

GRIDGUARD-AI: Scenario-Held-Out Cyberattack Detection with Explainable Assurance and Bounded Agentic Investigation

Stephen Kamran

kamran.msbf@gmail.com

Government College University, Lahore <https://orcid.org/0009-0000-3063-5611>

Farah Saman

University of Exeter <https://orcid.org/0000-0003-0933-4368>

Research Article

Keywords: smart-grid cybersecurity, Data4Cyber, temporal convolutional networks, SHAP, runtime assurance, AI agents, human-in-the-loop

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11447297/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

The increasing digitalization of smart grids exposes cyber-physical infrastructures to sophisticated false-data injection, communication spoofing, and supply-chain attacks, challenging the reliability of conventional intrusion detection under previously unseen operational conditions. Although machine-learning models can identify anomalous patterns, predictive discrimination alone does not establish whether automated decisions are safe, explainable, or operationally accountable. This study introduces GRIDGUARD-AI, an explainable, safety-assured agentic investigation framework integrating random forests (RF), temporal convolutional networks (TCN), and autoencoder (AE)-based anomaly detection with deterministic risk escalation, evidence-grounded AI investigations, and human-in-the-loop (HITL) governance. Using the Data4Cyber dataset, eight-fold leave-one-scenario-out evaluation examines 14,202 rolling windows containing 141 variables across eight operational scenarios. Results reveal substantial cross-scenario variability in detector discrimination, reinforcing the importance of evaluating generalization beyond familiar attack conditions. An exploratory three-way assurance policy classified 1,697 windows as SAFE, 501 as AUTO ALERT, and 12,004 as HUMAN REVIEW. Of 8,455 attack windows, 138 were incorrectly released SAFE (1.63%), yielding 98.37% attack escalation into automatic alert or human review rather than equivalent detection accuracy. Sequential AI-agent investigations completed 24/24 cases after quota-related retries, while automatic tool-calling workflows completed 8/8 scenario probes. Despite demonstrating traceable evidence handling and bounded investigation capabilities, the substantial human-review burden (84.52%), concentrated false-safe failures, and post-hoc policy refinement prevent claims of deployment-ready protection. GRIDGUARD-AI advances an auditable integration of scenario-held-out cyberattack detection, explainable risk assurance, and human-governed agentic decision support, providing an empirical foundation for future trustworthy smart-grid cybersecurity research.

I. INTRODUCTION

Modern distribution grids rely on interdependent supervisory communication, metering, storage, and distributed-energy interfaces. The resulting cyber-physical attack surface exposes operational telemetry to false-data injection, ARP spoofing, malware-inspired controller compromise, and supply-chain manipulation. Reviews of smart-grid attacks describe how corrupted measurements can disguise system states or degrade detection and operational decisions [1]–[3]. Such risks motivate a distinction between identifying anomalous windows and deciding what an operator may safely disregard. A confidently labelled SAFE state can be particularly consequential when the underlying detector has not encountered the attack configuration during training.

Recent machine-learning approaches combine supervised classifiers, temporal feature extraction, and anomaly reconstruction to recognize deviations in industrial and power-system data [4]–[6]. Yet their apparent discrimination can depend strongly on the operating scenario, class balance, measurement semantics, and separation of fitting from evaluation. A random train-test partition may expose related windows of the same attack episode to both training and testing. GRIDGUARD-AI therefore treats

scenario transfer as an explicit challenge: each named operational scenario is held out entirely, while model fitting and validation within the remaining scenarios use chronology-aware partitions and a purge around split boundaries.

Prediction confidence is not itself a safety argument. Anomaly thresholds learned from benign validation data may be poorly calibrated when attack mechanisms and background behavior shift; threshold fragility and alarm-budget research illustrate why both missed unsafe releases and operator workload deserve attention [7]–[11]. A useful assurance policy must expose not only automatic alerts but also the cases it declines to resolve. For this reason, the proposed evaluation separately reports false-SAFE attacks, benign SAFE release, and the volume routed to simulated human review. These outcomes should not be conflated with ROC-AUC, detector recall, or the number of independently mitigated incidents.

Tool-using language-model agents offer a complementary route to evidence retrieval and case summarization, but their explanatory fluency should not grant authority over critical infrastructure. Research on system-level assurance, agentic risk management, and human–AI security operations calls for explicit permission boundaries, inspectability, and human responsibility [12]–[16]. Accordingly, GRIDGUARD-AI places a deterministic safety gate outside the Gemini investigation layer; the latter may retrieve stored evidence, inspect available explanations, and recommend routing, but it neither changes the gate verdict nor sends commands to grid equipment.

This study asks: can heterogeneous detection evidence support transparent cyber-threat escalation under scenario shift while permitting bounded, auditable AI investigation? Its four contributions are: (i) a common eight-fold leave-one-scenario-out (LOSO) evaluation with chronological, purge-separated model development; (ii) comparative random forest (RF), temporal convolutional network (TCN), and autoencoder (AE) evidence with explicit undefined metrics on single-class folds; (iii) a three-way, percentile-based assurance prototype evaluated using attack false-SAFE, escalation, and review-workload measures; and (iv) executable evidence-retrieval and tool-calling traces with single-case TreeSHAP verification, without claims of live human intervention or autonomous grid protection.

II. RELATED WORK AND RESEARCH GAP

Cybersecurity research establishes a wide spectrum of measurement and operational-technology threats, including false-data injection, replay, and network compromise [1]–[3]. Hybrid autoencoder–random forest approaches [4] and deep sequential SCADA modeling [5] offer relevant detector baselines; the TCN literature supports temporal convolution as a general sequence-learning mechanism [6]. These contributions motivate complementary inductive biases, but performance within a studied corpus cannot automatically be interpreted as transfer to a previously unseen attack setup. GRIDGUARD-AI examines that distinction through complete scenario exclusion rather than assuming that combining detectors produces superior discrimination.

Recent evidence also emphasizes why data provenance and process semantics matter. Zhou et al. [17] document real-world industrial-control network attacks with fully labeled features, while Chahine and Noura [18] show that control-informed features and event-sensitive evaluation can reveal limits hidden by aggregate scores. Moradpour et al. [19] examine ensemble detection of false-data injection in power plants, illustrating a complementary approach to attack-specific variation. These studies reinforce the need to distinguish dataset realism, feature meaning, model ranking and risk-sensitive decisions; none establishes transferability of the present Data4Cyber policy.

Explainable machine learning has made attack-score attribution more inspectable. SHAP and domain-specific smart-meter XAI studies provide methods for ranking feature contributions, while research on explanation-guided power-system cyberattacks highlights that explanations themselves can be studied as attack surfaces [20]–[22]. However, feature contribution is not proof of causal physical behavior, nor does an explanation establish that a particular response is safe. Our implementation accordingly limits its empirical SHAP claim to one reconstructed RF case; it does not imply that every agent-reviewed event received a verified attribution.

A second research stream examines leakage-aware time-series evaluation, uncertainty calibration, fixed-threshold failures, and alarm-limited anomaly monitoring [7]–[11]. Its central lesson for cyber–physical decision support is that a single discrimination metric cannot characterize the consequences of operational routing. Even high escalation can be achieved by reviewing nearly every observation. We therefore report conservative escalation jointly with false-SAFE errors and the fraction of benign windows released. The resulting policy is explicitly exploratory because its final specification was revised after the held-out predictions had been inspected; no prospective calibration guarantee is claimed.

Finally, trustworthy-system literature distinguishes component reliability from whole-workflow assurance [12], [13], while agentic-AI taxonomies and human–AI security operations examine tool use, organizational accountability, and control limits [14]–[16]. These streams provide governance principles but do not themselves validate this particular grid dataset or its policy. The gap addressed here is their empirical connection: scenario-held-out detector outputs are converted into auditable SAFE, AUTO ALERT, and HUMAN REVIEW states, while tool-bounded agents produce evidence-based summaries without acquiring actuation authority. The contribution is a measured integration and its observed limitations, not a claim to invent its component algorithms or demonstrate operational readiness.

III. GRIDGUARD-AI METHODOLOGY

A. Data and leave-one-scenario-out design

Data4Cyber supplies 14,354 raw samples across eight evaluated scenario variants: benign baseline (S0), Industroyer PV (S1 and alternate), BSS (S2), ARP-spoofing variants (S3–S5) and MQTT supply-chain compromise (S6) [23]. Strict 20-step rolling construction produces 14,202 held-out windows with 141 channels. In each LOSO fold the scenario under evaluation is excluded entirely; fit and validation portions

of the remaining scenarios follow chronological order with a 19-step purge. These folds test named scenario shift, not transfer across independent physical grids.

The evaluation unit is a rolling-window observation, not an independent physical incident: adjacent windows can share measurements. Therefore window-level rates must not be interpreted as event-level detection performance or as a count of independently mitigated attacks. The 19-step purge limits temporal overlap across chronological fit and validation partitions, while scenario exclusion controls the identity of the held-out attack condition. Feature definitions and scenario labels are inherited from the dataset; no claim is made that anonymous channels map uniquely to causal power-system assets.

TABLE I

DATA4CYBER EVALUATION DESIGN

Quantity	Setting
Scenarios / test windows	8 / 14,202
Input window / variables	20 steps / 141
TCN ensemble	3 seeds $\times$ 8 folds
RF / AE	8 LOSO folds each
Data partitions	Chronological; purge = 19

B. Detection and assurance

The RF baseline has 120 trees, depth 18, balanced classes and minimum leaf size two. TCN models train with seeds 11, 42 and 77, early stopping and at most 25 epochs; AE reconstruction is fit on benign training cases with a 95th-percentile benign-validation flag. Using detector score p and fold-specific benign validation set V , the lower and upper score cutoffs are $Q_{0.05}(p|V)$ and $Q_{0.95}(p|V)$. Let a_{AE} be the binary anomaly flag:

Three modeling roles are deliberately separated. RF provides a supervised tabular decision score and an interpretable tree-based baseline. The TCN captures order-sensitive patterns in 20-step multichannel sequences and is trained repeatedly to expose seed sensitivity. The AE learns reconstruction structure from benign fit data, yielding a complementary deviation flag instead of a directly comparable probability. These outputs are not naively averaged because they need not be calibrated to equivalent scales, and the exploratory validation blend failed to improve the TCN benchmark.

$$A = (p_{\text{TCN}} \geq \tau_{\text{H,TCN}}) \wedge (a_{\text{AE}} = 1) \quad (1)$$

$$S = (p_{\text{TCN}} < \tau_{\text{L,TCN}}) \wedge (p_{\text{RF}} < \tau_{\text{L,RF}}) \wedge (a_{\text{AE}} = 0) \quad (2)$$

AUTO ALERT is assigned when $A = 1$; otherwise SAFE when $S = 1$; all remaining windows require HUMAN REVIEW. This deterministic gate does not allow Gemini agents to override routing. The final 5th/95th-percentile policy was revised after inspecting test outcomes; every policy result below is exploratory, not a prospective unbiased generalization estimate.

The validation quantiles represent operational score thresholds rather than guaranteed confidence intervals. On any held-out scenario, the resulting three-way policy may abstain extensively and still release a subset of attacks. This design intentionally exposes a trade-off between missed threats, automatic alerts and human triage. It does not optimize an explicit intervention cost, and an AUTO ALERT is merely an offline route label, not a physical protective action.

C. Explanation and bounded agents

A sequential Gemini pipeline performs evidence retrieval, explanation availability checks and risk summarization before deterministic HITL routing. Separate automatic function-calling probes invoke `retrieve_evidence`, `retrieve_xai` and `route_to_hitl`. RF TreeSHAP was recomputed and score-matched for one case (248); per-case SHAP was not available for the other agent cases. The review queue is simulated: no human analyst response, live attack interruption, agent privilege escalation or grid actuation is claimed [22], [14]–[16].

For transparency, the final RF comparator was refitted from training-only partitions after an earlier leakage-prone prototype. Naive RF–TCN mean blending reduced validation macro-AUC (0.490 versus TCN 0.660; RF 0.483), so ensemble superiority is not claimed. These are validation aggregates and must not be confused with scenario-held-out test AUCs. Policy performance is analyzed separately from discrimination.

IV. EXPERIMENTAL RESULTS

A. Cross-scenario discrimination

The strict RF achieved higher test ROC-AUC than the TCN on all six evaluable mixed-class scenarios, but performance was unstable (Table II; Fig. 2a), falling to 0.346 on S1 PV. ROC-AUC is undefined for benign-only S0 and attack-only S3. This is a warning against selecting algorithms solely on pooled results.

Because ROC-AUC measures ranking across threshold choices, it cannot reveal the rate of unsafe SAFE decisions generated by a specific routing policy. Conversely, a conservative gate can attain high escalation by reviewing most cases even when classifier ranking is weak. GRIDGUARD-AI therefore reports scenario-discrimination metrics alongside denominator-explicit policy outcomes, avoiding the common but misleading substitution of escalation for attack recall.

The strong S6 MQTT AUC contrasts sharply with the S1 Industroyer-PV fold, where both classifiers rank attacks weakly. This distributional variability is substantively important: the selected detector and fixed

operating threshold may perform differently on protocols, attack timing and component configurations absent from training. Hence a single pooled AUC would mask exactly the transfer risk probed by LOSO.

TABLE II

SCENARIO-HELD-OUT DISCRIMINATION

Holdout	RF AUC	TCN AUC	AE rec.
S1 PV	0.346	0.162	30.0%
S1 PV alt	0.807	0.756	10.3%
S2 BSS	0.727	0.673	9.3%
S4 ARP/PV	0.839	0.517	90.5%
S5 ARP/PV/BSS	0.886	0.815	53.1%
S6 MQTT	0.953	0.812	100%
S0 benign	N/A	N/A	N/A (no attacks)
S3 ARP/BSS	N/A	N/A	0.06%

AE attack recall was 30.0% (S1 PV), 10.3% (S1 alternate), 9.3% (S2), 0.06% (S3), 90.5% (S4), 53.1% (S5), and 100% (S6); the last fold simultaneously marked all available benign examples anomalous. Thus AE complements, but cannot independently validate, an attack alert.

AE false alarms likewise varied: approximately 0.90% on benign S0, 21.48% on S1 PV, 2.67% on S1 alternate, 1.54% on S2, 2.49% on S4 and 1.14% on S5; S3 contains no benign examples, making its benign false-alarm rate undefined. These outcomes reinforce why a reconstruction score should be treated as corroborating evidence, not an oracle.

B. Three-way assurance and failure modes

Across 14,202 held-out windows (8,455 attacks; 5,747 benign), the exploratory gate produced 1,697 SAFE (11.95%), 501 AUTO ALERT (3.53%) and 12,004 HUMAN REVIEW (84.52%). Among attacks, 138 were incorrectly released SAFE (1.63%), entirely in the S3 ARP/BSS fold; the remaining 98.37% entered alert or review. Benign SAFE release was 27.13%. Crucially, escalation includes human review and is not automated detection sensitivity.

Under the three-way policy, the attack false-SAFE rate is $FSR = N(\text{attack} \cap \text{SAFE})/N(\text{attack}) = 138/8,455 = 1.63\%$, with attack escalation $ER = 1 - FSR = 98.37\%$. This identity follows from the policy definition that

anything not SAFE is escalated, and therefore cannot demonstrate model classification precision or independent detection recall.

TABLE III
EXPLORATORY ASSURANCE OUTCOMES

Outcome / metric	Result
SAFE / AUTO ALERT	1,697 / 501
HUMAN REVIEW	12,004 (84.52%)
Attack false-SAFE	138 / 8,455 (1.63%)
Attack escalation	98.37%
Benign SAFE release	27.13%

Review or alert accounted for 88.05% of all evaluated windows, including many benign cases. Although 98.37% of attacks were escalated, an 84.52% overall review workload would be difficult to operationalize without triage budgets, case deduplication, and cost-sensitive threshold selection [11]. The concentrated S3 false-SAFE failure also shows that conservatism does not eliminate blind spots.

C. Bounded agent execution and auditability

With Gemini free-tier quota retries, the staged investigation pipeline completed 24/24 cases. Direct automatic function calling completed 8/8 scenario probes. These figures describe software execution, not classification accuracy, incident-response correctness, reviewer agreement, or autonomous security effectiveness. For case 248, a retrained LOSO RF matched the stored probability and a genuine TreeSHAP explanation identified leading anonymous predictors (feature_139, feature_136 and feature_52). Physical feature meaning was not independently established.

The simulated agent tests evaluate two distinct configurations. First, sequential role-specific prompting tests whether evidence, explanation limits, risk synthesis and routing can be assembled across 24 preselected cases. Second, eight one-case-per-scenario probes test automatic function calls with a fixed tool inventory. Passing either test means the programmed steps executed and produced outputs. Neither experiment measures analyst satisfaction, precision of forensic assertions, resistance to malicious tool responses, or response latency under operational load.

The tool-based validation supplied structured scenario evidence to the language model and exposed narrowly typed functions; tool outputs supported case summaries while the predetermined assurance state controlled the action recommendation. The local TreeSHAP demonstration found positive RF attack-class contributions for feature_139 (0.06469), feature_136 (0.05673) and feature_52 (0.05068).

Their anonymous names prohibit substantive attribution to a physical device without a verified feature map.

TABLE IV

AGENT EXECUTION AND AUDIT

Agent evaluation	Observed
Sequential workflow	24/24 after retries
Tool-driven scenario probes	8/8 successful
RF TreeSHAP	1 reconstructed case
Grid actuation / human trials	None / none

V. DISCUSSION AND LIMITATIONS

The substantive contribution is separation of statistical discrimination, uncertainty-sensitive routing, explainable evidence and bounded agent permissions. The review-heavy gate reduces immediate SAFE releases, yet makes no calibrated safety guarantee. Test-informed revision of score thresholds makes its 1.63% false-SAFE result descriptive rather than confirmatory. No field operators or live grid devices were evaluated; agent prompt misuse, tool-output injection, and physical consequence of alerts remain untested.

The findings demonstrate that detector diversity does not necessarily ensure superior cross-scenario generalization. Although TCN captures temporal dependencies and AE identifies reconstruction anomalies, RF achieved higher ROC-AUC across all evaluable mixed-class held-out scenarios, while naive RF–TCN averaging underperformed TCN alone during validation. This highlights the importance of treating detector disagreement as evidence of uncertainty rather than assuming ensemble superiority. GRIDGUARD-AI therefore prioritizes transparent, risk-sensitive escalation over predictive agreement, retaining HUMAN REVIEW when available evidence cannot justify SAFE release or automatic alerting.

However, conservative escalation introduces substantial operational costs. The framework routed 12,004 of 14,202 overlapping windows (84.52%) to simulated human review, revealing a significant potential for alert fatigue without representing 12,004 independent incidents. Future research should investigate event-level aggregation, cost-sensitive threshold calibration, risk-prioritized triage, and operator-in-the-loop evaluation, including detection latency, false-SAFE decisions, and review efficiency. These remain prospective research priorities rather than demonstrated operational benefits.

For Pakistan, the World Bank BEST-PAK Grid Stability Enhancement Project and Asian Development Bank transmission-strengthening proposal contextualize efforts to modernize grid infrastructure [24], [25].

Increasing digital interconnection makes independently verified telemetry assurance and human-governed cyber triage relevant research priorities; these reports do not evidence GRIDGUARD-AI deployment or effectiveness in Pakistan.

The practical application pathway would require mapping SCADA and distribution-system telemetry onto consistent, access-controlled features; documenting benign operating regimes; and designing escalation budgets around local operators' staffing and response obligations. These are proposed deployment requirements rather than outcomes demonstrated by the present experiments.

LOSO transfers across predefined scenario traces, not substations or utilities. Benign-quantile calibration under distribution shift lacks conformal or other coverage guarantees [7]–[10]. The agent workflow comprises sequential role prompts and tool calls, not an independently collaborating multi-agent society. Prior to any operational use, future studies need a sealed test set, fixed prospective thresholds, external grid datasets, instrumented operator trials, realistic attack costs, and audited tool permissions.

Even successful repeated tool calls do not prove correct cyber-forensic reasoning, justified response selection or resistance to adversarial inputs. A prospective study should score agents against gold-standard incident facts and explicit error severities, independently of the deterministic policy gate. An operator-in-the-loop trial would additionally measure triage time, alert fatigue and overrides before any deployment claim.

Finally, auditability requires stronger tests than an execution-success fraction. Future adversarial exercises should evaluate maliciously altered evidence, prompt injection in retrieved logs, tool-response inconsistencies, unavailable SHAP explanations, API quota interruptions, and disagreement between risk narratives and deterministic gate decisions. Each case should retain immutable detector scores, validation thresholds, provenance for retrieved artifacts, time-stamped tool calls and final human dispositions. Controlled comparisons against a non-agent evidence dashboard would help determine whether the language-model layer improves analyst decision quality rather than merely producing longer summaries. These safeguards would preserve a narrow, reviewable interface between probabilistic models and safety-sensitive workflows.

VI. CONCLUSION

GRIDGUARD-AI demonstrates an inspectable connection between scenario-held-out cyberattack detection, deterministic assurance, explainable evidence and permission-bounded AI investigation. In eight Data4Cyber folds comprising 14,202 test windows, strict RF and TCN rankings varied substantially across attack configurations, while autoencoder recall was inconsistent. These findings show why detector diversity cannot by itself be treated as verified cross-scenario resilience. The exploratory three-way policy released 138 of 8,455 attack windows as SAFE (1.63%) and escalated the remaining 98.37% to an alert or review pathway. However, 12,004 of all 14,202 windows (84.52%) required simulated HUMAN REVIEW, and all false-SAFE attacks were concentrated in the S3 spoofing scenario. Escalation is therefore a routing outcome—not confirmed detection recall or incident mitigation. Bounded Gemini

workflows completed 24/24 retried case executions and 8/8 tool-calling probes; an RF TreeSHAP reconstruction provided one auditable case explanation. These results establish technical feasibility of offline evidence routing, not reliable forensic judgment, live analyst assistance, or autonomous grid protection. The practical contribution is a measurable safety–workload trade-off with explicit separation of model evidence, agent recommendations and final routing authority. For deployment-oriented research, the priorities are preregistered thresholds and untouched external test sets, reduced operator workload through event-level triage, adversarial testing of retrieved evidence and tool permissions, and operator-in-the-loop assessment. Such validation would also be essential before considering transfer to modernizing power systems in Pakistan or other emerging markets.

Declarations

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the research reported in this manuscript.

FUNDING

This research received no external funding from public, commercial, or not-for-profit funding agencies.

DATA AND CODE AVAILABILITY

This study utilizes the Data4Cyber dataset, developed and released by its original creators and cited in the manuscript (Reference [23]). The dataset is attributed to its original source, and no claim of ownership is made by the authors. The study's experimental methodology, model configurations, evaluation procedures, and reported results are described in the manuscript. Additional experimental artifacts, analysis scripts, and reproducibility materials may be shared through a dedicated research repository, subject to verification, applicable licensing conditions, and disclosure requirements.

Dataset DOI: <https://doi.org/10.5281/zenodo.20004612>

References

1. Ishfaq H et al (2026) A comprehensive survey of false data injection attacks modeling and ML/DL-based detection in smart grids. *ICT Express*. 10.1016/j.ictex.2026.07.027
2. Caleb TD, Shao S, Kaabouch N (2026) False data injection attacks on smart grids: Attack models, challenges and future directions, *Int. J. Inf. Secur.*, vol. 25, Art. no. 94. 10.1007/s10207-026-01262-w
3. Tadese YB et al (2026) Cyber-physical microgrids under false data injection, denial-of-service, and replay attacks: A systematic review. *IEEE Access* 14:147521–147550.

10.1109/ACCESS.2026.3735264

4. Albarrak AM et al (2026) A hybrid deep autoencoders and random forest framework for false data injection attack detection in industrial Internet of Things networks, *Sensors*, vol. 26, Art. no. 5110. 10.3390/s26165110
5. Yilmaz MT et al (2026) Deep tabular and sequential models for robust SCADA cybersecurity using the custom-built SCADANet dataset. *IEEE Access* 14:77736–77755. 10.1109/ACCESS.2026.3691986
6. Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*. 10.48550/arXiv.1803.01271
7. İnan E (2026) Proximity-based explainable anomaly detection for time-series data with calibration, leakage safety, *Internet Things*, vol. 37, Art. no. 101920. 10.1016/j.iot.2026.101920
8. Yuan S et al (2026) Conformal machine learning for reliable anomaly detection in industrial cyber-physical systems, *Reliab. Eng. Syst. Saf.*, vol. 274, Art. no. 112417. 10.1016/j.ress.2026.112417
9. Dabthong H et al (2026) When fixed thresholds fail: A negative result study of reconstruction-based GNN anomaly detection on SWaT, in *Proc. IEEE ICCI*. 10.1109/ICCI68752.2026.11506458
10. Sreejith A et al (2026) Uncertainty-aware anomaly prediction using conformal multivariate time series clustering. *IEEE Trans Ind Cyber-Phys Syst* 4:280–284. 10.1109/TICPS.2026.3686147
11. Heydari V, Nyarko K (2026) Alarm-budgeted event-level evaluation of ICS anomaly detection: Lessons from SWaT and WADI. *IEEE Access* 14:17585–17609. 10.1109/ACCESS.2026.3659034
12. Hämmäläinen A et al (2026) Building trustworthy AI systems for condition monitoring. *IEEE Access* 14:67023–67037. 10.1109/ACCESS.2026.3688909
13. Huang Y (2026) Towards trustworthy AI-enabled systems: From model-level reliability to systemic assurance, in *Proc. ACM FSE Companion*, pp. 17–20. 10.1145/3803437.3804867
14. Sapkota R et al (2026) AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges, *Inf. Fusion*, vol. 126, Art. no. 103599. 10.1016/j.inffus.2025.103599
15. Raza S et al (2026) TRiSM for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems. *AI Open* 7:71–95. 10.1016/j.aiopen.2026.02.006
16. Kumar M et al (2026) Adaptive trust-aware SOC human–AI teaming for resilient operations, *Comput. Secur.*, vol. 169, Art. no. 104971. 10.1016/j.cose.2026.104971
17. Zhou X, Cheng Z, Wang C, Wang S, Tao C, Zhou Z, Chen X, Luo J, Wang D, Zhou H (2026) A dataset collected in real-world industrial control systems for network attack detection, *Sci. Data*, vol. 13, Art. no. 399. 10.1038/s41597-026-06738-x
18. Chahine K, Noura HN (2026) Features over architecture: Physics-informed anomaly detection in industrial control systems, *Future Internet*, vol. 18, no. 6, Art. no. 308. 10.3390/fi18060308
19. Moradpour A, Delkhosh H, Seifi H (2026) Detecting false data injection attacks on power plants based on voting ensemble learning utilizing data diversification, *Eng. Appl. Artif. Intell.*, vol. 164, Art. no. 113302. 10.1016/j.engappai.2025.113302

20. Zhang B et al (2026) Interpretable smart meter anomaly detection based on Bayesian-optimized XGBoost and SHAP, *Information*, vol. 17, Art. no. 810. 10.3390/info17080810
21. Paphitis G et al (2026) Explainability-guided counterfactual cyberattacks through surrogate modeling in power systems, *Sustain. Energy Grids Netw.*, vol. 47, Art. no. 102500. 10.1016/j.segan.2026.102500
22. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. in *Adv Neural Inf Process Syst*, 30
23. Sen Ö et al (2026) Data4Cyber: A labeled cyber-physical dataset for distribution-grid management systems under representative OT attacks, *Zenodo*, ver. 2. 10.5281/zenodo.20004612
24. World Bank, Appraisal Environmental and Social Review Summary BEST-PAK MPA Phase 1—Grid Stability Enhancement Project (P504536), Rep. ESRSA04732, Jan. 2026. [Online]. Available: <https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099012726092014169>
25. Asian Development Bank, Second Power Transmission Strengthening Project Report and Recommendation of the President, Project No. 59002-001, Oct. 2025. [Online]. Available: <https://www.adb.org/projects/documents/pak-59002-001-rrp>

Figures

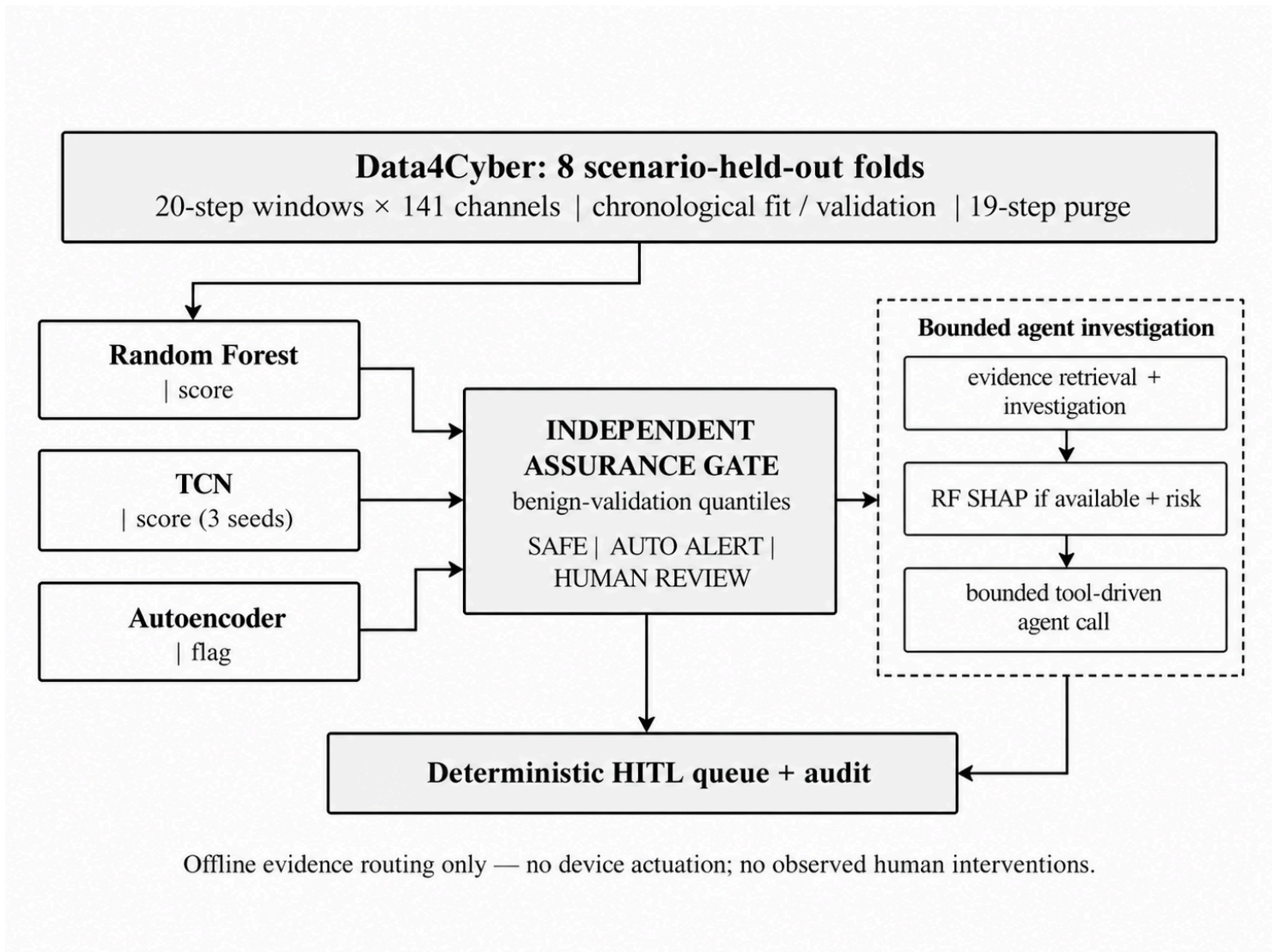


Figure 1

GRIDGUARD-AI architecture: detector evidence, independent assurance, bounded investigation, and simulated human-in-the-loop (HITL) routing.

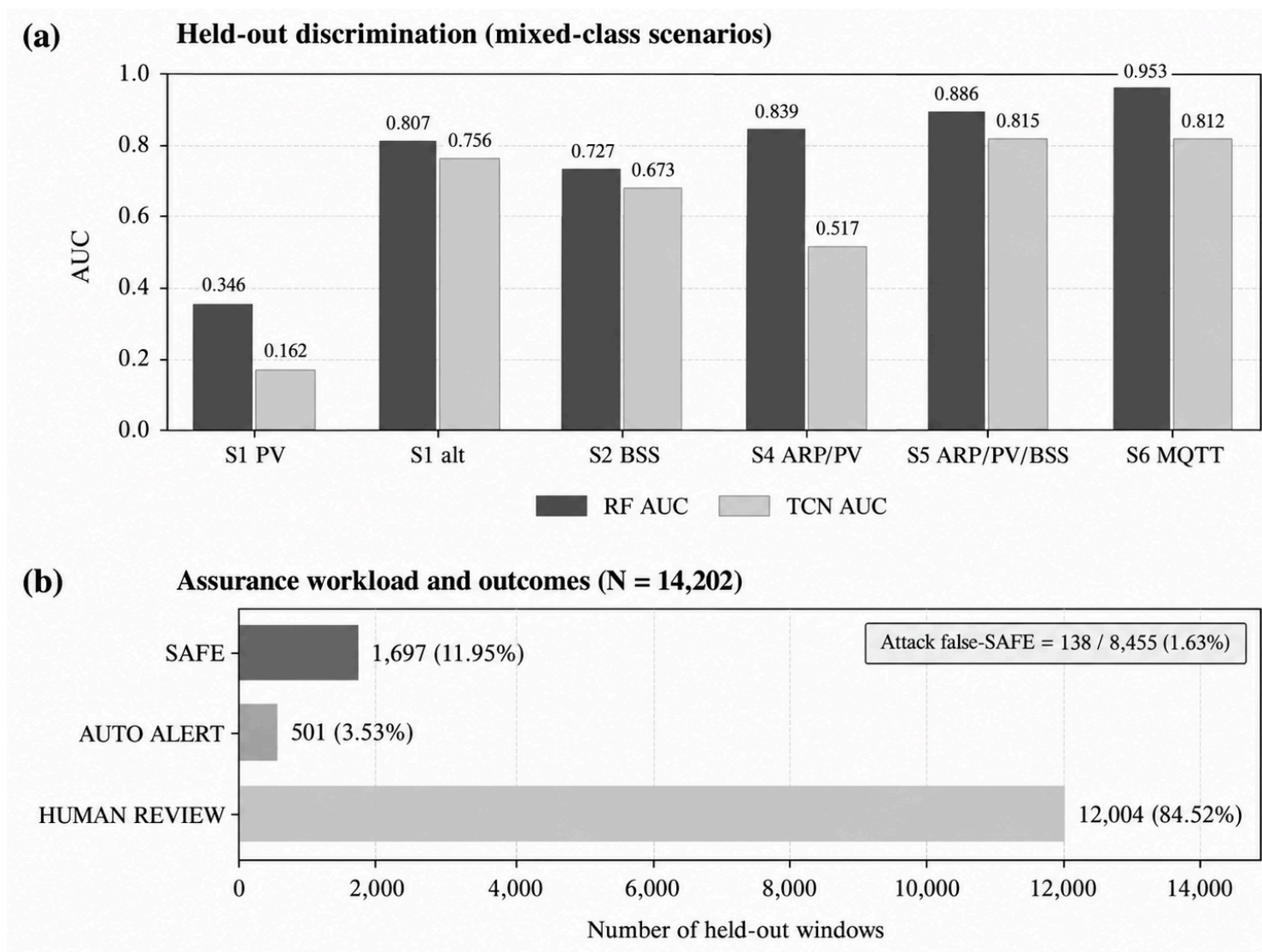


Figure 2

Held-out RF/TCN discrimination (a) and exploratory three-way assurance workload (b); S3 false-SAFE errors are reported separately.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [GRIDGUARD AISupplementaryFileS1.docx](#)