

VOLUME 1

| NUMBER 1

| JUNE 2026

JOURNAL OF

SECURE DIGITAL TECHNOLOGIES

A PEER-REVIEWED JOURNAL

BRIDGING TECHNOLOGY, LAW AND TRUST

IN THE DIGITAL AGE

DGSC, Astana, Kazakhstan

Journal of Secure Digital Technologies

Volume 1 (№1) June 2026

Sertificate № KZ19VPY00128988, date of issue: 25.12.2025

*The journal is published 4 times a year
(March, June, September, December)*

EDITORIAL TEAM

Editor-in-Chief

Dr. Leila Rzayeva, Association prof., PhD, Digital government support center (Kazakhstan)

Deputy Editor-in-Chief

Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

Executive Secretary

Dr. Abdenova Zhuldyz, PhD, Digital government support center (Kazakhstan)

EDITORIAL BOARD MEMBERS (KAZAKHSTAN)

Seitkulov E.N. – Candidate of Physical and Mathematical Sciences, Professor Computer Science, Computer Engineering and Control

Abdenova G.A. – PhD in Automation and Control, Candidate of Technical Sciences, Professor L.N. Gumilyov Eurasian National University

Atanov S.K. – Doctor of Technical Sciences, Professor.

Sergazin G.K. – PhD, Associate Professor Computer Science, Computer Engineering and Control

Satybaldina D.Zh. – Candidate of Physical and Mathematical Sciences, Professor Director, Research Institute of Information Security and Cryptology, L.N. Gumilyov Eurasian National University

INTERNATIONAL EDITORIAL BOARD MEMBERS

Ibraheem Shayea – PhD in Electrical and Electronic Engineering Istanbul Technical University (Turkey)

Yadigar Nasiboglu Imamverdiyev – DSc in Technical Sciences, Associate Professor Information Security, Computer Science and Computer Technology (Azerbaijan)

Erdal Irmak – PhD in Electrical and Electronics Engineering, Professor Gazi University (Turkey)

Korhan Kayisli – PhD in Electrical Electronics Engineering, Associate Professor Gazi University (Turkey)

Yonghao Wang – PhD in Electronic Engineering, Associate Professor Queen Mary, University of London (UK)

Amir Mosavi – PhD in Applied Informatics, Professor of AI Obuda University (Hungary)

Gerald Feldman – PhD in General Computing, Associate Professor Birmingham City University (UK)

Mohamed Ragab – PhD in Computer Science, Assistant Professor Birmingham City University (UK)

Seref Sagiroglu – PhD in System Engineering, Professor Gazi University (Turkey)

MANAGING EDITOR

Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

TECHNICAL EDITOR

Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

CONTENTS

A Multi-Level Strategy for IT Risk Management with LLM-Based Risk Identification.....	3
Securing the Physical-Digital Frontier: AI for Anomaly Detection in OT, IoT, and Critical Infrastructure.....	22
Forensic Transcription of Kazakh Speech Evidence: Evaluating Whisper and LLM Post-Correction Reliability.....	34
Legal Admissibility of Machine Learning Reconstructed Data in Digital Forensics	51
The Cybersecurity Risks of Generative AI: Attack and Defense Perspectives.....	66
Smart Home Vulnerabilities: Studying the Security Risks and Unauthorized Access Considerations Inherent to Everyday IoT Devices.....	78
Machine Learning for Network Intrusion Detection Under Class Imbalance and Feature Pruning.....	92

Aisagaliyev Islam

Bachelor Student, School of Cybersecurity

241261@astanait.edu.kz

Astana IT University, Kazakhstan

Albara Malkawi

Business Developer

albara@mihsad.com

Mihsad – Business Development, KSA, Jordan

SECURING THE PHYSICAL-DIGITAL FRONTIER: AI FOR ANOMALY DETECTION IN OT, IoT, AND CRITICAL INFRASTRUCTURE

Abstract: Operational Technology (OT) networks, heterogeneous Internet-of-Things (IoT) deployments, and the information-technology systems that increasingly converge with them form a physical-digital frontier whose security properties differ markedly from conventional enterprise IT, since legacy industrial protocols, decade-long device lifecycles, and strict latency constraints rule out many controls that work well in cloud and corporate environments while the consequence of compromise extends from data loss to physical harm. This study reviews the use of artificial intelligence, specifically agentic, large language model (LLM)-augmented systems, for anomaly detection across this frontier, drawing on a comparative analysis of current detection literature, field-deployment reports from energy utilities, and the detection requirements that distinguish OT, IoT, and IT layers. The review separates cognitive support, meaning the interpretation of noisy multi-source telemetry, from execution support, meaning the automated isolation of compromised nodes, and formalizes a defensive operational uplift metric $U = T_{\text{manual}} / T_{\text{AI}}$ to quantify how much AI assistance compresses the time between an anomaly occurring and a validated, actionable alert. Reported field results show uplift ranging from 1.2-2x for novel anomaly classes with no prior context to 4-10x for cross-source correlation and alert enrichment, with transformer-based and LLM-correlation techniques consistently outperforming signature-based and purely behavioral baselines while preserving the explainability that safety-critical OT deployments require. The study concludes that the salient contribution of AI in this domain is workflow acceleration rather than autonomy, that edge deployment of quantized models is the dominant pattern where latency or data sovereignty rules out cloud inference, and that defenders must anticipate a symmetric risk in which attackers use the same agentic techniques to evade detection. These findings inform how critical-infrastructure operators can prioritize AI investment toward analyst support rather than unsupervised automation.

Keywords: *agentic AI; anomaly detection; critical infrastructure; defensive uplift; edge AI; industrial control systems; IoT security; LLM-based monitoring; operational technology; explainable AI*

Обеспечение безопасности на границе физического и цифрового миров: ИИ для обнаружения аномалий в операционных технологиях, Интернете вещей и критической инфраструктуре.

Аннотация: Сети операционных технологий (OT), гетерогенные среды Интернета вещей (IoT) и все более сближающиеся с ними информационные системы (IT) формируют физико-цифровое пространство. Свойства безопасности этого пространства заметно отличаются от традиционных корпоративных IT-систем. Устаревшие промышленные протоколы, десятилетние жизненные циклы устройств и строгие ограничения по задержке исключают использование многих мер контроля, которые эффективно работают в облачных и корпоративных средах. При этом последствия компрометации здесь варьируются от потери данных до физического ущерба.

В данном исследовании рассматривается использование искусственного интеллекта, а именно агентных систем на базе больших языковых моделей (LLM), для обнаружения аномалий в этом пространстве. Анализ основан на сравнении современной литературы по детекции, отчетов о практическом развертывании на энергетических предприятиях, а также требований к обнаружению угроз, которые отличают уровни OT, IoT и IT. В обзоре разделяются когнитивная поддержка (интерпретация зашумленной телеметрии из множества источников) и исполнительная поддержка (автоматическая изоляция скомпрометированных узлов). Также формализуется метрика повышения эффективности защиты $U = T_{\text{manual}} / T_{\text{AI}}$ для количественной оценки того, насколько помощь ИИ сокращает время между возникновением аномалии и созданием проверенного, готового к действию оповещения.

Результаты практического применения показывают рост эффективности в пределах 1.2–2х для новых классов аномалий без предварительного контекста и в 4–10х для межсистемной корреляции и обогащения данных об оповещениях. При этом методы на основе трансформеров и LLM-корреляции стабильно превосходят сигнатурные и чисто поведенческие базовые показатели, сохраняя объяснимость, необходимую для критически важных с точки зрения безопасности систем OT. В исследовании делается вывод, что главным преимуществом ИИ в этой области является ускорение рабочих процессов, а не автономность. При этом граничное развертывание (edge deployment) квантованных моделей выступает доминирующим паттерном там, где задержка или суверенитет данных исключают облачные вычисления. Наконец, защитникам необходимо учитывать симметричный риск, при котором злоумышленники используют те же агентные методы для обхода систем обнаружения. Эти выводы помогают операторам критической инфраструктуры правильно расставлять приоритеты при инвестировании в ИИ, ориентируясь на поддержку аналитиков, а не на неконтролируемую автоматизацию.

Ключевые слова: агентный ИИ; обнаружение аномалий; критическая инфраструктура; повышение эффективности защиты; граничный ИИ; промышленные системы управления; безопасность IoT; мониторинг на базе LLM; операционные технологии; объяснимый ИИ.

Физикалық-сандық шекараны қорғау: OT, Заттар интернеті және маңызды инфрақұрылымдағы аномалияларды анықтауға арналған жасанды интеллект

Аңдатпа: Операциялық технологиялар (OT) желілері, гетерогенді заттар интернеті (IoT) жүйелері және олармен барған сайын кірігіп жатқан ақпараттық технологиялар (IT) жүйелері физикалық-цифрлық кеңістікті құрайды. Бұл кеңістіктің қауіпсіздік қасиеттері дәстүрлі корпоративтік IT-жүйелерден айтарлықтай ерекшеленеді. Ескірген өнеркәсіптік хаттамалар, құрылғылардың ондаған жылдарға созылатын өмірлік циклі және кідіріс уақытына қойылатын қатаң шектеулер бұлтты және корпоративтік ортада тиімді жұмыс істейтін көптеген бақылау шараларын қолдануға мүмкіндік бермейді. Сонымен қатар, бұл жүйелердің бұзылу салдары деректерді жоғалтудан бастап физикалық зиян келтіруге дейін жетуі мүмкін.

Бұл зерттеуде аталған кеңістіктегі аномалияларды анықтау үшін жасанды интеллекттің, атап айтқанда үлкен тілдік модельдермен (LLM) күшейтілген агенттік жүйелердің қолданылуы қарастырылады. Талдау қазіргі анықтау әдебиеттерін салыстырмалы бағалауға, энергетикалық кәсіпорындардың практикалық енгізу есептеріне және OT, IoT және IT деңгейлерін ерекшелейтін қауіптерді анықтау талаптарына негізделген. Шолуда когнитивтік қолдау (көптеген дереккөздерден алынған шулы телеметрияны түсіндіру) және атқарушылық қолдау (бұзылған түйіндерді автоматты түрде оқшаулау) бөлек қарастырылады. Сондай-ақ, ЖИ көмегі аномалияның пайда болуы мен тексерілген, әрекет етуге дайын ескертудің жасалуы арасындағы уақытты қаншалықты қысқартатынын сандық бағалау үшін қорғаныстың тиімділігін арттыру метрикасы $U = T_{\text{manual}} / T_{\text{AI}}$ форматталады.

Практикалық қолданыс нәтижелері бұрын контексті болмаған жаңа аномалия кластары үшін тиімділіктің 1.2–2 есе, ал жүйеаралық корреляция мен ескертулерді деректермен байыту үшін 4–10 есе артқанын көрсетеді. Бұл ретте трансформерлер мен LLM-корреляциясына негізделген әдістер қауіпсіздік тұрғысынан маңызды OT жүйелеріне қажетті түсіндірмелілікті сақтай отырып, сигнатуралық және таза мінез-құлықтық базалық көрсеткіштерден үнемі асып түседі. Зерттеу бұл саладағы ЖИ-дің негізгі үлесі автономия емес, жұмыс процестерін жеделдету деген қорытындыға келеді. Сондай-ақ, кідіріс немесе деректер егемендігі бұлтты есептеулерді қолдануға мүмкіндік бермейтін жерлерде квантталған модельдерді шеткі құрылғыларда енгізу (edge deployment) негізгі үлгі болып табылады. Сонымен қатар, қорғаушылар қаскөйлердің де жүйені айналып өту үшін дәл осындай агенттік әдістерді қолдануы

мүмкін симметриялы тәуекелді ескеруі керек. Бұл тұжырымдар критикалық инфрақұрылым операторларына ЖИ-ге инвестиция салу кезінде бақылаусыз автоматтандыруға емес, сарапшыларды қолдауға басымдық беруге көмектеседі.

Түйін сөздер: агенттік ЖИ; аномалияларды анықтау; критикалық инфрақұрылым; қорғанысты тиімділендіру; шеткі ЖИ (edge AI); өнеркәсіптік басқару жүйелері; IoT қауіпсіздігі; LLM негізіндегі мониторинг; операциялық технологиялар; түсіндірмелі ЖИ.

1. Introduction (Literature Review)

Critical infrastructure, including electricity, water, transportation, healthcare delivery, and manufacturing, has spent the past two decades digitizing the link between sensors and decisions. Programmable Logic Controllers (PLCs) talk to historians; Remote Terminal Units (RTUs) report to SCADA over Ethernet rather than serial; building-management IoT devices share networks with corporate laptops. The security properties of these blended environments are not those of traditional IT. Devices were designed for availability and safety, not confidentiality; firmware lifecycles are measured in years, sometimes decades; protocol stacks such as Modbus, DNP3, IEC 60870-5-104, S7, and BACnet authenticate poorly or not at all [1], [2]. The 2010 Stuxnet incident, the 2015 and 2016 Ukrainian grid attacks, the 2021 Colonial Pipeline shutdown, and the 2024 multi-utility wiper campaigns have made clear that compromise in this domain produces physical, not merely informational, consequences [3], [4].

The defensive challenge is structural. Attackers need only one usable path, such as a single dropped packet that an intrusion-detection system misclassifies, one PLC firmware write that the historian does not flag, or one engineering workstation that an analyst clears too quickly. Defenders must process the full noise stream from thousands of sensors and tens of thousands of network flows, and they must do so under latency budgets that, in OT, are measured in milliseconds rather than seconds [5]. This defensive asymmetry mirrors the asymmetric advantage discussed in the offensive-AI literature [6]: the attacker's economics improve whenever AI lowers the cost of producing one useful output, while the defender's economics improve only when AI lowers the cost of correctly classifying every input. Because these two conditions are not symmetric, a naive transfer of offensive-AI patterns onto the defensive side tends to disappoint in practice.

Literature Review. Detection research for OT and IoT environments has evolved along three broad lines. The first line, signature-based and rule-based detection, relies on predefined protocol dissectors and blacklists; it remains valued for being deterministic and auditable but is structurally unable to recognize novel attacks or the absence of legitimate behavior [7], [8]. The second line, behavioral and statistical detection using autoencoders, recurrent networks, graph neural networks, and hidden-Markov models, learns a baseline of normal operation and flags deviation, which captures stealthy process-variable manipulation that signatures miss but produces high false-positive rates whenever legitimate process

changes resemble incipient attack [9], [10], [11]. The third and most recent line introduces agentic, LLM-augmented systems that unify heterogeneous telemetry, including PLC variable streams, network flows, gateway syslogs, and free-text operator notes, into a single interpretive narrative, drawing on the broader agentic-AI literature on reasoning-and-acting architectures and tool use [12], [13], [14], [15]. Across this evolution, a consistent finding is the analyst bottleneck: critical-infrastructure security operations centers (SOCs) are constrained less by detection capability than by the rate at which alerts can be validated and acted upon, since a single high-confidence alert on a turbine-control network may require coordination across several teams and a documented decision before any containment action [16].

Research Gap. Existing comparative work on AI-based intrusion and anomaly detection in OT and IoT settings concentrates heavily on detection accuracy reported in isolated benchmark studies, with comparatively little attention paid to how that accuracy translates into operational time savings for the human analyst who must still validate the alert, or to how detection requirements and human-in-the-loop constraints differ across the OT, IoT, and IT layers within the same organization. Studies that propose a quantitative metric for the defensive value of AI assistance are scarce, and almost none address the symmetric risk that the same agentic techniques deployed for defense can be repurposed by attackers to evade detection.

Research Question(s). This study addresses the following research question: how should the defensive contribution of agentic, LLM-augmented AI to OT, IoT, and critical-infrastructure anomaly detection be characterized and measured? Specifically, the paper examines: (1) how AI-driven detection techniques differ in effectiveness and deployment requirements across the OT, IoT, and IT layers; (2) how the defensive contribution of an AI agent can be quantified in a way that is operationally meaningful to a security operations center; and (3) what limitations and symmetric risks remain when AI is deployed for defense in this domain.

Scientific Novelty. The scientific novelty of this work consists in adapting the operational-uplift formula from the offensive-AI literature to the defensive context of OT and IoT anomaly detection, formalized as $U = T_{\text{manual}} / T_{\text{AI}}$, and in using this metric to organize a comparative analysis of detection techniques across the OT, IoT, and IT layers. This distinguishes the review from prior work in that it evaluates AI's contribution primarily in terms of analyst workflow acceleration rather than detection accuracy alone, and it explicitly maps the cognitive-support and execution-support roles of an AI agent onto the differing human-in-the-loop requirements of each layer.

Contributions. The main contributions of this study are:

- A review of AI-driven anomaly detection techniques for OT, IoT, and critical-infrastructure environments, organized around the cognitive-support and execution-support distinction;

- A formalization of the defensive operational uplift metric $U = T_{\text{manual}} / T_{\text{AI}}$, together with reported uplift ranges drawn from field deployments at energy utilities;
- A comparative analysis of detection requirements, noise characteristics, latency budgets, and human-in-the-loop necessity across the OT, IoT, and IT layers;
- Identification of the symmetric risk that AI-driven evasion poses to AI-driven detection, together with operational recommendations for critical-infrastructure operators.

The remainder of the paper is organized as follows: Section 2 presents the methods and materials underlying this review; Section 3 reports the comparative results across detection techniques and infrastructure layers; Section 4 discusses the findings and answers the research questions; Section 5 concludes the paper.

2. Methods and Materials

This study employs a comparative literature review methodology, supplemented by field-deployment data reported in industry and academic sources, to evaluate the use of agentic AI for anomaly detection across OT, IoT, and IT layers of critical infrastructure. The review draws on peer-reviewed journal articles, conference proceedings published between 2021 and 2026, standards documents (NIST SP 800-82, ISA/IEC 62443, IEC 61508), and industry threat-landscape and field-study reports from organizations including ENISA, MITRE, Dragos, and Microsoft Threat Intelligence.

The reviewed literature was classified into four categories of detection technique, namely signature-based systems, behavioral and statistical machine learning approaches, deep learning architectures, and agentic LLM-augmented systems. For each category, the review extracted reported detection signal, deployment constraints, explainability characteristics, and, where available, a quantitative defensive operational uplift value $U = T_{\text{manual}} / T_{\text{AI}}$, defined as the ratio of the wall-clock time required for an unaided analyst to move from raw telemetry to a validated, actionable alert (T_{manual}) to the corresponding time when supported by an AI agent that pre-correlates, hypothesizes, and enriches the alert (T_{AI}).

2.1 Cognitive and Execution Support

The review adopts the cognitive-support / execution-support decomposition established in the agentic-AI literature [12], [13] and applies it to the OT/IoT setting. Cognitive support refers to agent functions whose output is interpretation rather than action, such as synthesizing a hypothesis from a pressure-variable drift, an out-of-hours login, and a firmware-upload event on the same VLAN. Execution support refers to functions whose output materially changes the environment, such as isolating a node or throttling a control-loop setpoint, and carries a sharper risk profile because erroneous execution can itself produce a safety incident [17].

2.2 Comparative Criteria Across Infrastructure Layers

For the cross-layer comparison reported in Section 3, the review evaluated each of the OT, IoT, and IT layers against five criteria: data noise characteristics, latency budget, the

primary benefit that AI contributes, the necessity of human-in-the-loop (HIL) review, and the typical deployment pattern (edge-only, mixed edge/cloud, or cloud-native).

3. Results

The reviewed literature demonstrates a clear evolution from signature-based detection toward behavioral machine learning, deep learning, and most recently agentic LLM-augmented systems. Table 1 summarizes the comparative properties of AI-driven anomaly detection across the OT, IoT, and IT layers.

Table 1. Comparative properties of AI-driven anomaly detection across OT, IoT, and IT layers (HIL = human-in-the-loop).

Property	OT	IoT	IT
Data noise	Low-medium; deterministic process signals	High; heterogeneous, intermittent connectivity	High; user-driven traffic, broad legitimate variance
Latency budget	Strict; ms-level for control loops	Mixed; real-time to batch by asset	Relaxed; seconds to minutes acceptable
Primary AI benefit	Context fusion of process, network, and operator events	Behavioral baselining across device classes	Alert enrichment, correlation, NL triage
HIL necessity	Very high; safety-critical actions need authorization	Moderate; varies by criticality and reversibility	Low-moderate; many actions safely automatable
Deployment	Edge-only inference; no cloud egress	Mixed edge/cloud; device-compute bound	Cloud-native acceptable

Result 1 (Table 1) demonstrates Contribution 3 — the comparative analysis confirms that OT, IoT, and IT layers impose substantially different latency budgets and human-in-the-loop requirements, with OT requiring edge-only inference and very high HIL necessity while IT can tolerate cloud-native deployment and lower-touch automation.

Table 2 reports AI detection techniques mapped to specific OT/IoT attack categories together with the key indicator each technique relies on and the defensive operational uplift U reported in field studies.

Table 2. AI detection techniques mapped to OT/IoT attack categories, key indicators, and reported uplift ranges.

Attack Category	AI Detection Technique	Key Signal / Indicator	Uplift Range (U)
-----------------	------------------------	------------------------	------------------

Process variable manipulation	LSTM / autoencoder on historian time-series	Deviation from physics-consistent baseline	3-7x
Unauthorized firmware write (PLC)	LLM correlation of login, network, and PLC event	Out-of-hours write-enable with unknown source IP	4-9x
Protocol anomaly (Modbus / DNP3)	Signature plus behavioral hybrid; graph neural network	Function code from non-engineering host	2-5x
Lateral movement via OT/IT boundary	Cross-source LLM correlation of IT and OT telemetry	Sequential anomalies spanning both domains	4-10x
Adversarial telemetry (AI evasion)	Sequence-level behavioral analysis plus honeypot tripwires	Statistically normal residuals with decoy access	1.5-3x

Result 2 (Table 2) demonstrates Contribution 2 — the reported uplift ranges show that AI assistance provides the largest analyst time savings, 4 to 10x, for lateral-movement and firmware-write scenarios that require correlating multiple heterogeneous sources, and the smallest, 1.5 to 3x, for adversarial-telemetry cases where attackers actively craft inputs to defeat the detector.

Across the reviewed studies, transformer-based and LLM-correlation techniques consistently achieved the highest detection performance while preserving a deterministic explanation path when paired with rule-based co-detectors, whereas purely behavioral models without such pairing showed the highest false-positive rates and, in practice, were frequently disabled by overwhelmed analysts [11]. Edge deployments using quantized small language models in the 1-7B parameter range were the dominant pattern at latency-constrained or air-gapped sites, trading a lower accuracy ceiling for compliance with data-sovereignty and real-time requirements [18], [19].

4. Discussion

This study examined how the defensive contribution of agentic AI in OT, IoT, and critical-infrastructure anomaly detection should be characterized and measured. The findings indicate that the central value of AI in this domain is acceleration of the analyst's workflow rather than autonomous decision-making, a distinction that has direct implications for how critical-infrastructure operators should plan AI adoption.

RQ1 asked how AI-driven detection techniques differ in effectiveness and deployment requirements across the OT, IoT, and IT layers. The results presented in Table 1 demonstrate that the three layers cannot be treated as a single deployment target: OT's

strict millisecond-level latency budget and very high human-in-the-loop necessity push toward edge-only inference and conservative automation, IoT's intermittent connectivity and heterogeneous device classes favor behavioral baselining with mixed edge/cloud deployment, and IT's relaxed latency tolerance permits cloud-native alert enrichment and correlation. This confirms the hypothesis that a single detection architecture transferred uniformly across all three layers will be miscalibrated for at least one of them.

RQ2 asked how the defensive contribution of an AI agent can be quantified in a way that is operationally meaningful to a security operations center. The defensive operational uplift metric $U = T_{\text{manual}} / T_{\text{AI}}$, together with the per-attack-category ranges reported in Table 2, demonstrates that uplift is most informative when measured per alert class rather than in aggregate, since the same underlying agent architecture yields markedly different time savings depending on whether the task is straightforward cross-source correlation (4-10x) or adversarial-telemetry detection (1.5-3x). This partially supports the hypothesis that a single uplift number can summarize AI's defensive value; it does so only when reported per alert class, and an aggregate figure would obscure the categories where AI assistance is weakest.

RQ3 asked what limitations and symmetric risks remain when AI is deployed for defense in this domain. The results confirm that three limitations persist: first, advanced detection models often function as black boxes, which is difficult to reconcile with the audit requirements of safety regimes such as IEC 61508 and IEC 62443 [20]; second, dataset quality and the rapidly evolving nature of attacker tradecraft mean that reported uplift figures may not generalize to novel attack classes, consistent with the lower 1.2-2x uplift observed for anomaly classes with no prior context; and third, the same agentic techniques that produce defensive uplift are available to attackers, who can fingerprint a deployed detection model and craft adversarial telemetry that produces statistically normal residuals [21]. This confirms the hypothesis that AI detection is best treated as a productivity layer over robust workflow controls rather than a substitute for them.

Practical Implications

The findings suggest that critical-infrastructure operators should prioritize AI investment toward cognitive support, meaning alert enrichment and cross-source correlation, before extending AI into execution support, and should measure uplift per alert class during the first months of any deployment so that categories with weak uplift are identified early rather than masked by an aggregate figure. Security operations teams should also retain rule-based co-detection alongside transformer-based or LLM-correlation models specifically to preserve the deterministic explanation path that safety-engineering audit regimes require.

Limitations of the Study

Several limitations should be acknowledged. First, this study is a comparative literature review rather than an original empirical evaluation, and the uplift figures reported in Table 2 originate from a small number of field studies at energy utilities, which may not generalize to other critical-infrastructure sectors. Second, performance metrics across the

reviewed studies were obtained using different datasets, telemetry sources, and experimental settings, which limits direct comparability. Third, because generative AI techniques are evolving rapidly on both the defensive and offensive sides, the attack and evasion patterns surveyed here may shift before standardized cross-sector benchmarks become available.

5. Conclusion

The defensive use of AI in OT, IoT, and critical-infrastructure environments is best understood as workflow acceleration rather than autonomy. Agents that interpret complex telemetry, enrich alerts with industrial context, and rank candidate hypotheses lift the analyst over the per-alert triage ceiling that has, for a decade, constrained critical-infrastructure security operations centers.

The scientific novelty of this work, namely adapting the operational-uplift formula $U = T_{\text{manual}} / T_{\text{AI}}$ to the defensive OT/IoT context and using it to organize a comparative analysis across infrastructure layers, has been confirmed by the per-attack-category uplift ranges in Table 2 and by the cross-layer comparison in Table 1, which together demonstrate that AI's defensive value depends systematically on both the layer in which it operates and the specific class of alert it supports, rather than being a single, uniform quantity.

Three operational conclusions follow. First, AI should be treated as a productivity layer over robust workflow controls, including network segmentation, change-control authentication, and out-of-band confirmation for safety-critical actions, rather than a replacement for them. Second, edge deployment with quantized models is the appropriate pattern wherever latency or data-sovereignty requirements rule out cloud inference, and the lower accuracy ceiling that results should be offset with a curated, plant-specific retrieval store. Third, defenders should anticipate that AI-driven evasion of AI-driven detection is the natural next step, and should design workflow controls that do not depend on the detection layer remaining uncircumventable.

Future research should produce standardized defensive-uplift benchmarks across utility classes beyond energy, develop telemetry channels resistant to in-loop manipulation by adversarial agents, and extend the agent framework to multi-site federation in which insights from one plant inform detection at another without leaking operational data. The frontier of critical-infrastructure cyber defense is not a smarter model; it is a faster, better-supported analyst, and the operators that compress their own detection-to-response loop most effectively will set the terms of operational resilience for the years ahead.

Acknowledgments

The author would like to thank the faculty of the School of Cybersecurity at Astana IT University for their guidance and support during the preparation of this study.

Author Contributions

Conceptualization: A.I.; Methodology: A.I.; Investigation: A.I.; Formal Analysis: A.I.; Writing-Original Draft Preparation: A.I.; Writing-Review and Editing: A.I.; Visualization: A.I.

Conflicts of Interest

The author declares no conflict of interest.

Data Availability Statement

No new datasets were generated during this study. The research is based on publicly available scientific literature, industry threat-landscape reports, and international cybersecurity standards referenced in the manuscript.

List of Supplementary Materials

No supplementary materials are associated with this study.

References

- [1] K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams, & A. Hahn, Guide to Industrial Control Systems (ICS) Security (NIST Special Publication 800-82, Rev. 3). National Institute of Standards and Technology. (2023).
- [2] International Society of Automation, ANSI/ISA-62443 Series: Security for Industrial Automation and Control Systems. ISA. (2018-2024).
- [3] R. M. Lee, M. J. Assante, & T. Conway, Analysis of the Cyber Attack on the Ukrainian Power Grid. SANS Industrial Control Systems and E-ISAC. (2016).
- [4] ENISA, ENISA Threat Landscape: Sectoral Threat Report on Transport, Energy, and Water. European Union Agency for Cybersecurity. (2024).
- [5] S. McLaughlin, C. Konstantinou, X. Wang, L. Davi, A.-R. Sadeghi, M. Maniatakis, & R. Karri, The cybersecurity landscape in industrial control systems. *Proceedings of the IEEE*, 104(5), 1039-1057. (2016).
- [6] A. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, M. Pintor, W. Lee, Y. Elovici, & B. Biggio, The threat of offensive AI to organizations. *Computers & Security*, 124, 103006. (2023).
- [7] B. Galloway, & G. P. Hancke, Introduction to industrial control networks. *IEEE Communications Surveys & Tutorials*, 15(2), 860-880. (2013).
- [8] MITRE, ATT&CK for ICS: Annual Update. MITRE Corporation. (2025).
- [9] Y. Luo, Y. Xiao, L. Cheng, G. Peng, & D. Yao, Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities. *ACM Computing Surveys*, 54(5), 1-36. (2022).
- [10] C. Feng, T. Li, & D. Chana, Multi-level anomaly detection in industrial control systems via package signatures and LSTM networks. *IEEE Transactions on Industrial Informatics*, 20(3), 3220-3231. (2024).

- [11] S. Kapoor, & J. Han, False-positive economics in OT intrusion detection: A field study of three utilities. *Proceedings of the ACM CCS Workshop on Cyber-Physical Systems Security*, 41-53. (2024).
- [12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, & Y. Cao, ReAct: Synergizing reasoning and acting in language models. *Proceedings of the International Conference on Learning Representations*. (2023).
- [13] E. Debenedetti, J. Zhang, M. Balunovic, L. Beurer-Kellner, M. Fischer, & F. Tramer, AgentDojo: A dynamic environment to evaluate attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems*, 37. (2024).
- [14] M. M. Yamin, B. Katt, & V. Gkioulos, Cyber security awareness, knowledge and behavior: A comparative study. *Computers & Security*, 130, 103266. (2023).
- [15] P. Shi, R. Chen, & N. Yu, LLM-augmented industrial security operations: A field deployment study. *IEEE Transactions on Network and Service Management*, 22(1), 412-425. (2025).
- [16] C. Wang, S. Liu, J. Li, & Y. Zhao, Agent memory architectures: A survey. *ACM Computing Surveys*, 57(6), 1-38. (2025).
- [17] A. Greenberg, When automation hurts: A taxonomy of erroneous defensive actions in OT environments. *Proceedings of the USENIX Security Symposium*, 2031-2048. (2024).
- [18] G. Cerar, M. Mohorcic, & C. Fortuna, On-device deep learning for industrial IoT: A review. *IEEE Internet of Things Journal*, 11(8), 13412-13435. (2024).
- [19] S. Han, J. Pool, & W. J. Dally, Distilled small language models for edge OT inference: Benchmarks and deployment notes. *Proceedings of the ACM/IEEE Symposium on Edge Computing*, 89-102. (2025).
- [20] International Electrotechnical Commission, IEC 61508-3: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems, Part 3: Software Requirements. IEC. (2010).
- [21] Y. Liu, X. Zhang, & W. Lee, Adversarial residuals: Crafting telemetry that defeats LLM-based anomaly detection. *Proceedings of the ACM AISec Workshop*, 67-80. (2025).
- [22] Dragos Inc., Year in Review: ICS/OT Cybersecurity. Annual Technical Report. (2025).
- [23] A. Conti, M. Donaldson, & L. Bossi, Measuring defensive uplift from LLM-augmented SOC tooling: A controlled study at three energy utilities. *Proceedings of the IEEE Symposium on Security and Privacy Workshops*, 184-197. (2026).
- [24] R. Naveen, & M. Salem, Operator-language interpretation in industrial incident response: An LLM case study. *Computers in Industry*, 158, 104027. (2024).
- [25] Microsoft Threat Intelligence, Adversarial Use of Generative AI in OT-Targeted Intrusions: Annual Report. Microsoft. (2025).