Djurović, S., & Mouss, L.-H. (2021). Machine learning-based condition monitoring for PV systems: State of the art and future prospects. *Energies*, *14*(19), 6316.

Bradl, H., Hofer-Schmitz, K., Grippa, P., & Hofer, G. (2026, March). *FHC Twin-Plant Photovoltaic Anomaly Detection Dataset.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.18979877` doi: 10.5281/zenodo.18979877

Branco, P., Gonçalves, F., & Costa, A. C. (2020). Tailored algorithms for anomaly detection in photovoltaic systems. *Energies*, *13*(1), 225.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104).

Chen, Q., Zhang, A., Huang, T., He, Q., & Song, Y. (2020). Imbalanced dataset-based echo state networks for anomaly detection. *Neural Computing and Applications*, *32*(8), 3685–3694.

Cheng, Z., Zou, C., & Dong, J. (2019). Outlier detection using isolation forest and local outlier factor. In *Proceedings of the conference on research in adaptive and convergent systems* (p. 161-168). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3338840.3355641

Ember. (2026, January). *European electricity review 2026* (Tech. Rep.). Ember. Retrieved from `{https://ember-energy.org/app/uploads/2026/01/EMBER-Report-European-Electricity-Review-2026.pdf}` (Accessed: 2026-02-23)

Gaviria, J. F., Narváez, G., Guillen, C., Giraldo, L. F., & Bressan, M. (2022). Machine learning in photovoltaic systems: A review. *Renewable Energy*, *196*, 298–318.

Goldstein, M., & Dengel, A. (2012). Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. In *Proceedings of the German Conference on Artificial Intelligence (KI) Poster and Demo Track* (p. 59-63).

Hariri, S., Kind, M. C., & Brunner, R. J. (2021). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1479-1489. doi: 10.1109/TKDE.2019.2947676

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern recognition letters*, *24*(9-10), 1641–1650.

Heim, N., & Avery, J. E. (2019). Adaptive anomaly detection

in chaotic time series with a spatially aware echo state network. *arXiv preprint arXiv:1909.01709*.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Soderstrom, T. (2018). Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 387–395).

Khandeparkar, V., Shreshtha, & Ramu, S. K. (2025). Effectiveness of supervised machine learning models for electrical fault detection in solar PV systems. *Scientific Reports*, *15*(1), 34919.

Lazzaretti, A. E., Costa, C. H. d., Rodrigues, M. P., Yamada, G. D., Lexinoski, G., Moritz, G. L., ... others (2020). A monitoring system for online fault detection and classification in photovoltaic plants. *Sensors*, *20*(17), 4688.

Li, B., Delpha, C., Diallo, D., & Migan-Dubois, A. (2021). Application of artificial neural networks to photovoltaic fault detection and diagnosis: A review. *Renewable and Sustainable Energy Reviews*, *138*, 110512.

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). COPOD: Copula-based outlier detection. In *2020 IEEE international conference on data mining (ICDM)* (p. 1118-1123). doi: 10.1109/ICDM50108.2020.00135

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth IEEE international conference on data mining* (p. 413-422). doi: 10.1109/ICDM.2008.17

Liu, Y., Duran, E., Bruce, A., Yildiz, B., Severiano, B. M., Ibrahim, I. A., ... Rougieux, F. (2025). A methodological review of cost-effective data-driven fault detection and diagnosis in distributed photovoltaic systems. *Applied Energy*, *401*, 126636.

Lu, Y., Wu, R., Mueen, A., Zuluaga, M. A., & Keogh, E. (2022). Matrix profile XXIV: Scaling time series anomaly detection to trillions of datapoints and ultra-fast arriving data streams. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (p. 1173–1182). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3534678.3539271

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*.

Malhotra, P., Vig, L., Shroff, G., Agarwal, P., et al. (2015). Long short term memory networks for anomaly detection in time series. In *Proceedings* (Vol. 89, p. 94).

Miraftabzadeh, S. M., Longo, M., Leva, S., & Matera, N. (2025). Data anomaly detection in photovoltaic power time-series via unsupervised deep learning with insufficient information. *Sustainable Energy, Grids and Networks*, *43*, 101769.

Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2018). Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE access*, *7*, 1991–2005.

Paffenroth, R., Kay, K., & Servi, L. (2018). Robust PCA for anomaly detection in cyber networks. *arXiv preprint arXiv:1801.01571*.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 427–438).

Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212-223. doi: 10.1080/00401706.1999.10485670

Sakurada, M., & Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis* (pp. 4–11).

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, *15*(9), 1779–1797.

Seghiour, A., Bendjeddou, Y., Mostefaoui, I. M., Chouder, A., Alharbi, H., Humayd, A. S. B., ... Babqi, A. (2026). Fault detection and diagnosis in photovoltaic systems using artificial intelligence and time–frequency analysis. *Scientific Reports*.

Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A. P., ... Frankenstein, S. (2015). Time series anomaly discovery with grammar-based compression. In *Proceedings of the international conference on extending database technology (EDBT)*. doi: 10.5441/002/edbt.2015.42

Shyu, M.-l., Chen, S.-c., Sarinnapakorn, K., & Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. In *Proceedings of the IEEE international conference on data mining workshops (ICDMW)* (pp. 172–179).

Song, H., Jiang, Z., Men, A., & Yang, B. (2017). A hybrid semi-supervised anomaly detection model for high-dimensional data. *Computational intelligence and neuroscience*, *2017*(1), 8501683.

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2828–2837).

Tang, J., Chen, Z., Fu, A. W.-C., & Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 535–548).

Toche Tchio, G. M., Kenfack, J., Kassegne, D., Menga, F.-D., & Ouro-Djobo, S. S. (2024). A comprehensive review of supervised learning algorithms for the diagnosis of photovoltaic systems, proposing a new approach using an ensemble learning algorithm. *Applied Sciences*, *14*(5), 2072.

Wenig, P., Schmidl, S., & Papenbrock, T. (2022). TimeEval: A benchmarking toolkit for time series anomaly detection algorithms. , *15*(12), 3678–3681. doi: 10.14778/3554821.3554873

Yairi, T., Kato, Y., & Hori, K. (2001). Fault detection by mining association rules from house-keeping data. In *proceedings of the 6th international symposium on artificial intelligence, robotics and automation in space* (Vol. 18, p. 21).

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., … Keogh, E. (2016). Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)* (p. 1317-1322). doi: 10.1109/ICDM.2016.0179

Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., & Salehi, M. (2024). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, *57*(1), 1–42.

Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., … Zhang, Q. (2020). Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE international conference on data mining (ICDM)* (pp. 841–850).

Ziegelmeir, J. (2019). *Development and comparison of self-learning modules for automated bench test data analysis of transient flight engine development tests* (Master's thesis). Technische Universität Berlin.