Robust Machine Learning Models for Cybersecurity in Critical Infrastructures: A Systematic Review, Empirical Synthesis, and Framework Proposal

Stephen Kofi Dotse
Department of Information Technology
University of Professional Studies, Accra

Samuel Yao Sebuabe
Department of Computer Science
Valley View University

Is-haq Kofi Mohammed
Department of Computing and Information Systems
Wisconsin International University College, Accra, Ghana

ABSTRACT
ML-based intrusion detection systems for critical infrastructure work well in controlled research settings. The gap between that and reliable operation against adversaries who probe, adapt, and know the detection layer is where this paper focuses. We reviewed 44 primary sources via a PRISMA-adapted protocol covering ensemble learning, GNNs, transformer architectures, deep reinforcement learning, adversarial defences, concept drift adaptation, federated learning, and XAI across three CI sectors: energy and water, healthcare IoMT, and intelligent transportation. The main findings: ensemble hybrids are the most deployable near-term architecture. Certified robustness methods are theoretically principled but fall short of operational security guarantees under realistic CI threat models, as Cullen et al. (2025) demonstrated at ICML. SHAP and LIME measurably improve analyst trust but enable adversaries to reconstruct model decision boundaries with over 90% success; this paper calls that tension the Adversarial XAI Paradox, and no reviewed study resolves it. Transportation sector evaluation is the weakest of the three, with no public V2X-specific benchmark comparable to SWaT or CICIoMT2024. We identify five testable research gaps, construct a five-layer framework aligned to NIST CSF 2.0, and close with specific recommendations.

General Terms
Security, Machine Learning, Algorithms, Performance.

Keywords
Critical infrastructure cybersecurity, machine learning robustness, adversarial ML, explainable AI, concept drift, intrusion detection, graph neural networks, certified robustness, NIST CSF 2.0

1. INTRODUCTION
The Colonial Pipeline attack in 2021 is a useful benchmark for how the threat to critical infrastructure has changed. Attackers hit the billing system. They never touched an ICS. The operator shut the pipeline down voluntarily because, with the corporate network compromised, they had no confident picture of what else might be reachable. No Stuxnet-level knowledge of industrial protocols was required. The attacker just needed access.

That shift matters for how we think about ML-based intrusion detection in CI. The field has spent most of the past decade demonstrating that ML can detect attacks on benchmark datasets. That case is made. What it has not spent nearly enough time on is whether those systems hold up when the adversary specifically knows ML is the detection layer, can probe it, and in some cases understands how it makes decisions. Signature-based rules cannot keep pace with novel attacks. That argument for ML is still valid. But the operational bar has moved.

The evidence reviewed here points to four specific places where current ML-based IDS fail. They are adversarially fragile: inputs crafted to exploit decision boundaries can deceive them, and even formally certified models provide weaker operational guarantees than their certificates suggest (Cullen et al., 2025; Sitawarin, 2024). They degrade silently when network behaviour, device populations, and attacker tactics shift over time (Yang & Shami, 2023; Sebopelo, 2026). Most published performance results come from generic IT network datasets that look nothing like the OT protocols and physics-constrained traffic of real CI environments (Zhevnenko et al., 2026). And the explainability methods designed to help operators understand model decisions also help adversaries reverse-engineer them (Khan et al., 2025).

This paper works through those four problems. We reviewed 44 sources, identified five testable gaps, and propose a five-layer framework for robust CI cybersecurity ML mapped to NIST CSF 2.0. Section 2 covers the threat landscape and related work. Section 3 describes the methodology and framework. Section 4 presents result by research question. Section 5 concludes.

2. BACKGROUND AND RELATED WORK

2.1 WHAT THE ATTACKS ACTUALLY LOOK LIKE
Healthcare had the highest mean breach cost of any sector in 2024: USD 9.77 million, the fourteenth consecutive year at the top (IBM Security, 2024). Sophos (2024) found that 67% of healthcare organisations were hit by ransomware. The February 2024 Change Healthcare breach compromised 190 million patients' records and cost UnitedHealth Group over USD 2.4 billion. The entry point was a Citrix portal with no MFA. No ICS expertise required. The Verizon DBIR (2024) counted 30,458 incidents, with vulnerability exploitation as an initial vector up roughly 180% from the year before.

Energy and transportation follow the same trajectory with different specifics. ENISA (2024) logged transportation as Europe's second most attacked sector, with hacktivist DDoS accounting for 79% of incidents the following year. Nation-state actors have been documented maintaining persistent OT access at US water and power utilities throughout 2024 without detection. The Cisco Systems (2025) readiness survey found

only 4% of 8,000 surveyed organisations globally had reached a Mature security posture; the AI Fortification pillar scored worst of all dimensions at 7% mature. Palo Alto Networks (2024) forecast that generative AI would compress attack timelines further, a development already visible in the 2025 threat data.

These numbers set the baseline for evaluating any detection improvement claim. A system achieving 99% accuracy on a benchmark dataset is less useful than it looks if the baseline operational detection rate is far lower and the failure modes are adversarial.

2.2 ML ARCHITECTURES FOR CI IDS

Random forests and gradient-boosted ensembles have dominated tabular network traffic benchmarks since Buczak and Guven (2016) surveyed the field. Their finding that ensembles outperform single classifiers on tabular IDS data has held up; on CICIDS2017 and NSL-KDD, XGBoost still exceeds 99% accuracy once class imbalance is addressed (Kasprzyk & Rychlicki, 2025).

GNNs are built for the topological structure of OT communication graphs, where deviations from expected PLC-to-sensor-to-historian paths are often the earliest sign of compromise. Gamage et al. (2025) reviewed 36 studies and found GNNs consistently outperform per-flow classifiers on lateral movement and reconnaissance, specifically because they model inter-node relationships rather than individual flows. Transformer architectures (Vaswani et al., 2017) capture long-range temporal patterns in alert sequences and are better suited to multi-stage APTs that unfold across hours than CNNs or LSTMs, which lose signal at longer horizons. Yang and Shami (2023) combined online Transformer-based pattern recognition with drift-adaptive ensemble learning in the MSANA framework, reaching 99.32% accuracy at below 4 ms on IIoT data, which matters for the real-time constraints of ICS monitoring. Deep reinforcement learning goes further still: DRL agents like those studied by Lutjens et al. (2020) can select defensive responses in real time with certified bounds on policy correctness under adversarial sensor manipulation, not just classify traffic.

For deployment today, ensemble methods are still the practical choice. Alharbi et al. (2025) reached 99.92% precision on CICIoMT2024 healthcare traffic with RF+AdaBoost+Bagging, beating heavier alternatives within edge medical device computational budgets. Arslan et al. (2026) showed HCABS-NID achieving 98.20% accuracy at 3.40 ms on V2X transportation data. Muzibuddin et al. (2026) showed that the USAD dual-decoder architecture reaches AUC 0.93 on SWaT without any labelled attack data, a practical advantage in OT environments where labelled incidents are scarce. The GNN and Transformer results are promising; the ensemble results are in production-adjacent territory.

2.3 ADVERSARIAL ROBUSTNESS: THE REAL PICTURE

Sitawarin (2024) puts it plainly: neural networks are the weakest link in the modern software stack under adversarial conditions, and this is a geometric property of high-dimensional learned boundaries, not an implementation problem with an architectural fix. FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), and C&W (Carlini & Wagner, 2017) are the canonical attack families; adversarial variants calibrated to network feature domains are now documented for IDS specifically (Khan et al., 2025).

The certified robustness literature is the principled response. Pfrommer et al. (2023a) showed that projecting randomised smoothing onto the data manifold yields certified volumes orders of magnitude larger than vanilla smoothing, which is more practical for CI sensor data that lives on a low-dimensional physical manifold. Pfrommer et al. (2023b) introduced asymmetric certification via feature-convex networks, enforcing robustness specifically against false-negative adversaries and directly addressing the cost asymmetry of CI detection: a missed attack is catastrophic, a false alarm is manageable. These are real contributions. But Cullen et al. (2025) demonstrated at ICML that certified classifiers can be forced to abstain within their certified radius, which is a denial-of-service effect that does not technically violate the certificate but defeats the purpose of detection. ℓ_p-norm certification bounds do not match the protocol-valid, semantically constrained manipulations available to real CI attackers. Any publicly known certification scheme gives adaptive adversaries a map of its gaps. Certification is useful. It is not sufficient on its own.

Concept drift compounds matter further. Unadapted models degrade to below-baseline performance within hours of drift onset on IIoT data (Yang & Shami, 2023). Sebopelo (2026) identified a specific ADWIN pathology called sensitivity oscillation and resolved it with Trinity-Controller ADWIN, sustaining post-drift accuracy above 90% with near-zero recovery delay. Zhevnenko et al. (2026) provide the most important aggregate finding: across 14 models and seven industrial datasets under five stress conditions, no single architecture dominates all of them, and graph-structured models are the most robust when sensors drop out.

2.4 XAI: USEFUL AND DANGEROUS IN EQUAL MEASURE

The case for XAI in CI IDS is well supported. Oriaro and Mishra (2025) synthesised 169 studies and found 40% higher analyst trust and 30% lower false-positive investigation time from SHAP and attention-based integration. Sunkara (2025) measured a 4.3/5 trust rating for XGBoost with SHAP in an actual SOC evaluation, with analysts describing the system as functioning like a pre-screening stage for alerts. Prasad et al. (2026) found that explanation format matters in practice: framing outputs in terms of MITRE ATT&CK TTPs was substantially more useful to operators than raw Shapley values, because operators already understood the attack taxonomy.

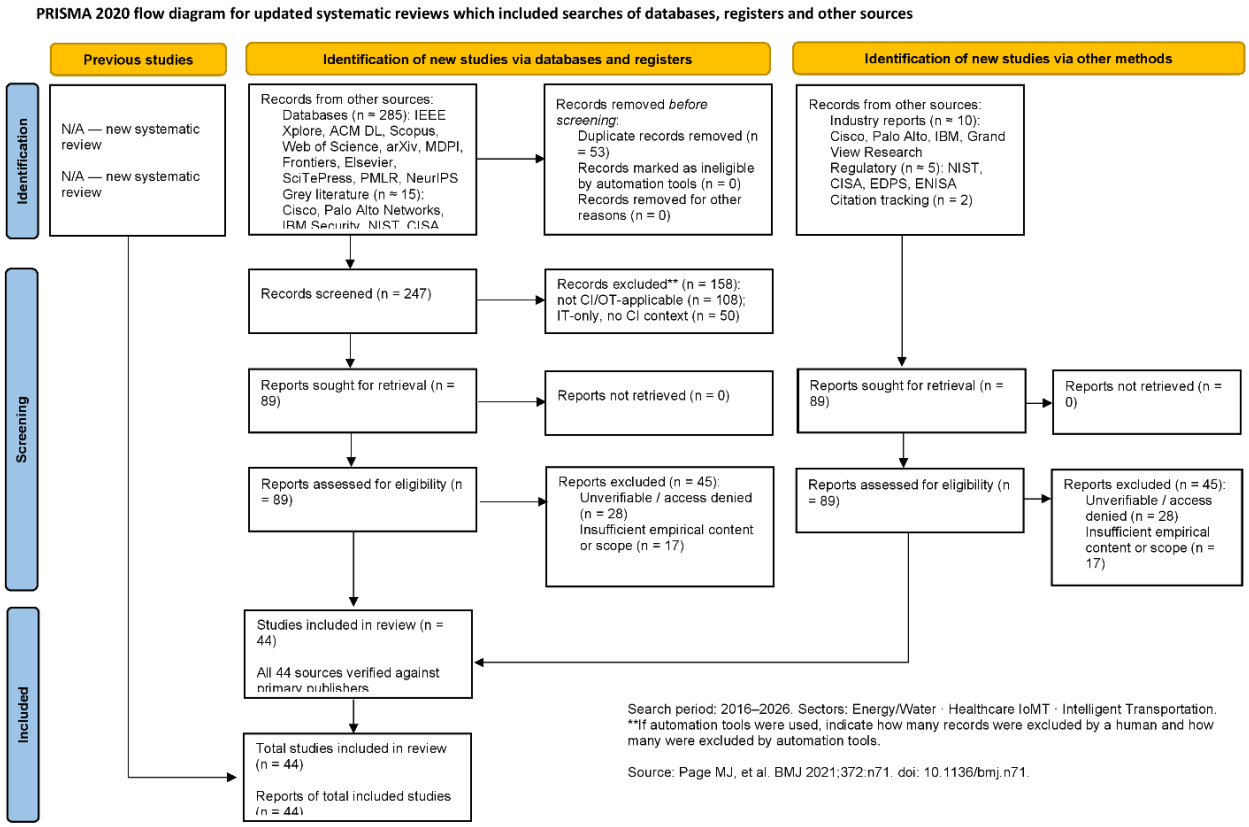
Khan et al. (2025) found the problem in a review of 135 XAI-IDS studies. SHAP and LIME outputs, when queryable without restriction, can reconstruct model decision boundaries with over 90% success. This is not a theoretical concern. It describes a live exploit pathway against currently deployed systems. Deploying XAI-enabled IDS without access controls on explanation endpoints is a net security regression: the transparency that helps operators helps adversaries equally. Batishchev and Saad (2025) reviewed CI incident databases and AI governance documents and concluded that current XAI techniques are not robust enough for safety-critical deployment. Neither paper offers a technical resolution.

3. METHODOLOGY

3.1 LITERATURE REVIEW

We followed PRISMA 2020 (Page et al., 2021). Databases searched: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, arXiv, and publisher platforms for MDPI, Frontiers, Elsevier, SciTePress, PMLR, and NeurIPS proceedings, plus industry and regulatory sources including Cisco, Palo Alto

Networks, IBM Security, NIST, CISA, EDPS, and ENISA. The date range is 2016 to 2026.

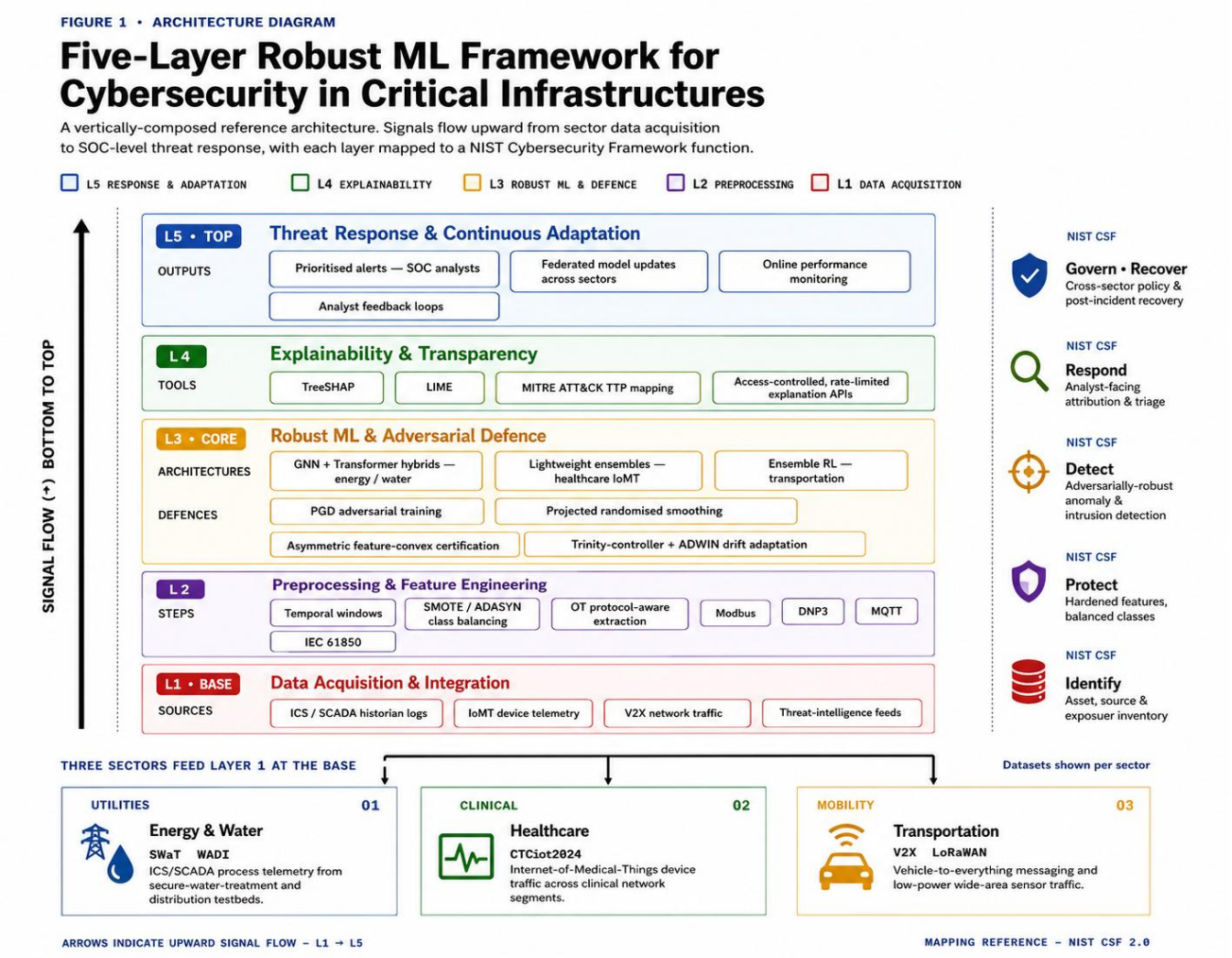


Search terms ran across three Boolean clusters: a problem-domain cluster targeting CI cybersecurity and ICS/OT security; a technical-approach cluster covering ML methods, adversarial robustness, XAI, and concept drift; and a sector-specificity cluster covering energy, healthcare IoMT, and transportation. Studies were included if they addressed ML or DL-based detection in CI or ICS/OT environments, provided empirical results or systematic synthesis, and covered at least one research theme. Studies were excluded if they covered only conventional IT security without CI applicability, were unverifiable against publisher databases, or were duplicates. One source attributed to a university blog could not be located in any publisher database and was excluded; all 44 retained sources were confirmed against primary publishers before citation.

3.2 FRAMEWORK DESIGN

The five-layer framework is grounded in the evidence: every design choice maps to at least one verified source, and the whole is aligned to NIST CSF 2.0 (NIST, 2024). The five gaps drive the design directly. Gap 1, the adversarial-ecological validity mismatch, determines the sector-calibrated architecture selection and adversarial defence layer. Gap 2, the Adversarial XAI Paradox, determines the access-controlled explanation design in Layer 4. Gap 3, the certified robustness and drift incompatibility, determines the layered defence strategy, combining empirical, certified, and adaptive mechanisms rather than depending on any single one. Gap 4, the absence of evaluated cross-sector federated IDS, determines the federated update protocol in Layer 5. Gap 5, the evaluation standardisation problem, determines the explicit assessment criteria.

Layer 1 aggregates CI telemetry with protocol-aware parsing for OT-specific formats (Modbus, DNP3, IEC 61850, MQTT) alongside conventional IP flow features. The ecological validity argument justifies this: a system extracting only generic flow statistics from OT traffic discards the protocol-level signal most relevant to ICS-specific attacks. Layer 2 handles preprocessing: temporal windowing, SMOTE/ADASYN class balancing, sector-specific feature extraction. Layer 3 is the detection engine. Architecture selection is sector-calibrated based on the Chapter 2 evidence: GNN+Transformer hybrids for energy and water, where both communication topology and temporal sequence carry attack signal; lightweight hard-voting ensembles for healthcare IoMT, given edge constraints and alert-fatigue priority; ensemble plus RL for transportation, where classification and adaptive response are both needed. Adversarial defences are layered deliberately. PGD adversarial training provides empirical broad-spectrum robustness. Projected randomised smoothing provides certified L2 robustness on



manifold-constrained inputs. Asymmetric feature-convex certification provides provable false-negative protection. No single mechanism addresses all attack vectors. Trinity-Controller ADWIN or DD-FS handles concept drift.

Layer 4 generates SHAP/TreeSHAP explanations mapped to MITRE ATT&CK TTPs. Explanation endpoints are rate-limited, authentication-gated, and monitored for systematic probing patterns. This is an operational mitigation for the Adversarial XAI Paradox. It is not a technical resolution. Layer 5 handles alert dispatch, federated model updates with Byzantine-tolerant aggregation, and continuous drift monitoring.

Figure 3.3 maps the threats the framework addresses. Evasion attacks are defended by Layers 2 and 3. Poisoning through federated learning is defended by Byzantine-tolerant aggregation in Layer 5. Concept drift is handled by Trinity-Controller ADWIN in Layer 3. Layer 4 access controls mitigate the Adversarial XAI threat; they do not solve it.

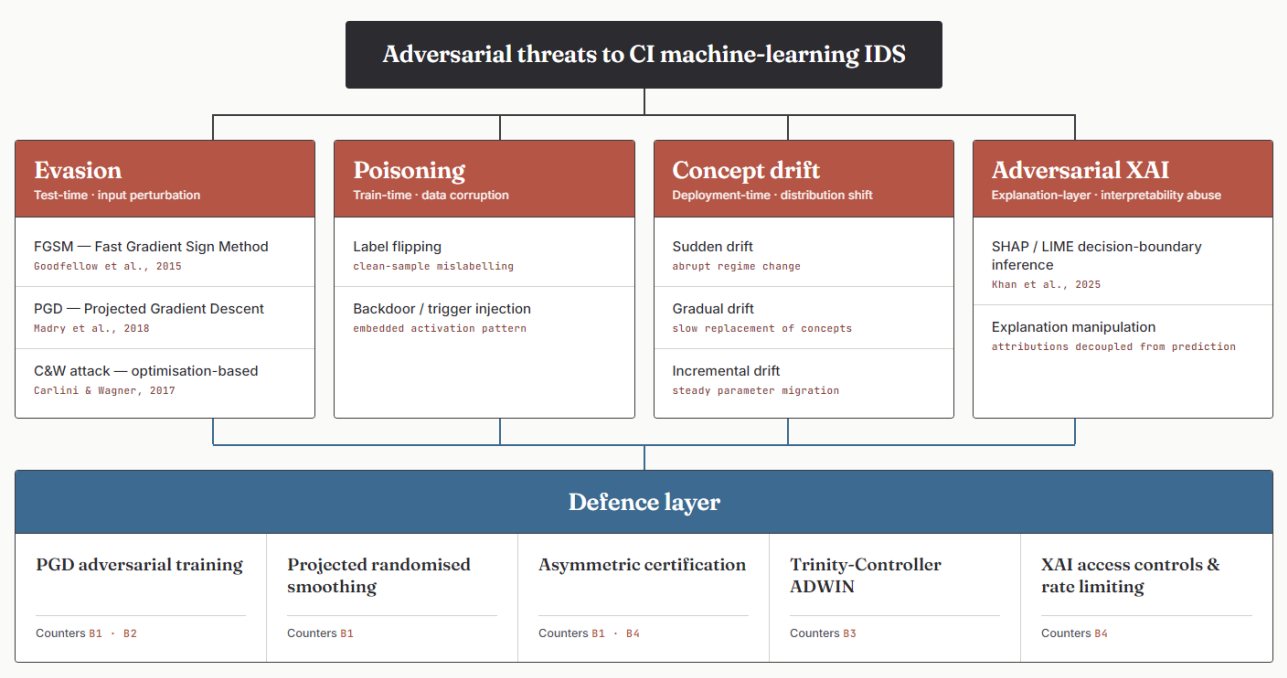


Table 3.1: Framework layer to NIST CSF 2.0 mapping

Layer	NIST CSF 2.0	Key sub-categories	Primary sources
L1: Data acquisition	Identify (ID)	ID.AM-01, ID.RA-01	Mathur & Tippenhauer (2016); Dadkhah et al. (2024)
L2: Preprocessing	Protect (PR)	PR.DS-02, PR.PS-01	Zhevnenko et al. (2026); Yang & Shami (2023)
L3: Robust ML	Detect (DE)	DE.AE-02, DE.CM-01	Pfrommer et al. (2023a,b); Cullen et al. (2025)
L4: XAI	Respond (RS)	RS.AN-03, RS.CO-02	Khan et al. (2025); Sunkara (2025)
L5: Response & adapt	Govern + Recover (GV+RC)	GV.PO-01, RC.RP-01	EDPS (2025); Lavaur & Busnel (2025)

4. RESULTS AND DISCUSSION

4.1 ARCHITECTURE FOR CI: THE SECTOR-SPECIFIC CASE

The clearest finding from the corpus is that architecture selection for CI IDS is a sector-specific engineering decision, not a general optimisation problem. It depends on the threat profile, computational constraints, and data ecology of each specific deployment context. A lot of published work optimises for a generic IT network benchmark and then describes the result as a CI security solution. That framing is where the problem starts.

For energy and water, GNN+Transformer hybrids make structural sense. Gamage et al. (2025) reviewed 36 studies and found GNNs consistently outperform per-flow classifiers on lateral movement and reconnaissance, because they model communication topology rather than individual flows. The Transformer element adds temporal depth for APT campaigns that unfold across shifts. Yang and Shami's (2023) MSANA showed online versions can sustain 99.32% accuracy below 4 ms while adapting to drift. Muzibuddin et al. (2026) add the label-scarcity dimension: USAD reached AUC 0.93 on SWaT without labelled attack data, which is practically important when historical attack data either does not exist or cannot be shared.

Healthcare IoMT is a different context entirely. Alharbi et al. (2025) achieved 99.92% precision on CICIoMT2024 with an ensemble of Random Forest, AdaBoost, and Bagging. IoMT traffic is categorical and tabular; it does not have the spatial or temporal structure that GNNs and Transformers are built to exploit. Alert fatigue is the dominant clinical operational concern, so precision matters more than recall, and the hard-voting mechanism delivers lower false positives than soft-voting alternatives. Heavier architectures may match the accuracy but not within edge device computational constraints.

Transportation is where the evidence is thinnest. Arslan et al. (2026) showed strong HCABS-NID performance on UNSW-NB15 (98.20%, 3.40 ms, 96.8% under FGSM), and Kasprzyk and Rychlicki (2025) showed XGBoost performing adequately on synthetic LoRaWAN data (95%, 0.02 FPR). Neither used a V2X-specific dataset. Neither Arslan et al. nor Kasprzyk and Rychlicki are at fault here. The benchmark simply does not exist.

Latency deserves explicit attention. Sub-4ms is achievable for network-layer ICS monitoring (Yang & Shami, 2023; Arslan et al., 2026). Singh et al. (2025) reported 1.2 seconds for RF+LSTM with XAI and policy enforcement, which is acceptable for network-layer detection but not for ICS control-loop contexts. GNN+Transformer hybrids and certified smoothing add overhead that has not been benchmarked in real-time CI settings. For V2X applications where collision avoidance operates within 100ms communication cycles, only the lightest architectures are feasible without hardware acceleration.

4.2 ADVERSARIAL RESILIENCE: WHAT ACTUALLY WORKS

Sitawarin (2024) is the most useful starting point. His analysis concludes that adversarial fragility is a structural property of learned classifiers in high-dimensional spaces, not an implementation problem with an architectural fix. That claim, backed by a decade of adversarial ML research, sets realistic expectations for any single defence mechanism.

Cullen et al. (2025) showed the specific ways certified robustness falls short in practice. Certified classifiers can be forced to abstain within their certified radius, a denial-of-service effect that does not technically violate the certificate but defeats the purpose of detection. Certification bounds computed over ℓp-norm perturbations do not reflect the protocol-valid, semantically constrained manipulations available to real CI attackers. Any publicly known certification scheme gives adaptive adversaries a map of its gaps. None of this makes certification useless. Pfrommer et al.'s (2023a) projected randomised smoothing produces certified volumes orders of magnitude larger than vanilla smoothing for CI sensor data on a physical manifold. Pfrommer et al.'s (2023b) asymmetric certification guarantees robustness specifically against false-negative adversaries, aligning the mathematical defence with the actual cost asymmetry of CI detection. These are real contributions. They are tools for defence-in-depth, not standalone solutions.

On the empirical side, Arslan et al. (2026) maintained 96.8% accuracy under 5% FGSM corruption with HCABS-NID. The ensemble diversity is doing work here: multiple heterogeneous base classifiers are harder to fool simultaneously than a single model. Singh et al. (2025) took a pipeline approach, intercepting adversarial perturbations through input validation before they reach the classifier.

The unresolved problem sits at the intersection of adversarial robustness and concept drift. Post-drift model retraining invalidates pre-existing certified robustness guarantees. Sebopelo's (2026) Trinity-Controller ADWIN result is the current best for drift adaptation. A system that is simultaneously drift-adaptive and certified robust does not exist in the reviewed literature. In a long-running CI monitoring deployment, that means the system loses either its robustness certification or its drift adaptation every time one of them operates.

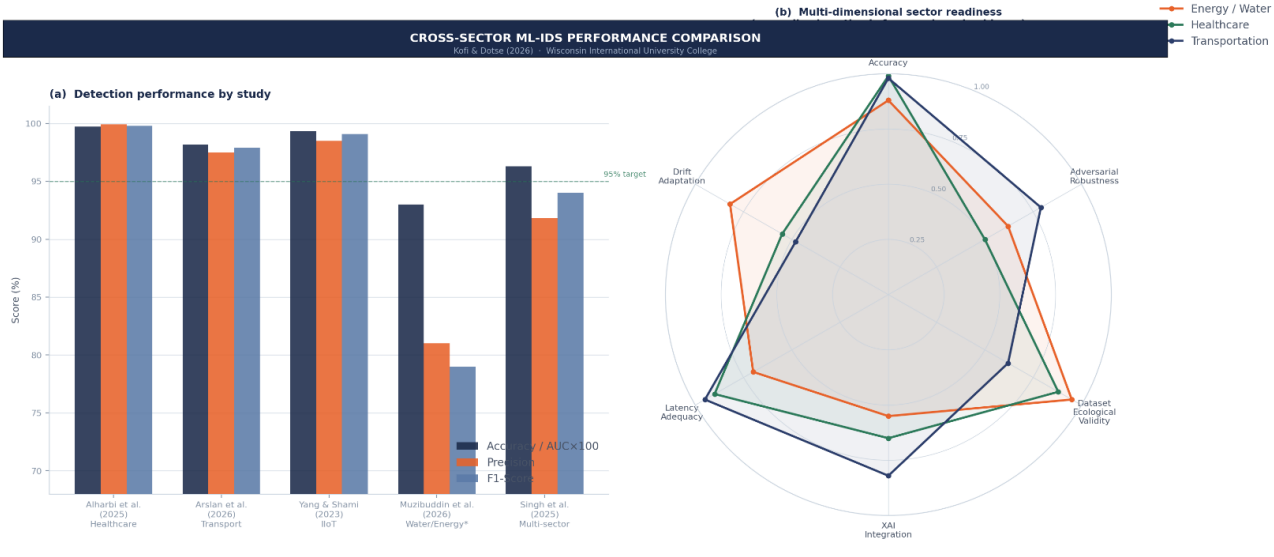
4.3 XAI: USEFUL AND DANGEROUS IN EQUAL MEASURE

The trust benefits from XAI integration are real. Oriaro and Mishra (2025) found 40% higher trust and 30% lower false-positive investigation time across 169 studies. Sunkara (2025) measured a 4.3/5 trust rating for XGBoost with SHAP in an actual SOC evaluation, with analysts describing the explanations as functioning like a pre-screening stage. Prasad et al. (2026) identified that framing outputs in terms of MITRE ATT&CK TTPs was substantially more useful to operators than raw Shapley values, because operators already understood the attack taxonomy.

Khan et al.'s (2025) finding is the counterweight. SHAP and LIME outputs, when queryable without restriction, can reconstruct model decision boundaries with over 90% success. An adversary who studies the XAI-enabled alert system learns what an uncredentialed SOC analyst learns. Deploying explainability without access controls is not a security improvement. The Adversarial XAI Paradox is not a future risk; it describes exploits possible against currently deployed systems.

Pfrommer et al.'s (2023b) feature-convex architecture is the nearest technical mitigation. By enforcing robustness structurally, there is no post-hoc boundary information to extract. But this applies only to false-negative adversaries and trades off model expressivity. For general CI IDS, the response is operational: rate-limit explanation endpoints, gate them behind authentication, log queries, and monitor for systematic probing patterns. This raises the cost of the attack without eliminating it. A general technical resolution to the Adversarial XAI Paradox remains the field's most pressing open problem.

4.4 CROSS-SECTOR VALIDATION: HOW MUCH CONFIDENCE CAN WE HAVE?



Energy and water have the SWaT testbed (Mathur & Tippenhauer, 2016): six-stage physical water treatment plant, 51 sensors and actuators at 1 Hz, 11 days of operation, 36 physically meaningful attacks. Zhevnenko et al.'s (2026) event-level benchmarking across seven industrial datasets under five deployment stressor conditions is the methodological ceiling

for this domain. The confidence level for energy and water recommendations is high.

Healthcare has CICIoMT2024 (Dadkhah et al., 2024): 40 IoMT devices, 18 attacks, three protocols. Alharbi et al. (2025) evaluated it thoroughly. The caveat is 15 simulated out of 40 devices, which may not fully capture real clinical device behaviour. Confidence is high with that qualification.

Transportation has UNSW-NB15 and synthetic LoRaWAN traffic. Neither captures V2X protocol semantics, vehicular mobility patterns, or DSRC/C-V2X timing constraints. Deploying a transportation CI IDS based solely on these results would require additional sector-specific validation before operational use.

Table 4.1: Key results by sector

Sector	Study	Dataset	Method	Accuracy	Adv. robust?	Drift-adaptive?
Energy/Water	Muzibuddin et al. (2026)	SWaT	USAD (unsupervised)	AUC 0.93	Partial	No
Healthcare	Alharbi et al. (2025)	CICIoMT2024	RF+AdaBoost+Bagging	99.72%	Not tested	No
Transport	Arslan et al. (2026)	UNSW-NB15*	HCABS-NID	98.20%	96.8% (FGSM)	No
Transport	Kasprzyk & Rychlicki (2025)	Synth. LoRaWAN*	XGBoost	95.00%	Not tested	No
Multi-sector	Singh et al. (2025)	Mininet/SCADA sim	RF+LSTM+XAI	96.30%	Input validation	No
IIoT (multi)	Yang & Shami (2023)	CICIDS2017, IoTID20	MSANA	99.32%	Not tested	Yes

** Not a sector-specific CI dataset; ecological validity caveat applies.*

The last column of Table 2 is the one to focus on. Every study testing adversarial robustness skips drift adaptation. Every study testing drift adaptation skips adversarial robustness. In a live CI environment, those are not separate events. The absence of any study addressing both together is an engineering gap with real operational consequences.

4.5 THE FIVE GAPS, VALIDATED

Gap 1, adversarial-ecological validity: no reviewed study evaluates adversarially trained or certified CI IDS on ICS-specific datasets under problem-space threat models. Arslan et al. (2026) came closest, testing FGSM on UNSW-NB15, which is still a generic IT benchmark with feature-space perturbations.

Gap 2, the Adversarial XAI Paradox: Khan et al. (2025) establish it at over 90% decision-boundary inference success. Pfrommer et al. (2023b) provides a partial architectural response for asymmetric adversaries only. No study provides a general resolution.

Gap 3, certified robustness and drift incompatibility: the two literatures are completely separate. Post-drift retraining invalidates existing certificates. No reviewed work addresses the integration.

Gap 4, cross-sector federated IDS: Lavaur and Busnel (2025) show Byzantine-tolerant FL aggregation degrades above approximately 30% compromised participants. No evaluated federated CI IDS spans all three sectors. Censinet et al. (2026) find that 70% of healthcare organisations lack enterprise AI inventories, meaning the governance preconditions for federated healthcare IDS are not in place in most real deployments.

Gap 5, standardised evaluation: Table 2 demonstrates this directly. Different metrics, different datasets, different threat conditions, no drift testing, no standardised latency measurement. Cross-study comparison is not reliable without a protocol specifying all of these simultaneously.

5. CONCLUSION

The performance problem for CI IDS is largely solved. What remains is the operational robustness problem. Current systems are adversarially fragile, drift-vulnerable, ecologically undervalidated, and caught in an XAI paradox that reduces security while improving usability. We think these are solvable engineering problems, but solving them requires shifting the field's attention from benchmark performance toward the conditions of actual deployment. That shift is not yet happening at the scale the problem requires.

The framework proposed here is a structured account of what a genuinely robust CI IDS would need. It is grounded in 44 verified sources, mapped to NIST CSF 2.0, and explicit about where the evidence runs out. Its purpose is to make the required design choices visible so they can be tested, argued with, and improved.

5.1 FOR RESEARCHERS

The most urgent priority is adversarial evaluation on ICS-specific datasets under problem-space threat models: attacks that respect OT protocol constraints, not just ℓp-norm feature perturbations. Every current adversarial evaluation uses feature-space attacks on IT network data. Until evaluations run on SWaT, WADI, or equivalents against protocol-valid adversarial traffic, published robustness claims about CI IDS cannot be verified in any operationally meaningful sense.

The incompatibility of certified robustness and concept drift adaptation is the second priority. These are separate communities working on mechanisms that cannot currently coexist in the same deployed system. Either a robustness-preserving drift adaptation algorithm or continuous re-certification in the streaming setting would resolve this. Both are tractable problems. Neither has been attempted.

The Adversarial XAI Paradox needs a technical response, not just an operational one. Intrinsic interpretability architectures are the more promising direction: concept-based models, prototype networks, and rule-extracting networks all provide explanations without exposing decision boundaries to external queries. And someone needs to build a public V2X or ITS-specific network security benchmark with the ecological validity of SWaT. That is a research infrastructure problem, but it is holding up progress in transportation CI security more than any modelling gap is.

5.2 FOR PRACTITIONERS

Architecture selection should start with a threat-profile analysis

of the specific deployment context, not benchmark ranking. The evidence shows clearly that what works in healthcare IoMT does not automatically transfer to energy OT environments. Trained models should be treated as perishable assets with production monitoring, not as static software components deployed once and left. Trinity-Controller ADWIN (Sebopelo, 2026) is a practical open-source starting point for streaming drift management. SHAP and LIME should be deployed only in access-controlled, rate-limited SOC environments with explanation query audit logs. Map ML security investments to NIST CSF 2.0 and the relevant sector-specific framework (NERC CIP, HIPAA/SP 800-66, TSA Security Directives) before procurement, not after.

5.3 FOR REGULATORS

Ecological validity standards for ML IDS certification would substantially raise the quality of evidence submitted in support of CI deployment claims. Energy ICS certifications should require evaluation on physics-based datasets. Healthcare certifications should require multi-protocol IoMT evaluation. Transportation certifications should require V2X or ITS protocol-specific evaluation once a suitable benchmark exists. The EDPS (2025) recommendation for multi-party computation and trusted execution environments in federated learning deployments should be a baseline governance requirement, not an advisory. Khan et al. (2025) established the decision-boundary inference risk empirically; regulators should require adversarial red-teaming of queryable XAI interfaces as a certification precondition.

5.4 LIMITATIONS

This paper is a synthesis, not an experiment. No models were trained. No datasets were generated. The conclusions about model performance depend on published results, and publication bias likely over-represents approaches that worked. Several ResearchGate-linked sources could not be verified in full text and their findings are excluded from substantive claims. The generative AI attack surface is developing faster than the peer-reviewed corpus currently reflects; the empirical IDS literature has not caught up. The five gaps this paper identifies are testable in future experimental work, which is the appropriate next step.

REFERENCES

[1] Alharbi, A. A., Alharby, M., & Hanandeh, A. A. (2025). Securing healthcare systems and optimizing data analytics through IoMT threat detection. *AIMS Mathematics, 10*(11), 25274--25306. https://doi.org/10.3934/math.20251119

[2] Arslan, R., Özseven, T., Aydın, M. M., & Çelik, Y. (2026). Cybersecurity in intelligent transportation systems: A comparative study on AI-based anomaly detection and threat analysis. *Mechatronics and Intelligent Transportation Systems, 5*(1), 11--30. https://doi.org/10.56578/mits050102

[3] Batishchev, D., & Saad, M. (2025). The black box problem: AI decision-making in critical infrastructure and its implications [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202511.2276.v1

[4] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5--32. https://doi.org/10.1023/A:1010933404324

[5] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153--1176. https://doi.org/10.1109/COMST.2015.2494502

[6] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy* (pp. 39--57). https://doi.org/10.1109/SP.2017.49

[7] Censinet, American Hospital Association, Health-ISAC, Health Sector Coordinating Council, Scottsdale Institute, & University of Texas at Austin. (2026). *The 2026 healthcare cybersecurity benchmarking study*. https://www.censinet.com

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785--794). https://doi.org/10.1145/2939672.2939785

[9] Cisco Systems. (2025). *2025 Cisco cybersecurity readiness index*. https://www.cisco.com

[10] Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th ICML* (PMLR Vol. 97, pp. 1310--1320).

[11] Cullen, A. C., Montague, P., Erfani, S. M., & Rubinstein, B. I. P. (2025). Position: Certified robustness does not (yet) imply model security. In *Proceedings of the 42nd ICML* (PMLR Vol. 267, pp. 81185--81198). https://proceedings.mlr.press/v267/cullen25a.html

[12] Dadkhah, S., Carlos Pinto Neto, E., Ferreira, R., Molokwu, R. C., Sadeghi, S., & Ghorbani, A. A. (2024). CICIoMT2024: A multi-protocol dataset for assessing IoMT device security. *Internet of Things, 28*, 101351. https://doi.org/10.1016/j.iot.2024.101351

[13] European Data Protection Supervisor. (2025). *TechDispatch #1/2025 -- Federated learning*. https://www.edps.europa.eu

[14] Gamage, T. P. D., Gutierrez, J. A., & Ray, S. K. (2025). The role of graph neural networks, transformers, and reinforcement learning in network threat detection: A systematic literature review. *Electronics, 14*(21), 4163. https://doi.org/10.3390/electronics14214163

[15] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *Proceedings of ICLR 2015*. arXiv. https://arxiv.org/abs/1412.6572

[16] Grand View Research. (2025). *Cybersecurity in critical infrastructure protection market report 2025--2033*. https://www.grandviewresearch.com

[17] IBM Security. (2024). *Cost of a data breach report 2024*. IBM.

[18] Kasprzyk, Z., & Rychlicki, M. (2025). Comparative analysis of machine learning algorithms for sustainable attack detection in intelligent transportation systems using long-range sensor network technology. *Sustainability, 17*(20), 8985. https://doi.org/10.3390/su17208985

[19] Khan, N., Ahmad, K., Al-Fuqaha, A., & Khalil, I. (2025). Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: A systematic review. *Information, 16*(12), 1036. https://doi.org/10.3390/info16121036

[20] Lavaur, L., & Busnel, Y. (2025). *Tutorial: Federated learning and network security* [Tutorial, IEEE ICDCS 2025]. https://hal.science/hal-05288649v1

[21] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in NeurIPS 30* (pp. 4765--4774).

[22] Lütjens, B., Everett, M., & How, J. P. (2020). Certified adversarial robustness for deep reinforcement learning. *PMLR, 100*, 1328--1337. https://proceedings.mlr.press/v100/lutjens20a.html

[23] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *Proceedings of ICLR 2018*. https://arxiv.org/abs/1706.06083

[24] Mathur, A. P., & Tippenhauer, N. O. (2016). SWaT: A water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks* (pp. 31--36). https://doi.org/10.1109/CySWater.2016.7469060

[25] Muzibuddin, S., Reddy, T. D. G., Reddy, V. J. S. P., & Srimani, K. (2026). Enhancing critical infrastructure security using USAD for unsupervised anomaly detection. *IJERT, 15*(3). https://www.ijert.org

[26] National Institute of Standards and Technology. (2023). *AI risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

[27] National Institute of Standards and Technology. (2024). *The NIST cybersecurity framework (CSF) 2.0* (NIST CSWP 29). https://doi.org/10.6028/NIST.CSWP.29

[28] Oriaro, S., & Mishra, S. (2025). Improving cybersecurity through explainable artificial intelligence: A systematic literature review. *Issues in Information Systems, 26*(3), 387--400. https://iacis.org/iis/2025/3_iis_2025_387-400.pdf

[29] Page, M. J., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

[30] Palo Alto Networks. (2024). *2025 cybersecurity predictions*. Unit 42. https://www.paloaltonetworks.com

[31] Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. In *Proceedings of IEEE EuroS&P 2018* (pp. 399--414). https://doi.org/10.1109/EuroSP.2018.00035

[32] Pfrommer, S., Anderson, B. G., & Sojoudi, S. (2023a). Projected randomized smoothing for certified adversarial robustness. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=FObkvLwNSo

[33] Pfrommer, S., Anderson, B., Piet, J., & Sojoudi, S. (2023b). Asymmetric certified robustness via feature-convex neural networks. In *Advances in NeurIPS 36* (pp. 52365--52400). https://proceedings.neurips.cc

[34] Pfrommer, S. I. (2025). *Safety, robustness, and interpretability in machine learning* [Doctoral dissertation, UC Berkeley]. EECS Tech Report UCB/EECS-2025-67. https://www2.eecs.berkeley.edu

[35] Prasad, P. W. C., Sayeed, M. S., Nguyen, D.-M., Hutabarat, D. P., & Mohiuddin, G. M. (2026). Explainable AI: Enhancing decision-making in the detection of cyber threats. *Frontiers in Computer Science, 8*, 1762332. https://doi.org/10.3389/fcomp.2026.1762332

[36] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM KDD* (pp. 1135--1144). https://doi.org/10.1145/2939672.2939778

[37] Sebopelo, R. B. (2026). Trinity-Controller ADWIN: An accuracy-guided sensitivity control framework for streaming intrusion detection. *Journal of Information Systems and Informatics, 8*(1), 501--529. https://doi.org/10.63158/journalisi.v8i1.1421

[38] Singh, A., Duvvada, S. R., Nisha, R. S., Parthiban, K., Niveditha, S. R., & Vineesha, M. (2025). A hybrid defense framework for critical infrastructure. In *Proceedings of ICRDICCT'25* (pp. 820--826). SciTePress. https://doi.org/10.5220/0013944300004919

[39] Sitawarin, C. (2024). *New perspectives on adversarially robust machine learning systems* [Doctoral dissertation, UC Berkeley]. EECS Tech Report UCB/EECS-2024-10. https://www2.eecs.berkeley.edu

[40] Sophos. (2024). *The state of ransomware in healthcare 2024*. Sophos Ltd.

[41] Sunkara, G. (2025). Explainable AI for cyber threat intelligence: Enhancing analyst trust. *Open Access Research Journal of Science and Technology, 14*(2), 29--40.

[42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in NeurIPS 30* (pp. 5998--6008).

[43] Verizon. (2024). *2024 data breach investigations report*. Verizon Business.

[44] Wickramasinghe Brahmana, C. S., Marino, D., De Silva, D., & Manic, M. (2025). Editorial: Machine learning for cybersecurity. *Frontiers in Artificial Intelligence, 8*, 1640609. https://doi.org/10.3389/frai.2025.1640609

[45] Yang, L., & Shami, A. (2023). A multi-stage automated online network data stream analytics framework for IIoT systems. *IEEE Transactions on Industrial Informatics, 19*(2), 2107--2116. https://doi.org/10.1109/TII.2022.3212003

[46] Yin, J., Xie, W., Liang, G., Zhang, L., & Zhang, X. (2025). Concept drift detection and adaptation method for IoT security framework. *China Communications, 22*(12), 137--147. https://doi.org/10.23919/JCC.fa.2022-0379.202512.

[47] Zhevnenko, D., Makarov, I., Kovalenko, A., Meshchaninov, F., Kozhukhov, A., Travnikov, V., Ippolitov, M., Yashunin, K., & Katser, I. (2026). Benchmarking IoT time-series anomaly detection with event-level augmentations [Preprint]. arXiv. https://arxiv.org/abs/2602.15457.