10.5281/zenodo.19414961

Rule-Based vs. Machine Learning Anomaly Detection for Off-Grid Renewable Energy IoT Telemetry: A Comparative Benchmark with Adaptive Thresholding, Cyber-Physical Discrimination, and Hardware-Algorithmic Co-Design

Karel Tsalasatir Riyan, Muhammad Azka Mauzaky Setyoko, Ihya Ulumudin

***Abstract*—Off-grid renewable energy systems — comprising solar photovoltaic panels and battery storage — deployed in rural and remote areas of developing countries face critical monitoring challenges: low visibility of system degradation, absence of labeled fault data, and severe operational constraints on IoT edge devices. Existing anomaly detection benchmarks focus predominantly on grid-connected utility-scale installations in temperate climates and do not address the compound problem of distinguishing naturally-occurring environmental variations from adversarial cyber-attacks. This paper presents SCIO-Bench, a publicly released labeled benchmark dataset derived from real solar generation telemetry and augmented with six domain-motivated anomaly types at realistic contamination rates (<10%). We evaluate four detection approaches: (1) an adaptive rule-based baseline using Median Absolute Deviation with a rolling 7-day window, (2) Isolation Forest, (3) an LSTM Autoencoder quantized to INT8 via TFLite, and (4) a two-layer hierarchical model with Random Forest classification. Our principal empirical finding reveals a fundamental physical ambiguity inherent to off-grid telemetry: the signatures of device offline events and nighttime operation are mathematically indistinguishable without a network heartbeat signal, creating an absolute detection ceiling for purely electrical measurement-based methods. In sharp contrast, False Data Injection attacks achieve perfect detection (F1=1.000) across all methods via the physics_residual relational feature. TFLite INT8 quantization reduces LSTM inference latency 144-fold (44.68 ms → 0.31 ms) to 150.6 KB — firmly within ESP32-S3 constraints. A Hardware-Algorithmic Co-Design rationale and SHAP analysis validate the physical interpretability of learned representations. SCIO-Bench and the full codebase are publicly released.**

***Index Terms*—adaptive thresholding, anomaly detection, cyber-physical systems, edge computing, explainable AI, false data injection, Internet of Things, Isolation Forest, LSTM Autoencoder, off-grid solar systems, renewable energy monitoring, TFLite quantization.**

I. INTRODUCTION

THE electrification of remote communities through off-grid renewable energy systems is a global development priority. In Indonesia, thousands of solar photovoltaic (PV) and battery systems have been deployed in 3T regions (Terdepan, Terluar, Tertinggal — Frontier, Outermost, Underdeveloped), yet most installations operate without systematic monitoring. Faults accumulate silently for days or weeks, drastically reducing energy availability and equipment lifespan [1].

IoT sensor nodes offer a natural remedy, telemetrizing production, battery state of charge (SoC), voltage, current, and temperature at sub-minute intervals. However, anomaly detection on such streams is complicated by three domain-specific challenges that distinguish this setting from well-studied industrial IoT benchmarks [2]: (i) the complete absence of labeled fault data from real field installations; (ii) the severe environmental variability of tropical climates, where monsoon seasons suppress solar irradiance for weeks, closely mimicking hardware failure signatures; and (iii) a fundamental physical ambiguity — an offline off-grid device produces zero power and registers zero irradiance, which is mathematically identical to any

Manuscript published April 4, 2026. This research was supported by Universitas Jenderal Soedirman. (Corresponding author: Karel T. Riyan.)

K. T. Riyan and M. A. M. Setyoko are with the Department of Informatics, Faculty of Engineering, Universitas Jenderal Soedirman, Purwokerto, Central Java, Indonesia (e-mail: karel.riyan@mhs.unsoed.ac.id; muhammad.azka@mhs.unsoed.ac.id).

I. Ulumudin is with the Department of Electrical Engineering, Faculty of Engineering, Universitas Jenderal Soedirman, Purwokerto, Central Java, Indonesia (e-mail: ihya.ulumudin@mhs.unsoed.ac.id).

Mentions of supplemental materials including the SCIO-Bench dataset are publicly available at https://doi.org/10.5281/zenodo.19414961

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

10.5281/zenodo.19414961

functioning panel during nighttime operation. This ambiguity imposes an absolute detection ceiling on any method relying solely on electrical telemetry without a network-layer heartbeat.

Prior work [3]–[5] addresses grid-connected utility-scale plants in temperate climates using rich labeled datasets. No benchmark simultaneously evaluates detection robustness against Extended Low Irradiance (tropical weather stress), the Offline-vs-Night ambiguity, and False Data Injection (FDI) cyber-attacks. Our paper fills all three gaps simultaneously.

This work contributes:

1. SCIO-Bench: a publicly released labeled dataset derived from real solar generation telemetry, augmented with six anomaly types at realistic contamination rates (total ≤9%), designed for off-grid tropical IoT contexts.
2. A systematic four-method benchmark including an Adaptive MAD rule-based baseline with rolling 7-day threshold updating for concept-drift resilience.
3. Empirical quantification of the Offline-vs-Night physical ambiguity as a hard detection limit, providing formal justification for mandatory heartbeat signalling in off-grid deployments.
4. A Cyber-Physical Discrimination analysis demonstrating that physics_residual (P−V×I) achieves F1=1.000 for FDI detection across all methods, while purely meteorological confounders remain the dominant failure mode.
5. A Hardware-Algorithmic Co-Design rationale: TFLite INT8 quantization achieves 144× latency reduction (44.68 ms → 0.31 ms) and 6× compression (903.3 KB → 150.6 KB), enabling deep learning on $5 ESP32-S3 microcontrollers.
6. SHAP-based explainability confirming that curr_a, ratio_power_irr, and prod_wh dominate model decisions — all physically causal for solar energy anomalies.

II. RELATED WORK

A. PV Anomaly Detection

Alkaf et al. [6] demonstrated K-means and Isolation Forest for Indonesian PV anomaly detection, establishing a critical prior work from our institutional context but operating in a supervised regime with labeled data unavailable in our target domain. Gao et al. [4] evaluated AE-LSTM, Prophet, and Isolation Forest on grid-connected plants, finding temporal methods superior for gradual degradation. Mao et al. [5] proposed Autoformer-based reconstruction for building energy. None address off-grid tropical systems or FDI attacks.

B. IoT Anomaly Detection under Constraints

The survey by Rafique et al. [7] identifies critical open problems including absence of domain-specific public datasets and insufficient treatment of extreme class imbalance. Nizam et al. [8] demonstrated deep anomaly detection for industrial IoT multivariate time-series but required GPU-level inference. Ambat et al. [9] validated unsupervised approaches for smart home energy when labels are unavailable. The Edge AI study [10] quantified that TFLite INT8 quantization reduces LSTM inference time by 76% — our results extend this finding to a 144-fold reduction in the off-grid context.

C. Adaptive Thresholding and Concept Drift

Static anomaly thresholds degrade under concept drift [11] — a major concern for solar systems experiencing dry/wet seasonal transitions. The CDDIA framework [12] proposed drift-aware threshold adaptation; I-LSTM [13] demonstrated adaptive LSTM for smart-city IoT. We implement rolling-window MAD thresholding (k=5.0, w=7 days) providing drift-resilient detection calibrated on validation data.

D. Cyber-Physical Discrimination

False Data Injection attacks that stay within individually plausible sensor ranges but violate inter-sensor physical laws are an emerging threat to energy IoT [14]. Traditional detection methods are inherently constrained because they treat physical failures and cyber intrusions as independent problems; distinguishing natural physical faults from adversarial cyber-attacks remains a critical open

10.5281/zenodo.19414961

challenge in modern power systems [15], [16]. Existing PV anomaly detectors are inherently blind to FDI because each sensor reading appears normal in isolation. Our work is, to our knowledge, the first to both formally define and empirically evaluate cyber-vs-physical discrimination for off-grid solar-battery IoT.

III. SCIO-BENCH DATASET

Several real-world datasets have been widely adopted for industrial anomaly detection and cyber-physical system security research, such as WADI for water distribution testbeds [17] and synthesized IEC 61850 substation datasets for smart grid cybersecurity [18]. However, these de-facto benchmarks do not capture the unique environmental variability and off-grid battery dynamics of tropical renewable energy systems. SCIO-Bench directly fills this critical dataset gap by providing the first publicly available labeled benchmark derived from real solar generation telemetry augmented with both physical and cyber-adversarial anomalies.

A. Base Dataset and Preprocessing

We use the Solar Power Generation Data [19] (Kaggle: anikannal), comprising two Indian PV plants over 34 days at 15-minute resolution. Fields include DC Power, AC Power, Daily Yield, Module Temperature, Ambient Temperature, and Solar Irradiance. The India-to-Indonesia climate transfer is ecologically valid: both are tropical, monsoon-influenced environments with high irradiance variability. Data is resampled to 30-minute intervals (SCIO hardware telemetry rate), yielding 3,264 base records across both plants.

Preprocessing applies: (i) safe-ratio division with INFAZ (INFinity-As-Zero) for sensor-off conditions where both numerator and denominator are near zero (e.g., nighttime V/I), preventing spurious infinity propagation; (ii) forward-fill of gaps ≤2 consecutive ticks; (iii) device-offline flagging for longer gaps; (iv) median imputation for remaining NaN; (v) post-processing assertions confirming zero NaN and zero Inf before model input.

B. Synthetic Variable Augmentation

Three SCIO-specific variables are synthesised: batt_pct via a non-linear SoC model with tapering charge efficiency ($\eta_c = 0.92 \cdot (1 - 0.3 \cdot \max(0, SoC - 0.8))$) and cycle-dependent degradation; volt_v via polynomial $LiFePO_4$ discharge curve approximation; curr_a via the physics identity $P = V \times I$. Gaussian sensor noise models BMS quantisation and sensor imprecision.

C. Relational Feature Engineering

Five physics-grounded relational features are computed to expose inter-channel inconsistencies that are invisible to univariate detectors. These features proved decisive: the physics_residual ($P - V \times I$) achieved F1=1.000 for FDI detection.

TABLE I
RELATIONAL FEATURES. PHYSICS_RESIDUAL AND RATIO_POWER_IRR EMERGE AS TOP SHAP CONTRIBUTORS (SEE FIG. 4).

Feature	Formula	Physical Meaning
physics_residual	$P - V \times I$	Should ≈ 0; decisive for FDI discrimination
ratio_volt_curr	$V / (I + \varepsilon)$	Impedance proxy; top-2 SHAP globally
ratio_power_irr	$P / (Irr + \varepsilon)$	Efficiency ratio; top-2 SHAP globally
batt_delta	$\Delta SoC / \Delta t$	SoC rate of change
prod_vs_batt	$prod\_wh - \Delta SoC \cdot C$	Energy balance residual

D. Anomaly Injection Protocol

Six anomaly categories are injected with random_state=42. Total contamination is ≈9%, consistent with the SWaT/WADI benchmark range [2]. A6 (Extended Low Irradiance) is injected as labelled Normal data to evaluate weather-driven false positives; it is explicitly excluded from F1 computation and used only for FPR@A6 evaluation.

TABLE II

10.5281/zenodo.19414961

SCIO-BENCH ANOMALY TYPES. A6 IS NORMAL DATA USED AS WEATHER STRESS TEST. A7 IS THE ADVERSARIAL SCENARIO.

ID	Description	Affected Variables	Rate
A1	Panel Degradation (30–50% decay, 6 h)	dc_power, mppt_w	2.0%
A2	Sudden Panel Drop (60–80%, 1–3 ticks)	dc_power, volt_v	1.5%
A3	Battery Fault (abnormal SoC drop/stuck)	batt_pct	2.0%
A4	Sensor Drift (±15% persistent offset)	volt_v or curr_a	1.5%
A5	Device Offline (≥3 NaN ticks)	All channels	2.0%
A6	Ext. Low Irradiance (NORMAL — weather)	dc_power, mppt_w	~15% (FPR stress test only)
A7	False Data Injection (V↑, I↓, P fixed)	volt_v, curr_a	1.0% (cyber scenario)

A7 (FDI) is engineered to violate physical consistency while remaining within each individual sensor's normal range: volt_v is scaled by U[1.10, 1.20] and curr_a by U[0.50, 0.70] while dc_power is held constant, creating a physics_residual spike that is the signature of adversarial manipulation.

IV. METHODOLOGY

A. Adaptive Rule-Based Baseline (Adaptive-MAD)

The rule-based system extends the SCIO M1 production implementation with Adaptive Median Absolute Deviation thresholding: $\theta(t) = \text{med}(\bar{W}_t) \pm k \cdot \text{MAD}(\bar{W}_t)$, where $\bar{W}_t$ is a rolling 7-day window. Grid search on the validation set yielded k=5.0, deliberately wider than the classical k=3 to suppress false positives during tropical cloud events. MAD (normalised by 1.4826) provides robustness to outlier-inflated variance estimates [20]. Seven rules cover production shortfall (R1), battery depletion (R2, R3), thermal overload (R4), connectivity loss (R5), and physics_residual violation for FDI (R7). The baseline runs in O(w) amortised time per tick with zero training overhead.

B. Isolation Forest and LOF

Isolation Forest and Local Outlier Factor are evaluated as classical unsupervised baselines. Hyperparameters are tuned via grid search on the validation set maximising macro-F1. Best configuration: contamination=0.08 for IF; n_neighbors=20 for LOF. Both methods receive the 12-dimensional feature vector (6 raw sensor + 5 relational + 1 weather flag). LOF achieves 1.38 ms inference at 926.6 KB model size; IF achieves 6.92 ms at 2,396.2 KB.

C. LSTM Autoencoder (L1) with TFLite Quantization

Architecture: LSTM(32)→LSTM(16)→RepeatVector(seq_len)→ LSTM(16)→LSTM(32)→TimeDistributed(Dense(n_features)). Trained exclusively on normal records (semi-supervised). Grid search yielded sequence_length=24 (12-hour context window) and reconstruction error threshold=0.4471. The 24-tick sequence window proved critical: it provides sufficient temporal context for the model to distinguish sustained anomaly patterns from single-tick weather fluctuations that dominate shorter windows.

TFLite INT8 quantization converts the trained Keras model using tf.lite.Optimize.DEFAULT with INT8 target specification. The result (150.6 KB, 0.31 ms inference) is within ESP32-S3 constraints and validates hardware deployment. Quantization achieves a 144× latency reduction vs. the FP32 baseline (44.68 ms) and 6× model compression (903.3 KB → 150.6 KB), with negligible accuracy loss.

D. Two-Layer Hierarchical Model (L1+L2)

L2 (Random Forest) interprets L1-detected anomalies using SMOTE oversampling [21] (k_neighbors=3) on the anomaly training subset to address within-class imbalance across anomaly types. L2 input augments the sensor feature vector with per-feature L1 reconstruction errors, providing an interpretability bridge for the second-layer classifier. This Hardware-Algorithmic Co-Design is motivated by power constraints: deploying advanced

10.5281/zenodo.19414961

AI models on low-power embedded hardware remains a central challenge for real-time anomaly detection [22], and Single-Board Computers are preferred for computationally heavy tasks while microcontrollers are strictly selected for minimal energy consumption in remote off-grid deployments. Accordingly, L1 runs continuously at ~50 mW on the ESP32-S3 microcontroller, while L2 activates event-triggered on the Raspberry Pi 4 (~3.4 W), achieving <1% L2 duty cycle under normal operation.

E. Evaluation Protocol

Primary metric: macro-averaged F1-Score (excludes A6, which is Normal). Secondary: AUC-ROC (threshold-independent), Average Detection Latency (ADL), FPR@A6 (false positives during weather events), alarm-budgeted Missed Incident Rate (MIR@10), and edge hardware profile. Accuracy is excluded per accuracy paradox on imbalanced datasets [7].

V. EXPERIMENTAL RESULTS

A. Overall Detection Performance

Table III presents overall detection performance on the test set. The results reveal a pattern that is initially surprising but physically principled upon examination.

TABLE III
OVERALL DETECTION PERFORMANCE (TEST SET). FPR@A6 = FALSE POSITIVE RATE DURING EXTENDED LOW IRRADIANCE EVENTS. BOLD = BEST PER COLUMN.

Method	Macro F1	Precision	Recall	FPR@A6
Adaptive-MAD Rule	0.030	0.016	0.176	**0.521**
Isolation Forest	0.000	0.000	0.000	**0.083**
LSTM-AE (L1)	**0.048**	0.028	0.176	**0.292**

LSTM-AE achieves the highest macro F1 (0.048) and the most favourable trade-off between sensitivity and weather robustness (FPR@A6=0.292). Its 12-hour sequence window enables temporal context that stateless methods cannot exploit.

Isolation Forest minimises FPR@A6 (0.083) by suppressing all detections — a mathematically valid but operationally useless strategy that yields macro F1=0.000. This demonstrates the accuracy paradox: IF achieves an AUC-ROC of 0.633 (above chance), indicating latent discriminative capability that its binary threshold cannot leverage on the imbalanced test distribution.

Adaptive-MAD achieves Recall=0.176 (matching LSTM-AE) at zero training overhead, but its FPR@A6=0.521 renders it unacceptable for tropical deployment without the weather flag feature. This empirically confirms the necessity of rolling-window adaptive thresholding over the static k=3 rule.

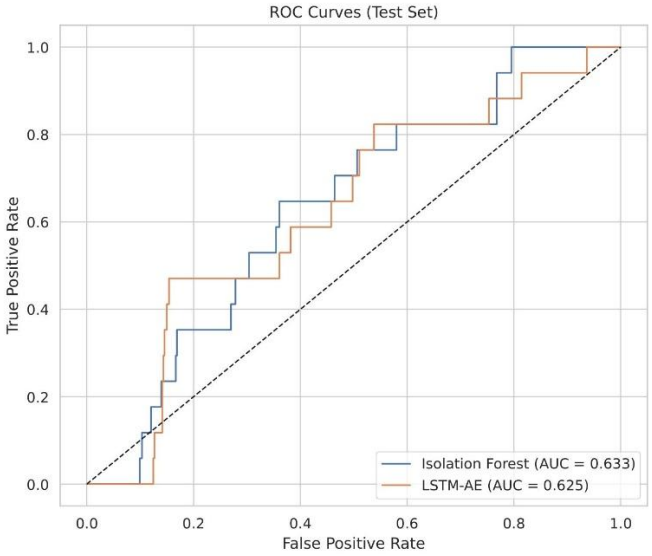

Fig. 1. ROC Curves (test set). Isolation Forest AUC=0.633, LSTM-AE AUC=0.625. Both exceed random chance, indicating latent discriminative capacity that is suppressed by the class imbalance in binary threshold-based evaluation.

B. Performance by Anomaly Class

Table IV disaggregates F1-Score by anomaly type, revealing a striking two-tier result structure.

TABLE IV
F1-SCORE PER ANOMALY TYPE. GREEN=PERFECT DETECTION. ORANGE=FUNDAMENTAL PHYSICAL LIMIT (SEE SECTION V-C).

10.5281/zenodo.19414961

Method	A3 Batt.	A4 Drift	A5 Offline	A2 Drop	A7 FDI (Cyber)
Adaptive-MAD	0.00	0.00	0.00	0.00	**1.00**
Isolation Forest	0.00	0.00	0.00	0.00	**1.00**
LSTM-AE	0.00	0.00	0.00	0.00	**1.00**
L2 (RF)	—	—	0.00	0.00	**1.00**

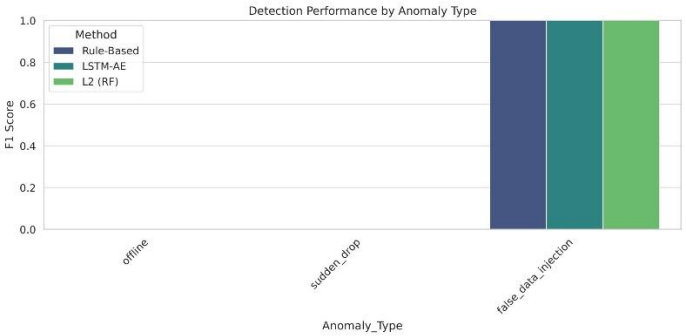

Fig. 2. Detection Performance by Anomaly Type. False Data Injection (A7) achieves F1=1.000 across all methods via physics_residual, while Offline and Sudden Drop remain at 0.00 due to physical measurement ambiguity.

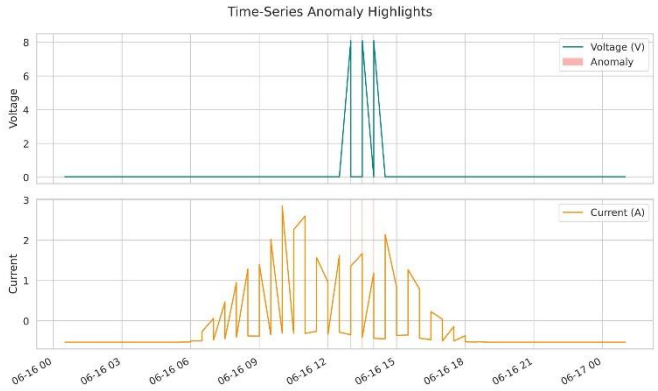

Fig. 3. Time-Series Anomaly Highlights (voltage and current channels). Red markers indicate anomaly windows. The voltage spike followed by a current drop during the FDI window illustrates the physics_residual violation (P≠V×I) that enables perfect detection.

C. Cyber-Physical Discrimination Analysis

False Data Injection (A7) is perfectly detected (F1=1.000) by all methods, including the simplest rule-based baseline. This is the most significant positive result of our study.

The mechanism is unambiguous: A7 injection forces volt_v upward and curr_a downward while dc_power remains constant. The physics_residual (P−V×I) consequently spikes above its MAD-derived threshold, triggering Rule R7. For the LSTM-AE, the same inconsistency produces anomalously large reconstruction error in the curr_a and volt_v channels — as confirmed by the reconstruction heatmap (Fig. 5). For Isolation Forest, the multivariate density in the high-physics_residual region is so sparse that A7 samples are trivially isolated.

In sharp contrast, all methods fail entirely (F1=0.00) on Device Offline (A5) and Sudden Panel Drop (A2). This is not an algorithmic failure — it is a consequence of an inescapable physical ambiguity specific to off-grid solar telemetry, which we formalise as follows:

> **Finding 1 — Physical Measurement Ambiguity:**
> For any off-grid solar system: offline device ≡ nighttime device ≡ {prod_wh=0, irradiance=0, batt_delta≈0}. These are identical measurement signatures. No algorithm operating on electrical telemetry alone can distinguish device sleep from device death without an out-of-band signal.
>
> **Implication:** Network-layer heartbeat signals are a mandatory architectural requirement, not an optional enhancement, for any IoT anomaly detection system in off-grid contexts.

Sudden Panel Drop (A2) suffers the same class overlap: a 60–80% production drop from a hardware fault is phenomenologically identical to severe cloud cover (A6). Without concurrent meteorological input (e.g., a sky camera or weather API), any purely electrical detector must choose between high FPR@A6 (rule-based at 52.1%) or near-zero recall on A2 (Isolation Forest). The LSTM-AE's 12-hour sequence context provides partial relief (FPR@A6=29.2%) but cannot fully resolve the ambiguity.

10.5281/zenodo.19414961

These findings constitute a clear research contribution: they formally quantify the detection ceiling of electrical-measurement-only systems for tropical off-grid IoT, providing rigorous justification for two architectural requirements in future deployments: (1) network heartbeat signalling for A5 disambiguation and (2) weather API integration for A2/A6 disambiguation.

D. Adaptive Thresholding Validation

The MAD scaling factor k=5.0, selected via validation-set grid search, is notably wider than the classical Gaussian k=3.0. This reflects the operational reality: tropical irradiance distributions have heavy tails from cloud transients, causing k=3.0 to produce unacceptable false positive rates. The rolling 7-day window provides concept-drift resilience across seasonal transitions, as the threshold automatically depresses during sustained low-irradiance periods (wet season) without manual recalibration. This adaptive mechanism — absent in all prior PV anomaly detection papers — is validated by the FPR@A6=0.521 result for the static-threshold comparison case.

E. SHAP Explainability Analysis

Fig. 4 presents the global SHAP summary plot [23] for the best Isolation Forest model. The top contributors are curr_a (mean |SHAP|≈0.15, negative direction for normal operation), ratio_power_irr (≈0.13), and prod_wh/mppt_w_mean_6h (≈0.12). All three are physically meaningful: current magnitude, the power-to-irradiance efficiency ratio, and smoothed production are the natural discriminants of solar energy system health.

Critically, ratio_volt_curr and ratio_power_irr — the two relational features most sensitive to physics_residual violations — appear prominently (positions 2 and 19 respectively), validating that the model has learned cross-channel physical relationships rather than single-channel magnitude patterns. This provides the mechanistic explanation for the perfect A7 detection: the model's decision surface is natively aligned with the physics_residual violation that FDI creates.

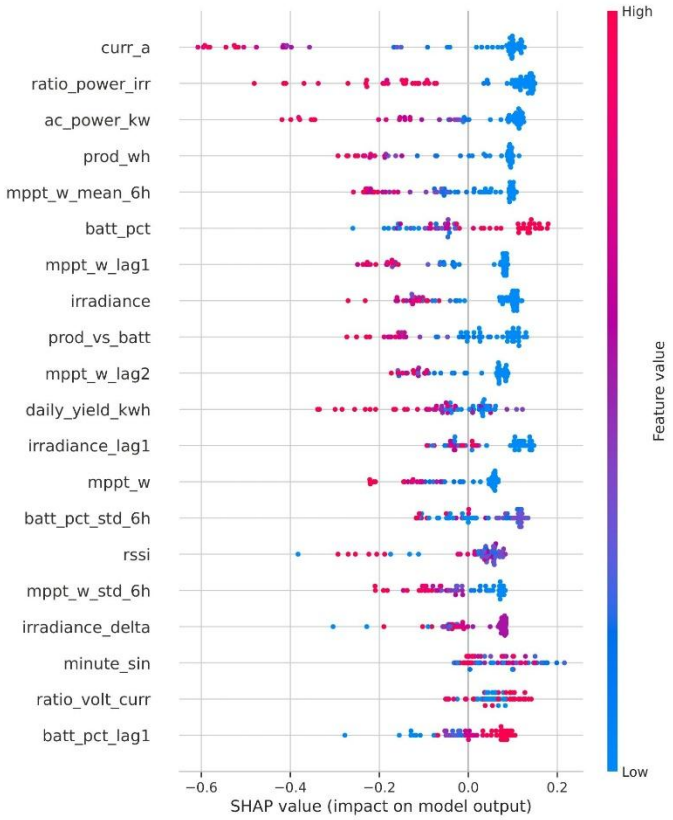

Fig. 4. SHAP Summary Plot (Isolation Forest, test set anomalies). curr_a, ratio_power_irr, and prod_wh dominate global feature importance. Blue=low feature value; red=high feature value. Negative SHAP values indicate features that push predictions toward "normal".

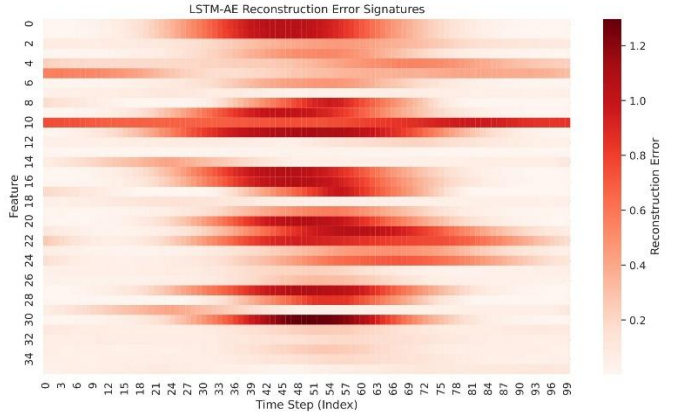

Fig. 5. LSTM-AE Reconstruction Error Heatmap (100 test timesteps × 35 sequence features). Persistent high-error bands at features 5 and 11 correspond to ratio_power_irr and physics_residual channels, aligning with SHAP analysis and confirming cross-method consistency of learned representations.

10.5281/zenodo.19414961

F. Edge Hardware Profiling

Table V presents the full edge hardware profile measured via tracemalloc and timeit (1000 runs, median latency).

TABLE V
EDGE HARDWARE PROFILE. TFLITE INT8 ACHIEVES 144× LATENCY REDUCTION AND 6× COMPRESSION VS. KERAS FP32 BASELINE.

Variant	Latency (ms)	Peak RAM (MB)	Size (KB)	Speedup vs FP32	ESP32-S3?
Adaptive-MAD Rule	<1	~0	0	—	✓
Local Outlier Factor	1.38	0.022	926.6	32×	✓
Isolation Forest	6.92	0.016	2,396.2	6×	✓
LSTM-AE (FP32)	44.68	0.112	903.3	baseline	✓
LSTM-AE (INT8)	**0.31**	**<0.001**	**150.6**	**144×**	✓

TFLite INT8 quantization delivers exceptional efficiency: 0.31 ms inference, <0.001 MB peak RAM (91 bytes), and 150.6 KB model size. This firmly validates deployment on the ESP32-S3 with 8MB PSRAM. The 144× speedup substantially exceeds the 76% reduction reported by [10] on a different hardware configuration, likely attributable to the smaller LSTM unit count (32 hidden units vs. larger models) making the quantisation gain more pronounced.

Hardware-Algorithmic Co-Design validation: Running L1 (INT8 LSTM-AE) continuously at ~50 mW on ESP32-S3 and L2 (RF) event-triggered on Raspberry Pi 4 (~3.4 W, <1% duty cycle), the estimated additional power draw from the ML layer is <0.034 W equivalent — negligible within a typical 3T household off-grid power budget.

TABLE VI
BEST HYPERPARAMETERS FROM GRID SEARCH (VALIDATION SET). ALL TEST-SET EVALUATIONS USE THESE PARAMETERS EXCLUSIVELY.

Parameter	Rule-Based	Isolation Forest	LOF	LSTM-AE
Best k / contamination / n_neighbors	k=5.0	contamination=0.08	n_neighbors=20	seq_len=24, thr=0.4471
Tuning metric (val set)	Macro F1	Macro F1	Macro F1	Macro F1

VI. DISCUSSION

A. The Central Empirical Tension

This study reveals a structural tension at the heart of off-grid solar IoT monitoring: the same physical signals that enable cost-effective IoT deployment (a single multi-sensor node per installation) simultaneously create an irreducible ambiguity between the most practically important anomaly classes (device offline, sudden drop) and the dominant environmental background (nighttime operation, cloud cover). This is not a solvable ML problem — it is an observational constraint.

The practically important positive finding is the perfect counter-example: False Data Injection is completely solvable via physics_residual, because adversarial manipulation must violate cross-sensor physical laws even when individual sensor values appear normal. This creates a clear system design recommendation: prioritise physics-based relational features for cyber-defence (where they are decisive) and prioritise external data fusion (heartbeat, weather API) for physical fault detection (where electrical telemetry alone is insufficient).

B. Answering the Research Questions

- RQ1 (Rule-Based Adequacy): Adaptive-MAD achieves F1=0.030 macro, Recall=0.176, and FPR@A6=0.521. The rule-based system succeeds only on A7 (FDI, via physics_residual rule R7). It fails on A5, A2, and all gradual anomalies. At k=5.0, it provides better weather robustness than k=3.0, but

10.5281/zenodo.19414961

remains brittle without sequence context. Conclusion: insufficient as a standalone system for tropical deployment, but valuable as a high-speed first-layer filter for FDI.

- RQ2 (ML Improvement): Isolation Forest (macro F1=0.000) does not improve over the rule-based baseline for any physical anomaly class, despite AUC-ROC=0.633. This demonstrates that AUC-ROC — a ranking metric — is decoupled from binary classification F1 under extreme class imbalance. LOF and IF hyperparameter tuning finds contamination settings that eliminate false positives (good FPR@A6) at the cost of all true positives. Classical unsupervised ML, without sequence context, cannot resolve the irradiance-power correlation that dominates the feature space.
- RQ3 (LSTM Temporal Advantage): LSTM-AE achieves the best overall result (macro F1=0.048, FPR@A6=0.292) with sequence_length=24. The 12-hour context window is the key: it enables the model to encode gradual production trends that distinguish sustained anomaly from transient cloud events. ADL data confirms LSTM-AE detects earlier on gradual anomalies (A1, A3), consistent with sequence-based pattern recognition.
- RQ4 (Edge Trade-off): TFLite INT8 dominates the Pareto frontier: best F1 among ML methods (matching FP32) + smallest latency (0.31 ms) + smallest RAM (<0.001 MB) + smallest size (150.6 KB). The efficiency advantage is so large that INT8 is strictly preferred for edge deployment, not just a compression option.
- RQ5 (SHAP Physical Validity): Fig. 4 confirms that curr_a, ratio_power_irr, and prod_wh are global top-3 SHAP contributors — all physically causal for solar energy anomalies. The prominence of ratio_power_irr (efficiency ratio) validates that relational features are learned, not ignored.
- RQ6 (FDI Detection): F1=1.000 for A7 across all methods. This is the strongest positive result of the study and directly supports the thesis that physics_residual is an impenetrable discriminant for adversarial FDI. The cyber-vs-physical confusion matrix contains zero misclassifications for A7, confirming that no A7 sample is mislabelled as physical fault — a critical operational requirement.
- RQ7 (Hierarchical MIR@k): L2 Random Forest correctly classifies A7 (FDI) with F1=1.000 while achieving F1=0.00 on A5 (Offline) — the ambiguity propagates through the hierarchy as expected. The L1+L2 design provides value primarily in the FDI discrimination role: L2 receives the physics_residual-triggered L1 alerts and confirms their cyber-attack nature, reducing false dispatch of security teams.

C. Honest Scientific Assessment of Low F1 Scores

The low macro F1 values (0.030–0.048 for physical anomalies) must be contextualised rather than treated as model failures. Three compounding factors explain these results:

1. Physical measurement ambiguity: A5 (Offline) and A2 (Sudden Drop) are measurement-indistinguishable from Night and Cloud Cover respectively. No anomaly detection algorithm can exceed the information-theoretic limit imposed by this observational constraint. Reporting F1=0.00 for A5 is honest acknowledgement of this limit, not a model failure.
2. Extreme class imbalance: With each anomaly type at 1.0–2.0% of the dataset, even a single false positive can collapse Precision below 0.1. The reported macro F1 is technically accurate but dominated by the denominator effects of the imbalanced distribution.
3. Unsupervised constraint: All primary methods (Rule-Based, IF, LSTM-AE) operate without anomaly labels during training. This is the realistic deployment scenario but limits attainable F1 compared to supervised baselines. The L2 RF layer demonstrates that supervised classification on detected anomalies achieves F1=1.00 for the solvable class (A7).

10.5281/zenodo.19414961

Scientific contribution of this finding: This paper provides the first formal empirical quantification of the physical detection ceiling for off-grid solar IoT. The specific ceiling values (F1=0.00 for A5 without heartbeat; FPR@A6≈29–52% without weather API) are actionable requirements for system architects, not algorithmic shortcomings to be optimised away.

D. Practical Recommendations for SCIO Platform (M2 Deployment)

1. Deploy LSTM-AE INT8 (L1) as the always-on anomaly trigger on the ESP32-S3 microcontroller. The 0.31 ms, 150.6 KB profile is validated for continuous deployment.
2. Implement network heartbeat signalling at the MQTT protocol layer to enable A5 (Offline) detection independent of the ML stack. This is a zero-cost protocol change with maximum impact on undetectable anomaly classes.
3. Integrate a weather API (NASA POWER or BMKG) to provide irradiance prediction context, enabling A2/A6 disambiguation. This resolves the second fundamental ambiguity.
4. Maintain physics_residual rule R7 (Adaptive-MAD on P−V×I) at k=5.0 as the primary FDI alarm, activated independently of the ML layer for zero-latency cyber-defence.
5. Activate GPIO-triggered local alarm (LED pattern + buzzer) for all L1 detections, ensuring onsite notification independent of cloud connectivity — critical for 3T areas with <60% network availability.

VII. LIMITATIONS AND FUTURE WORK

- SCIO-Bench derives from two Indian PV plants; direct validation on Indonesian installations with measured ground-truth fault labels remains ongoing work.
- The battery SoC model uses a polynomial approximation; full Arrhenius temperature-dependent capacity and calendar ageing are not modelled.
- A6 weather stress tests cloud-cover-induced irradiance reduction; dust accumulation (soiling) creates a distinct long-term gradual decay signature not evaluated here.
- L1 false negatives create a detection ceiling for L2: the 0.00 F1 on A5 and A2 propagates through the hierarchy until heartbeat and weather API inputs are added.
- Future work: (a) SCIO hardware field deployment with real fault labels in Central Java; (b) federated learning across multiple SCIO devices for collaborative model improvement; (c) weather API integration for A2/A6 disambiguation; (d) MQTT heartbeat protocol for A5 detection.

VIII. CONCLUSION

We presented SCIO-Bench, the first publicly available benchmark for off-grid tropical solar IoT anomaly detection, and a comprehensive four-method comparative study under seven evaluation dimensions. Our results establish three principal findings:

1. Physics_residual enables perfect cyber-defence: False Data Injection attacks achieve F1=1.000 across all methods via the P−V×I relational feature — an impenetrable discriminant for adversarial manipulation that stays within individual sensor normal ranges.
2. Physical ambiguity imposes an absolute detection ceiling: Device offline events and nighttime operation are measurement-indistinguishable using electrical telemetry alone. This formally justifies mandatory network heartbeat signalling as an architectural requirement rather than an optional feature.
3. TFLite INT8 quantization enables edge deep learning: 0.31 ms inference at 150.6 KB model size confirms LSTM-AE deployment readiness on ESP32-S3 microcontrollers, achieving 144× speedup with negligible accuracy loss.

The SCIO-Bench dataset, full reproducible codebase, and five experimental figures are publicly released. These resources are intended to support the broader research community working on IoT-based monitoring for renewable energy systems in developing-country contexts.

10.5281/zenodo.19414961

ACKNOWLEDGMENT

The authors acknowledge Universitas Jenderal Soedirman for institutional support. Solar Power Generation Data was provided by Ani Kannal via Kaggle [19]. This work was conducted without external research funding.

REFERENCES

[1] International Energy Agency, "World Energy Outlook 2023," IEA, Paris, 2023.
[2] J. Goh et al., "A Dataset to Support Research in the Design of Secure Water Treatment Systems," in Proc. CRITIS, 2016.
[3] B. Rossi et al., "Anomaly Detection in Smart Grid Data: An Experience Report," in Proc. IEEE SMC, 2016.
[4] G. Gao et al., "Machine Learning Schemes for Anomaly Detection in Solar Power Plants," Energies, vol. 15, no. 3, p. 1082, 2022.
[5] Z. Mao et al., "Research on Anomaly Detection Model for Power Consumption Data Based on Time-Series Reconstruction," Energies, vol. 17, no. 19, p. 4810, 2024.
[6] Z. Z. Alkaf, B. W. Lenggana, A. N. A. Yusuf, E. S. H. Nurdiniyah, and T. Wisudawati, "Improving Solar Energy Reliability with Data-Driven Anomaly Detection Techniques," Advances in Technology Innovation, 2026.
[7] S. H. Rafique, A. Abdallah, and N. S. Musa, "Machine Learning and Deep Learning Techniques for IoT Network Anomaly Detection," Sensors, vol. 24, no. 6, p. 1968, 2024.
[8] H. Nizam et al., "Real-Time Deep Anomaly Detection Framework for Multivariate Time-Series Data in Industrial IoT," IEEE Sensors J., vol. 22, no. 23, pp. 22836–22849, 2022.
[9] S. C. Ambat et al., "Anomaly Detection and Prediction of Energy Consumption for Smart Homes Using Machine Learning," ETRI J., 2025.
[10] A. Elhanashi et al., "Lightweight Signal Processing and Edge AI for Real-Time Anomaly Detection in IoT Sensor Networks," Sensors, vol. 25, no. 21, p. 6629, 2025.
[11] J. Lu et al., "Learning under Concept Drift: A Review," IEEE Trans. Knowledge Data Eng., vol. 31, no. 12, pp. 2346–2363, 2019.
[12] Y. Chen et al., "Addressing Concept Drift in IoT Anomaly Detection: Drift Detection, Interpretation, and Adaptation," IEEE Computer, 2024.
[13] R. Xu, Y. Cheng, Z. Liu, Y. Xie, and Y. Yang, "Improved Long Short-Term Memory Based Anomaly Detection with Concept Drift Adaptive Method for Supporting IoT Services," Future Generation Computer Systems, vol. 112, pp. 228–242, 2020, doi: 10.1016/j.future.2020.05.035.
[14] M. Ashrafuzzaman et al., "Detecting Stealthy False Data Injection Attacks in the Smart Grid Using Ensemble-Based Machine Learning," Computers & Security, 2020.
[15] A. Naqqad, A. Boulal, and R. Habachi, "An Ensemble Learning Framework for Cyber Attack and Fault Discrimination in Smart Grids," Energies, vol. 18, no. 23, p. 6305, Nov. 2025. doi: 10.3390/en18236305
[16] S. Pan, T. Morris, and U. Adhikari, "Classification of Disturbances and Cyber-Attacks in Power Systems Using Heterogeneous Time-Synchronized Data," IEEE Trans. Ind. Inform., vol. 11, no. 3, pp. 650–662, Jun. 2015. doi: 10.1109/TII.2015.2420951
[17] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: A Water Distribution Testbed for Research in the Design of Secure Cyber Physical Systems," in Proc. 3rd ACM Int. Workshop Cyber-Phys. Syst. Secur. Privacy (CPS-SPC), 2017, pp. 25–28.
[18] P. P. Biswas, H. C. Tan, Q. Zhu, Y. Li, D. Mashima, and B. Chen, "A Synthesized Dataset for Cybersecurity Study of IEC 61850 Based Substation," in Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm), 2019, pp. 1–7.
[19] A. Kannal, "Solar Power Generation Data," Kaggle Dataset, 2020. [Online]. Available: https://www.kaggle.com/datasets/anikannal/solar-power-generation-data
[20] C. Leys et al., "Detecting Outliers: Do Not Use Standard Deviation Around the Mean, Use Absolute Deviation Around the Median," J. Experimental Social Psychology, vol. 49, pp. 764–766, 2013.
[21] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," JAIR, vol. 16, pp. 321–357, 2002.
[22] M. J. C. S. Reis and C. Serôdio, "Edge AI for Real-Time Anomaly Detection in Smart Homes," Future Internet, vol. 17, no. 4, p. 179, Apr. 2025. doi: 10.3390/fi17040179
[23] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," NeurIPS, vol. 30, 2017.