



## OPEN ACCESS

### EDITED BY

Bernardi Pranggono,  
Anglia Ruskin University, United Kingdom

### REVIEWED BY

Piotr Gaj,  
Silesian University of Technology, Poland  
Muhammet Çakmak,  
Giresun University, Türkiye  
Alfan Presekala,  
University of Indonesia, Indonesia

### \*CORRESPONDENCE

Heinrihs Kristians Skrodelis  
✉ Heinrihs-Kristians.Skrodelis@rtu.lv

RECEIVED 13 October 2025  
REVISED 18 January 2026  
ACCEPTED 09 February 2026  
PUBLISHED 03 March 2026

### CITATION

Skrodelis HK and Romanovs A (2026)  
Explainable hybrid intrusion detection for  
SCADA/ICS: a review and research  
agenda. *Front. Comput. Sci.* 8:1724245.  
doi: 10.3389/fcomp.2026.1724245

### COPYRIGHT

© 2026 Skrodelis and Romanovs. This is  
an open-access article distributed under  
the terms of the [Creative Commons  
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,  
distribution or reproduction in other  
forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# Explainable hybrid intrusion detection for SCADA/ICS: a review and research agenda

Heinrihs Kristians Skrodelis\* and Andrejs Romanovs

Institute of Information Technology, Riga Technical University, Riga, Latvia

**Introduction:** Supervisory Control and Data Acquisition (SCADA) and Industrial Control System (ICS) networks underpin critical infrastructure across energy, water, transportation, and manufacturing sectors. Existing intrusion detection systems face inherent trade-offs: signature-based approaches achieve low false-positive rates but cannot detect zero-day attacks, while anomaly-based methods detect novel threats but generate ambiguous alerts that burden operators and erode trust. Recent empirical studies reveal persistent practical gaps in deployment, including the difficulty of obtaining labeled attack data for supervised methods, severe hyperparameter tuning challenges for one-class classifiers, and limited integration of protocol-aware features despite the prevalence of process-aware detection. Explainability techniques remain underimplemented in industrial intrusion detection despite their potential for improving operator understanding and security workflow integration. This article presents a systematic review and research agenda for explainable hybrid intrusion detection in SCADA/ICS environments; it synthesizes evidence on detection architectures, explainability mechanisms, and deployment challenges, but does not report original experimental results.

**Methods:** A systematic literature review was conducted across major databases for the period 2014–2025, yielding 40 studies for synthesis after screening.

**Results:** The review distills five practical gaps: limited zero-day coverage, false-positive control, process awareness vs. protocol blindness, explainability for operators, and deployment complexity including concept drift. Reported performance ranges (90%–99% accuracy, 0.8%–2.1% false-positive rates) and latency benchmarks represent summaries of prior work, not new experimental findings from this study.

**Discussion:** A conceptual reference architecture is proposed that fuses protocol-aware signatures with temporal anomaly detection and feature-attribution-based explanations. An evaluation checklist and research agenda guide future prototype development and pilot deployments under latency and safety constraints.

### KEYWORDS

concept drift, explainability, hybrid detection, industrial control systems, intrusion detection, SCADA security

## 1 Introduction

Industrial control systems encompass Supervisory Control and Data Acquisition (SCADA) networks and related architectures that monitor and control physical processes across multiple critical infrastructure sectors, including energy generation and distribution, water treatment and distribution, transportation networks, and manufacturing facilities. The convergence of Information Technology (IT) and Operational Technology (OT),

combined with the proliferation of Industrial Internet of Things (IIoT) devices, has expanded the attack surface beyond traditional network perimeters. Attackers can now manipulate control logic or process signals in ways that evade IT-centric defenses, as demonstrated by incidents targeting energy grids and water treatment facilities (Cali et al., 2024; Stouffer et al., 2015).

Intrusion Detection Systems (IDS) represent a crucial defensive layer in SCADA/ICS environments. However, the literature reveals fundamental limitations in existing approaches. Signature-based IDS match known malicious patterns, achieving low false-positive rates but providing limited coverage against novel or zero-day attacks. Anomaly-based IDS learn normal operational behavior and can detect previously unseen threats, yet they frequently generate ambiguous alerts that operators struggle to interpret and trust (Sommer and Paxson, 2010; Garcia-Teodoro et al., 2009). Human-Machine Interface (HMI) operators and Security Operations Center (SOC) analysts require not only accurate detection but also actionable explanations that map alerts to physical process context.

The literature suggests that hybrid IDS architectures—combining signature-based and anomaly-based methods—can leverage complementary strengths to improve both detection coverage and false-positive control (Khan et al., 2019; Rosa et al., 2021; Kwon et al., 2022). When augmented with Explainable AI (XAI) techniques such as SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME), these systems can provide interpretable rationales for alerts, supporting analyst decision-making and building operational trust (Lundberg and Lee, 2017; Ribeiro et al., 2016; Gaspar et al., 2024; Ustebay et al., 2024).

## 1.1 Scope and practical gaps

This review identifies five persistent practical gaps that motivate the research agenda:

- G1: Limited zero-day coverage.** Signature-based IDS cannot detect attacks exploiting previously unknown vulnerabilities or novel attack vectors. While anomaly-based methods offer broader coverage, reported zero-day detection rates drop to approximately 73% even with hybrid approaches (Pan et al., 2015).
- G2: False-positive control.** Anomaly-based IDS are sensitive to operational noise and concept drift, generating false alarms that burden SOC analysts. Hybrid approaches achieve false-positive rates of 0.8%–2.1% in controlled evaluations, but field performance remains less certain (Kayode Saheed et al., 2023; Mitseva et al., 2022).
- G3: Process awareness vs. protocol blindness.** Process-aware features (e.g., plant states, physical invariants) appear in eight of 10 recent ICS IDS studies, yet protocol-aware features leveraging industrial protocol semantics (Modbus function codes, DNP3 command structures, IEC 60870-5-104 information objects, OPC UA node semantics) remain rare (Ike et al., 2023; Ghazi et al., 2025). This asymmetry leaves IDS blind to protocol-level attack vectors.

**G4: Explainability for operators.** Black-box machine learning (ML) models hinder operator understanding of alert rationales. Only two of 10 recent ICS IDS studies explicitly implement XAI techniques (Oyedotun et al., 2025; Kenmogne and Mocanu, 2024), despite evidence that SHAP/LIME attributions improve analyst trust and triage speed (Eriksson and Grov, 2022; Gaspar et al., 2024).

**G5: Deployment and drift challenges.** Supervised IDS require extensive labeled attack data impractical to collect in operational environments; one-class classifiers face severe hyperparameter tuning difficulties (Wolsing et al., 2024). Concept drift from system updates, equipment changes, and workflow modifications degrades detection accuracy over time (Wadinger and Kvasnica, 2024).

## 1.2 Research questions

This systematic review addresses the following research questions:

- RQ1:** What hybrid IDS architectures have been proposed for SCADA/ICS, and how do they balance zero-day detection with false-positive control?
- RQ2:** To what extent do current ICS IDS incorporate protocol-aware and process-aware features, and how does this affect detection performance and explainability?
- RQ3:** What deployment challenges and constraints (latency, throughput, drift, maintenance, SOC integration) have been reported, and how do hybrid/XAI-based designs address them?
- RQ4:** What evaluation methodologies and platforms are suitable as reference implementations for explainable hybrid IDS?

## 1.3 Article scope and contributions

This article is a systematic review and research agenda; it does not present new experimental results but synthesizes prior evaluations and proposes a conceptual blueprint and evaluation plan for future work. In contrast to prior ICS IDS surveys that focus on deep learning methods without addressing protocol-aware feature engineering or explainability (Altunay et al., 2021), and to XAI-IDS studies in adjacent domains such as Internet of Medical Things that demonstrate effective SHAP-based explanations but do not address ICS-specific constraints (Dakhil and Çakmak, 2025), this work uniquely integrates protocol-aware detection, hybrid fusion, XAI-based explanations, and a concrete evaluation plan tailored to SCADA/ICS requirements. The contributions are:

- A systematic synthesis (2014–2025) of 40 studies on hybrid SCADA/ICS IDS and explainable IDS.
- A comparative analysis table contrasting recent hybrid/XAI ICS IDS studies on architecture, protocols, explainability, and limitations.
- A practitioner-oriented gap matrix mapping method families to the five practical gaps (G1–G5).

- A conceptual reference architecture with mathematical formalization of fusion logic, drift handling, and SHAP-based explanation pipelines.
- A consolidated evaluation checklist and research agenda mapping each agenda item to specific gaps.

*Roadmap.* Section 2 provides background on SCADA/ICS security and IDS method families. Section 3 details the systematic review methodology. Section 4 synthesizes the evidence. Section 5 presents the gap analysis. Section 6 describes the proposed conceptual architecture. Section 7 outlines the research agenda. Section 8 presents the evaluation plan. Section 9 concludes.

## 2 Background

This section introduces key terminology, the SCADA/ICS security context, IDS method families, and explainability concepts.

### 2.1 Terminology

The following acronyms and terms are used throughout this article:

- **SCADA:** Supervisory Control and Data Acquisition—systems for monitoring and controlling industrial processes.
- **ICS:** Industrial Control System—broader category encompassing SCADA, Distributed Control Systems (DCS), and Programmable Logic Controllers (PLCs).
- **HMI:** Human–Machine Interface—operator consoles for monitoring and controlling processes.
- **PLC:** Programmable Logic Controller—industrial computers that control physical processes.
- **RTU:** Remote Terminal Unit—devices that interface with sensors and actuators in the field.
- **SOC:** Security Operations Center—organizational unit responsible for security monitoring and incident response.
- **SIEM:** Security Information and Event Management—platforms aggregating security logs and alerts.
- **IDS:** Intrusion Detection System—software or hardware that monitors for malicious activity.
- **XAI:** Explainable Artificial Intelligence—methods providing interpretable rationales for ML model decisions.
- **SHAP:** SHapley Additive exPlanations—a game-theoretic approach to feature attribution (Lundberg and Lee, 2017).
- **LIME:** Local Interpretable Model-agnostic Explanations—a perturbation-based explanation method (Ribeiro et al., 2016).

### 2.2 SCADA/ICS and security constraints

SCADA systems manage physical processes under real-time, safety, and reliability constraints. Legacy industrial protocols—including Modbus, Distributed Network Protocol 3 (DNP3),

and IEC 60870-5-104—and vendor-specific communication stacks complicate security monitoring and protocol parsing. More modern protocols such as Open Platform Communications Unified Architecture (OPC UA) offer improved security features but coexist with legacy installations. NIST SP 800-82 emphasizes defense-in-depth tailored to industrial constraints (Stouffer et al., 2015). Complementary survey work systematically reviews cybersecurity vulnerabilities in SCADA architectures (open networks, legacy components, weak segmentation) and discusses mitigation measures such as encryption, network zoning, IDS and SIEM integration, and the growing use of AI/ML for threat detection (Skrodelis H. et al., 2024). Table 1 summarizes a Purdue-model reference architecture with monitoring points across zone boundaries. To support real-world adoption, the proposed monitoring placement can be explicitly mapped to established industrial security frameworks: Purdue levels motivate where passive sensors and analytics reside (field/SCADA/DMZ/enterprise), and IEC 62443 zone-and-conduit concepts motivate placing sensors at conduit boundaries to observe inter-zone traffic in a segmentation-aware manner; this complements NIST SP 800-82's defense-in-depth monitoring recommendations (Stouffer et al., 2015; Eißler et al., 2023).

### 2.3 IDS method families for ICS

Four primary IDS method families are considered in this review:

**Signature-based IDS** match network traffic or system events against databases of known malicious patterns. They achieve low false-positive rates but provide limited coverage against unknown attacks (Stouffer et al., 2015).

**Anomaly-based IDS** learn models of normal behavior from training data and flag deviations as potential attacks. They can detect novel threats but are susceptible to false positives from legitimate operational variations and concept drift (Sommer and Paxson, 2010; Garcia-Teodoro et al., 2009; Mitchell and Chen, 2014). Prior surveys on deep learning for SCADA anomaly detection catalog autoencoder, LSTM, and CNN architectures but largely omit protocol-aware feature engineering and XAI integration (Altunay et al., 2021).

**Specification-based IDS** encode expected behavior as formal invariants derived from process design or protocol specifications. They offer interpretable detection logic but require significant domain expertise to develop and maintain (Umer et al., 2017).

**Hybrid IDS** combine multiple detection methods—typically signature and anomaly approaches—to leverage complementary strengths. Recent ICS-focused hybrid IDS report accuracies of 90%–99% with false-positive rates of 0.8%–2.1% (Khan et al., 2019; Kayode Saheed et al., 2023; Kwon et al., 2022).

The term “fusion” denotes the mechanism within a hybrid IDS by which multiple detection channels or decision sources are combined (e.g., serial vs. parallel architectures, score-level vs. alert-level fusion).

TABLE 1 Purdue-model ICS reference architecture showing zones (enterprise to field), representative assets, passive monitoring points at zone boundaries, and typical attack surfaces.

Zone (level)	Representative assets	Monitoring point	Typical attack surface
Enterprise (L4)	IT systems; Cloud/SaaS; SOC/SIEM; AD/identity	SIEM (northbound)	Phishing or remote access
Industrial DMZ (L3.5)	Firewalls and proxies; replication historian; jump server	TAP at DMZ	Pivot across DMZ
SCADA & operations (L3–L2)	HMI; SCADA server; historian; engineering workstation	SPAN at SCADA	Protocol misuse or fuzzing
Basic control (L1)	PLCs; RTUs; IEDs or remote I/O	TAP at L1	Firmware or configuration tampering
Process & field (L0)	Sensors, actuators, drives/valves	— (typically observed via L1)	Physical access

## 2.4 Explainability in IDS

Explainable AI (XAI) provides *post-hoc* or intrinsic interpretability for machine learning models. SHAP computes feature attributions based on Shapley values from cooperative game theory, providing locally faithful explanations (Lundberg and Lee, 2017). LIME generates local surrogate models by perturbing input features (Ribeiro et al., 2016). In SOC contexts, explanations have been shown to improve analyst confidence and triage prioritization, though the relationship between explanation quality and decision accuracy remains nuanced (Eriksson and Grov, 2022; Weerts et al., 2019; Gaspar et al., 2024). Recent work in adjacent domains demonstrates that XAI-enhanced IDS can achieve high detection accuracy with interpretable outputs; for example, XGBoost with SHAP explanations reaches over 99% accuracy in Internet of Medical Things settings (Dakhil and Çakmak, 2025). For ICS applications, explainability must map model features to process context that operators understand, linking network-level observations to physical plant states and expected operational sequences (Neupane et al., 2022; Fung et al., 2024).

## 3 Methodology

This section describes the systematic review methodology following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines adapted for engineering research.

### 3.1 Search strategy and selection

We conducted a systematic search in five electronic databases—IEEE Xplore, ACM Digital Library, Scopus, Web of Science and Google Scholar—using combinations of terms related to industrial control systems, SCADA/OT security, hybrid intrusion detection, and explainable AI. The search was limited to peer-reviewed articles, full conference papers, and English-language publications in 2010–May 2025. The final search strings, applied consistently across databases with minor syntax adjustments, are described below.

**Search strings:** The core query combined Boolean terms across three concept groups:

1. *Domain:* (ICS OR SCADA OR “industrial control” OR “operational technology”)
2. *Detection:* (“intrusion detection” OR IDS OR “anomaly detection”)
3. *Approach:* (hybrid OR fusion OR correlation OR ensemble) OR (explain\* OR SHAP OR LIME OR XAI)

The full query was: (Group 1) AND (Group 2) AND (Group 3).

Across all databases, the final search queries returned 412 records, to which we added 23 records identified through backward and forward citation chasing of three seed papers on ICS datasets and hybrid IDS (Conti et al., 2021; Stouffer et al., 2015; Neupane et al., 2022). After removing 137 duplicates, 298 unique records remained for title and abstract screening. Of these, 187 were excluded as off-topic (e.g., purely IT-network IDS), non-empirical, or lacking an IDS component, leaving 111 full-text articles for eligibility assessment. At the full-text stage, 71 articles were excluded for reasons such as insufficient ICS/OT context, lack of a hybrid or XAI element, or missing evaluation metrics, resulting in 40 studies included in the final synthesis (Figure 1). Table 2 summarizes the number of records retrieved from each database and the reductions at each screening step (deduplication, title/abstract screening, and full-text screening). Of the 40 included studies, 36 were identified through database searching and 4 through backward and forward citation chasing.

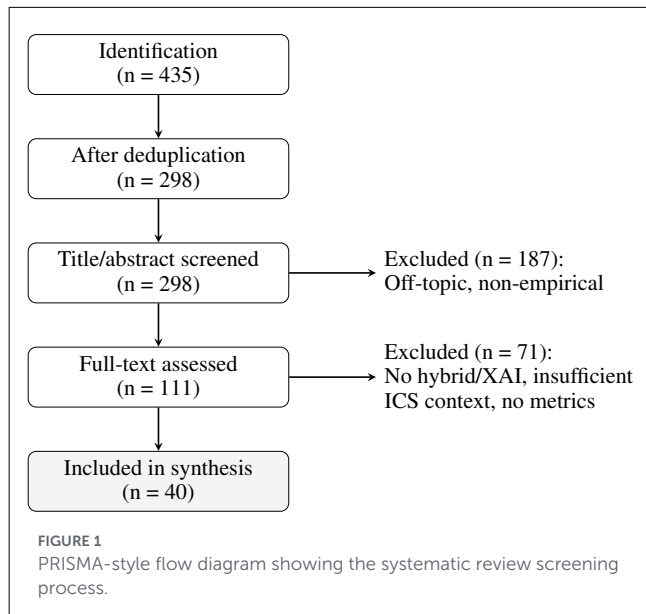
### 3.2 Inclusion and exclusion criteria

#### Inclusion criteria:

- Focus on SCADA, ICS, or OT network security.
- Proposes, implements, or evaluates hybrid IDS (combining signature/anomaly/specification methods) or XAI techniques for IDS.
- Reports empirical evaluation with quantitative performance metrics (accuracy, precision, recall, F1, false-positive rate, detection latency).
- Peer-reviewed publication (journal or conference).

#### Exclusion criteria:

- Purely IT-focused IDS without ICS/SCADA data or context.
- Surveys or reviews without new empirical evidence or system contributions.



- Non-English publications.
- Theses, book chapters, or non-peer-reviewed materials.

### 3.3 Screening and selection process

Figure 1 illustrates the PRISMA-style screening process:

1. **Identification:** database searches yielded 412 records; 23 additional records were identified through backward/forward citation chasing from seed papers (Conti et al., 2021; Stouffer et al., 2015; Neupane et al., 2022).
2. **Deduplication:** after removing duplicates, 298 unique records remained.
3. **Title/abstract screening:** records were screened against inclusion criteria; 187 records were excluded as off-topic (purely IT IDS, non-detection security topics, or non-empirical).
4. **Full-text assessment:** 111 full-text articles were assessed for eligibility; 71 were excluded (no hybrid/XAI component, insufficient ICS context, or no quantitative evaluation).
5. **Included studies:** 40 studies were included in the final synthesis.

### 3.4 Quality assessment

Included studies were assessed using a lightweight quality checklist:

- **Dataset realism:** use of recognized ICS datasets (e.g., SWaT, WADI, HAI) or real-world testbeds vs. purely synthetic data.
- **Metric completeness:** reporting of multiple metrics (accuracy, precision, recall, F1, false-positive rate) vs. single metrics.
- **ICS specificity:** clear focus on ICS protocols, processes, or operational constraints vs. generic network IDS.
- **Reproducibility:** availability of code, datasets, or detailed configuration descriptions.

Studies were not excluded based on quality scores but were weighted in the synthesis according to evidence strength. Of the 40 included studies, 18 were rated high quality (meeting all four criteria), 15 medium quality (meeting two or three criteria), and seven low quality (meeting one or fewer criteria). Full-text availability was limited for six of 40 studies (abstract-only access), constraining detailed analysis for those sources.

### 3.5 Data extraction

For each included study, the following data were extracted: hybrid architecture components, protocols covered, explainability mechanisms, detection performance metrics, evaluation methodology, deployment considerations, and identified limitations. Supplementary Tables 8, 9 provides detailed characteristics of all 40 included studies.

## 4 Findings

This section synthesizes evidence from the 40 included studies across four thematic areas corresponding to the research questions.

### 4.1 Hybrid IDS architectures (RQ1)

Three primary architectural patterns emerge from the literature:

**Sequential (serial) architectures** apply detection methods in stages, where earlier components filter or preprocess data before later methods perform deeper analysis. For example, statistical filtering followed by deep learning analysis reduces computational burden while maintaining detection accuracy (Kwon et al., 2022; Feng et al., 2017). The Bloom filter plus LSTM architecture achieves classification times of 0.03 ms by screening packets before temporal analysis (Feng et al., 2017).

**Parallel (ensemble) architectures** combine multiple concurrent detectors through voting or aggregation mechanisms. Ensemble methods using bagging, stacking, and AdaBoost with Naive Bayes and SVM classifiers achieve 99% accuracy with perfect precision and recall on some datasets (Kayode Saheed et al., 2023). However, computational costs are higher than serial designs.

**Hierarchical architectures** incorporate preprocessing, feature optimization, dimensionality reduction, and multiple detection layers. The most sophisticated designs combine autoencoders for anomaly detection, CNNs for feature extraction, and hybrid ResNet-LSTM for temporal pattern learning (Almalawi et al., 2025; Oyedotun et al., 2025).

**Performance summary:** Hybrid IDS report accuracies of 90.4%–99.5% and false-positive rates of 0.8%–2.1% where reported (Khan et al., 2019; Kayode Saheed et al., 2023; Pan et al., 2015). However, direct comparisons to pure signature-based or anomaly-based methods remain scarce; only two studies quantify improvements, showing precision gains of 0.008, recall improvements of 0.067, and F1-score increases of 0.039 over autoencoder-only approaches (Kwon et al., 2022). Zero-day

TABLE 2 Search results and screening per database.

Database/source	Initial hits	After keyword filters	After title/abstract screen	After full-text screen	Included
IEEE Xplore	150	90	28	11	11
ACM Digital Library	45	28	8	3	3
Scopus	220	145	36	13	13
Web of Science	190	103	24	7	7
Google Scholar	110	46	10	2	2
Citation chasing (backward/forward)	23	23	5	4	4
<b>Total</b>	<b>738</b>	<b>435</b>	<b>111</b>	<b>40</b>	<b>40</b>

Bold values indicate the final number of studies retained at each screening stage (i.e., the “Included” column).

detection rates reach 99%–100% for known attack patterns but drop to 73.43% for novel attack scenarios (Pan et al., 2015). This degradation likely reflects class imbalance in training data, overfitting to known attack signatures, and the inherent difficulty of generalizing to truly unseen attack vectors without protocol-semantic or physics-based priors.

## 4.2 Protocol and process awareness (RQ2)

A striking asymmetry characterizes current ICS IDS approaches: process-aware features are common (eight of 10 recent studies), while protocol-aware features are virtually absent (two of 10 studies with minimal integration) (Ike et al., 2023; Ghazi et al., 2025).

**Process awareness** encompasses plant state modeling, physics-based constraints, actuation-sensing relationships, and temporal dynamics. The most sophisticated approaches employ physics-informed neural networks that embed generic knowledge of inertial process dynamics, achieving 98.3% accuracy with 0.8% false positives—a 20 percentage point improvement over baseline methods (Ike et al., 2023). State-aware invariants derive tighter bounds specific to each system state, achieving 2% false-positive rates (Abbas et al., 2024).

**Protocol awareness** leveraging Modbus function codes, DNP3 command structures, IEC 60870-5-104 information objects, or OPC UA semantics remains rare. Only Ghazi et al. (2025) demonstrate substantive OPC UA integration, achieving F1-score improvements of 2.27% when incorporating protocol-specific temporal dependencies. This gap leaves IDS vulnerable to attacks exploiting protocol semantics. The scarcity of protocol-aware features can be attributed to the complexity of building and maintaining robust parsers for legacy industrial protocols, the lack of publicly available labeled datasets with protocol-level annotations, and the heterogeneity of vendor-specific protocol extensions.

Existing specification- and invariant-based IDS partly address protocol blindness by constraining either process evolution or permitted protocol interactions. For example, invariant mining for water treatment and distribution testbeds (Feng et al., 2019; Song et al., 2024) and automatic specification extraction from device

documentation (Esquivel-Vargas et al., 2017) encode expected system behavior beyond raw packet fields. Stateful protocol models for Modbus, IEC-104, and OPC UA similarly capture function-code sequences and protocol states using deterministic finite automata (DFAs), Markov chains, and probabilistic automata (Faisal and Sitnikova, 2016; Kleinmann and Wool, 2016; Matoušek et al., 2021; Burgetova et al., 2021; Ghazi et al., 2025). However, these approaches are typically deployed as standalone detectors and rarely fuse process invariants, protocol state, and data-driven anomaly scores in a single explainable pipeline—precisely the integration targeted by the proposed protocol-aware signature plus temporal anomaly architecture. Protocol-abstraction layers such as IPAL (Wolsing et al., 2022) further demonstrate that protocol-specific detectors can be lifted to a shared representation; the proposed architecture can be seen as a concrete way to exploit such abstractions in a hybrid IDS.

**Dataset limitations:** the heavy reliance on SWaT and WADI datasets in the literature introduces evaluation biases. These datasets feature relatively high attack-to-normal ratios compared to operational networks, cover only a narrow subset of industrial protocols (primarily process-level signals without rich protocol-layer annotations), and represent a single industrial domain (water treatment). Reported detection rates may therefore overestimate performance in heterogeneous, low-attack-frequency production environments. More broadly, many ICS benchmarks remain scenario-specific and often omit protocol-semantic annotations and diverse operating modes, so conclusions should be validated across datasets and operating conditions rather than inferred from single-testbed performance alone. For explainable IDS in particular, ground-truth rationales are rarely available, so benchmark evaluation should emphasize explanation fidelity/stability and human-grounded utility rather than assuming *post-hoc* explanations are correct (Carvalho et al., 2019; Jacovi and Goldberg, 2020).

## 4.3 Deployment challenges and XAI solutions (RQ3)

The most fundamental deployment challenge is the tension between data requirements and hyperparameter tuning. Supervised

IDS require extensive labeled attack data impractical to collect in operational environments (Wolsing et al., 2024). One-class classifiers that train only on benign data face severe hyperparameter tuning difficulties that can lead to devastating performance penalties (Wolsing et al., 2024). In operational ICS environments, data provenance and quality issues—sensor noise/quantization, partial logging or packet loss, and time-synchronization inconsistencies across PLC/HMI/historian sources—can degrade both detection accuracy and explanation reliability. When telemetry is corrupted, signature-based IDS may miss matches due to lost context, while ML-based anomaly detectors can inflate false positives by treating noise artifacts as deviations from learned baselines (Khraisat et al., 2019). Accordingly, evaluation should include targeted robustness checks (noise injection, missing-data/packet-loss simulation, and clock-skew perturbations) to characterize stability under realistic data-quality conditions.

**Latency constraints:** comparative testing shows Suricata achieving 2.5 s median detection latency vs. Snort's 5.1 s (Waagsnes and Ulltveit-Moe, 2018). Real-time SHAP/LIME computation adds overhead that must be considered in latency budgets; GPU-accelerated TreeSHAP and FastSHAP methods add approximately 10–50 ms per alert (Yang, 2021; Jethani et al., 2021).

**Concept drift:** ICS normal behavior evolves due to device updates, equipment replacements, and workflow modifications. Adaptive mechanisms including automatic threshold tuning and dynamic parameter updates address drift without human intervention (Wadinger and Kvasnica, 2024; Abdelaty et al., 2021).

**XAI implementation:** only Oyedotun et al. (2025) explicitly integrate SHAP and LIME in an ICS IDS context, achieving 92% accuracy with AUC of 0.97. SOC studies show that SHAP attributions improve analyst trust and triage speed (Eriksson and Grov, 2022; Gaspar et al., 2024), though the evidence base for ICS-specific operator studies remains limited. Explainability requirements also vary by ICS role and decision horizon: operators making seconds-to-minutes triage decisions need concise, prescriptive rationales (e.g., top contributing protocol/process features plus confidence), whereas control engineers diagnosing incidents over minutes-to-hours benefit from deeper causal chains and counterfactual “what-if” explanations tied to process variables; SOC analysts preparing reports can use richer narratives and exemplar cases. Evidence from decision-support studies suggests explanation effectiveness and trust calibration improve when explanation format and depth are matched to user expertise and time constraints (Dutta et al., 2025).

#### 4.4 Platforms and reference implementations (RQ4)

Established open-source platforms—Zeek, Suricata, Snort, and Wazuh—serve as the primary bases for ICS monitoring research (Joy et al., 2024; Waagsnes and Ulltveit-Moe, 2018; Haas et al., 2020). The Elastic stack with Kibana dashboards provides visualization capabilities. However, no widely adopted reference implementation bundles packet capture, ICS protocol

parsing, and XAI-enhanced dashboarding in a single open-source, containerized stack.

Malcolm, an open-source network traffic analysis suite maintained by CISA, integrates Arkime, Zeek, Suricata, and OpenSearch/Dashboards (CISA, 2025). It provides a pragmatic baseline for prototyping and evaluation, supporting ICS protocol parsing via Zeek ICSNPP plugins (Modbus, DNP3, S7, IEC 60870-5-104, OPC UA). The zeek-osquery platform demonstrates production readiness with 96% accuracy in network flow attribution and scalability to over 870 hosts (Haas et al., 2020).

#### 4.5 Comparative analysis of recent studies

Table 3 presents a comparative analysis of representative recent hybrid and XAI-enhanced ICS IDS studies (2019–2025), highlighting architectural patterns, protocol coverage, explainability mechanisms, and identified limitations.

The comparison reveals that protocol-specific detection remains underdeveloped, XAI techniques are underimplemented, and most evaluations rely on laboratory datasets rather than field deployments. The proposed research agenda and conceptual architecture address these limitations by: (1) adding protocol-aware features to the signature channel, (2) integrating SHAP-based explanations in the fusion layer, and (3) defining deployment-focused evaluation criteria including drift handling and latency budgets.

### 5 Gap analysis and open challenges

This section elaborates on the five practical gaps introduced in Section 1, synthesizing evidence from the included studies.

#### 5.1 G1: limited zero-day coverage

Signature-based IDS match known attack patterns with high precision but cannot detect novel threats. Anomaly-based methods offer broader coverage but face a fundamental trade-off: models trained on benign data may flag legitimate operational variations as attacks (false positives) or miss subtle attack patterns that resemble normal behavior (false negatives).

Hybrid approaches partially address this gap. Pan et al. (2015) report 90.4% overall accuracy for mixed scenarios including known attacks, disturbances, and normal operations, but zero-day detection accuracy drops to 73.43% under 10-fold cross-validation on unseen attack patterns. This suggests that even hybrid methods face substantial challenges with truly novel attack vectors.

#### 5.2 G2: false-positive control

False-positive rates are critical for SOC workflows: excessive alerts create analyst fatigue and erode trust in the IDS. Reported false-positive rates range from 0.8 to 2.1% in controlled evaluations

TABLE 3 Comparative analysis of recent hybrid/XAI ICS IDS studies (2019–2025).

Study/year	Architecture	Protocols	Explainability	Identified limitation
TwinSec-IDS (Krishnaveni et al., 2024)	SDN-Digital Twin; Bi-GRU-CNN/LSTM ensemble	Not specified	Feature importance (MI, chi-square)	Development stage unspecified; limited protocol coverage
Oyedotun et al. (2025)	Multimodal: CNN + LSTM + autoencoder	Modbus, DNP3, OPC-UA (mentioned)	SHAP, LIME	92% accuracy; precision 0.49 for attack class due to imbalance
Wolsing et al. (2024)	Ensemble unsupervised IIDSs	Generic (SWaT, WADI)	None	Hyperparameter tuning without attacks is difficult; lab-only evaluation
Ghazi et al. (2025)	Markov chain + ML; second-order memory	OPC UA	SHAP, causal inference (PC algorithm)	F1 improvement 2.27%; single protocol focus
Ike et al. (2023)	Physics-informed neural network	Generic (11 industrial processes)	Causal explanations (operations to effects)	Preprint; limited protocol semantics
Khan et al. (2019)	Sequential: preprocessing + bloom filter + instance-based learner	Modbus (gas pipeline)	None	97% accuracy; no XAI; limited protocol diversity
Kwon et al. (2022)	Sequential: statistical filtering + composite autoencoder	Generic (SWaT)	None	Modest improvements (+0.008 precision, +0.067 recall) over pure autoencoder

(Pan et al., 2015; Kayode Saheed et al., 2023), but laboratory conditions may not reflect operational noise.

The tension between data requirements and hyperparameter tuning exacerbates this challenge. Wolsing et al. (2024) demonstrate that one-class classifiers require careful hyperparameter selection, and configurations rarely transfer across different ICS contexts. Poor choices can lead to devastating performance degradation.

### 5.3 G3: process awareness vs. protocol blindness

Process-aware IDS incorporating plant state models, physics-based constraints, and actuation-sensing relationships achieve superior detection performance. Ike et al. (2023) report 98.3% accuracy with 0.8% false positives using physics-informed neural networks—a 20 percentage point improvement over baseline methods.

However, protocol-aware features remain rare. Industrial protocols (Modbus, DNP3, IEC 60870-5-104, OPC UA) encode semantic information—function codes, command types, information object addresses—that can distinguish legitimate operations from protocol-level attacks. The near-complete absence of protocol-aware features (present in only two of 10 recent studies) represents a critical vulnerability: attacks exploiting protocol semantics may evade detection entirely. As detailed in Section 4, this protocol blindness stems from parser complexity, dataset limitations, and vendor heterogeneity; the proposed architecture addresses it by treating protocol-semantic features as first-class inputs to both signature and anomaly channels.

Existing specification- and invariant-based IDS partly address protocol blindness by constraining either process evolution or

permitted protocol interactions. For example, invariant mining for water treatment and distribution testbeds (Feng et al., 2019; Song et al., 2024) and automatic specification extraction from device documentation (Esquivel-Vargas et al., 2017) encode expected system behavior beyond raw packet fields. Stateful protocol models for Modbus, IEC-104, and OPC UA similarly capture function-code sequences and protocol states using deterministic finite automata (DFAs), Markov chains, and probabilistic automata (Faisal and Sitnikova, 2016; Kleinmann and Wool, 2016; Matoušek et al., 2021; Burgetova et al., 2021; Ghazi et al., 2025). However, these approaches are typically deployed as standalone detectors and rarely fuse process invariants, protocol state, and data-driven anomaly scores in a single explainable pipeline—precisely the integration targeted by the proposed protocol-aware signature plus temporal anomaly architecture. Protocol-abstraction layers such as IPAL (Wolsing et al., 2022) further demonstrate that protocol-specific detectors can be lifted to a shared representation; the proposed architecture can be seen as a concrete way to exploit such abstractions in a hybrid IDS.

### 5.4 G4: explainability for operators

Black-box machine learning models hinder operator understanding and trust. SOC studies demonstrate that SHAP/LIME attributions improve analyst confidence and triage speed (Eriksson and Grov, 2022; Gaspar et al., 2024), but ICS-specific operator studies remain scarce.

Only two recent ICS IDS studies explicitly implement XAI techniques: Oyedotun et al. (2025) integrate SHAP and LIME, and Ghazi et al. (2025) combine SHAP with causal inference. This gap between research emphasis on interpretability and practical implementation motivates the XAI layer in the proposed architecture.

## 5.5 G5: deployment and drift challenges

Beyond explainability, broader deployment challenges limit the transition from laboratory prototypes to operational systems. Supervised IDS require labeled attack data impractical to collect in operational environments; one-class classifiers face hyperparameter tuning difficulties (Wolsing et al., 2024). Concept drift from system updates, equipment changes, and workflow modifications degrades detection accuracy over time (Wadinger and Kvasnica, 2024).

Latency constraints compound deployment challenges. Waagsnes and Ulltveit-Moe (2018) report Suricata achieving 2.5 s median detection latency vs. Snort's 5.1 s. Adding SHAP computation introduces additional overhead (10–50 ms per alert with GPU acceleration) (Yang, 2021; Jethani et al., 2021) that must be budgeted in real-time systems.

## 5.6 Gap matrix summary

Tables 4, 5 summarize method-family trade-offs and cross-cutting evidence for each gap.

Despite a decade of progress, the gap matrix also helps explain why hybrid and explainable IDS designs have only rarely translated into sustained deployments. Most reported gains are derived from laboratory datasets such as SWaT, WADI, and small protocol-specific corpora, with limited evidence from long-running field traces or multi-site evaluations. Protocol coverage remains uneven—Modbus and generic TCP/UDP traffic are well represented, whereas IEC-104, OPC UA, and vendor-specific stacks are underexplored—so many proposed methods cannot be dropped into heterogeneous plants without substantial engineering. Few studies include operator-in-the-loop experiments, leaving usability, workload, and governance questions unanswered; and only a minority explicitly address concept drift, model maintenance, or integration into existing SOC workflows. The research agenda in Section 7 is therefore structured to close this lab-to-field gap: Items 1–2 and 7 target prototype integration and benchmark evaluation on multiple ICS datasets; Items 3, 4, and 6 address explainability UX, drift-aware optimization, and SOC/ATT&CK alignment; and Item 8 elevates specification and stateful protocol models to first-class fusion inputs. Together, these steps are intended to turn the abstract capabilities summarized in Tables 4, 5 into deployable, maintainable systems in real SCADA/ICS environments.

## 6 Proposed conceptual architecture

This section presents a conceptual reference architecture for explainable hybrid IDS in SCADA/ICS environments. The architecture is a blueprint derived from the gap analysis and literature synthesis; it does not represent a fully implemented system. Similar components (signature IDS, anomaly models, SHAP explainers, ICS parsers) exist in prior work; the contribution here is organizing them into an evidence-based, explainable hybrid

IDS framework with mathematical formalization and an evaluation plan for future empirical validation.

## 6.1 Architecture overview

Figure 2 presents four core elements:

1. **Signature engine:** a protocol-aware signature IDS (e.g., Snort/Suricata/Zeek with ICS-specific rulesets) that matches known attack patterns and protocol violations.
2. **Anomaly engine:** time-series and multivariate anomaly detection models (e.g., autoencoders, LSTM networks, isolation forests) trained on benign operational data.
3. **Alert fusion:** a fusion layer that combines signature and anomaly outputs through deduplication, temporal correlation, and score aggregation.
4. **XAI layer:** a SHAP-based explanation module that computes feature attributions for anomaly alerts and attaches interpretable rationales for the operator console.

The operator console presents fused alerts with concise explanations mapped to process context (Figure 3).

## 6.2 Mathematical formalization

This section provides mathematical formulations for three key mechanisms: fusion logic, drift detection, and SHAP-based explanation pipelines. These formulations represent planned logic for future implementation.

### 6.2.1 Fusion logic

Let  $A(x) \in [0, 1]$  denote the anomaly score from the anomaly engine for event  $x$ , and let  $S(x) \in [0, 1]$  denote the signature confidence score (1 if a signature matches, 0 otherwise, or a normalized match confidence). The fused detection score  $F(x)$  is computed as:

$$F(x) = \alpha \cdot A(x) + (1 - \alpha) \cdot S(x) \quad (1)$$

where  $\alpha \in [0, 1]$  is a tunable weight parameter balancing anomaly and signature contributions. An alert is generated when  $F(x) > \tau_{\text{alert}}$ , where  $\tau_{\text{alert}}$  is a configurable threshold. In practice, the fusion weight  $\alpha$  can be tuned via grid search or Bayesian optimisation on a held-out validation set, optimizing for a target trade-off between detection rate and false-positive rate.

Alternative fusion mechanisms include:

- **Maximum rule:**  $F(x) = \max(A(x), S(x))$ —conservative, triggers on either channel.
- **Dempster-Shafer combination:** for evidence-theoretic fusion under uncertainty (Sahu and Siano, 2021).

TABLE 4 Gap matrix (Part A): practitioner gaps vs. method-family ratings (H, high capability; M, medium; L, low) with supporting evidence.

Gap	Signature	Anomaly	Hybrid (+XAI)
G1: zero-day coverage	L; rules match known patterns (Stouffer et al., 2015)	H; detects deviations; suffers drift (Wadinger and Kvasnica, 2024)	H; fusion improves coverage (Khan et al., 2019; Kwon et al., 2022)
G2: false positives	H; specific rules keep FP low (Stouffer et al., 2015)	L/M; sensitive to noise (Wolsing et al., 2024)	M; fusion filters noise; XAI aids triage (Gaspar et al., 2024)
G3: process awareness	L; network rules lack context (Stouffer et al., 2015)	M; temporal/state features help (Stefanidis and Voyiatzis, 2016)	M/H; invariants + XAI mapping (Umer et al., 2017; Ike et al., 2023)
G4: explainability	M; rule names explain matches (Stouffer et al., 2015)	L; black-box models (Weerts et al., 2019)	H; SHAP/LIME raise trust (Gaspar et al., 2024; Oyedotun et al., 2025)
G5: deployment	L; mature tooling (Stouffer et al., 2015)	M/H; drift monitoring needed (Wolsing et al., 2024)	H; fusion + XAI adds complexity (Wolsing et al., 2024)

Bold text indicates the highest capability rating (H = High) for each gap across method families.

TABLE 5 Gap matrix (Part B): cross-cutting evidence notes across gaps.

Gap	Dataset realism	Protocol coverage	XAI latency	Human-factors evidence
G1: zero-day	Most gains on SWaT/WADI (Goh et al., 2016; Mitseva et al., 2022)	Modbus/DNP3 strongest (Bhatia et al., 2014; Rodofile et al., 2017)	10–50 ms with GPU (Yang, 2021)	ICS studies scarce (Weerts et al., 2019)
G2: false positives	Lab FPRs optimistic (Mitseva et al., 2022)	IEC-104 gaps (Radoglou-Grammatikis et al., 2019)	Batchable (Jethani et al., 2021)	Triage faster with XAI (Eriksson and Grov, 2022)
G3: process	Water plants dominate (Taormina et al., 2018)	IEC-104 underused (Radoglou-Grammatikis et al., 2019)	Sub-second feasible (Yang, 2021)	Few controlled studies (Weerts et al., 2019)
G4: explain	Lab datasets (Mitseva et al., 2022)	Protocol-aware aids readability (Bhatia et al., 2014)	FastSHAP amortizes (Jethani et al., 2021)	Trust/triage improve (Gaspar et al., 2024)
G5: deploy	Repro kits rare (Mitseva et al., 2022)	Multi-protocol maintenance (Radoglou-Grammatikis et al., 2019)	GPU methods acceptable (Yang, 2021)	Training needed (Eriksson and Grov, 2022)

The weighted fusion rule in Equation 1 should be seen as a conceptual baseline rather than a fixed choice. Prior IDS work has successfully applied evidence-theoretic fusion, particularly Dempster–Shafer (D–S) theory, to combine heterogeneous sensors, including in ICS power networks (Wang and Ghorbani, 2004; Sahu and Davis, 2021), and Bayesian multi-model fusion across classifiers (Katar et al., 2006; Xu and Eckert, 2009). The proposed architecture is compatible with these schemes: the signature, anomaly, and specification-based scores can be treated as evidence sources and fused using D–S or learned logistic/ML-based gating, enabling future work to systematically compare conservative (max/OR), evidence-theoretic, and learned fusion strategies under adversarial conditions.

Threshold  $\tau_{\text{alert}}$  is calibrated during deployment using baseline traffic to achieve target false-positive budgets.

## 6.2.2 Drift detection and threshold update

To handle concept drift, continuous monitoring detects distribution shifts in input features or model residuals. Inspired by Wadinger and Kvasnica (2024), the following drift detection and threshold update rule is employed:

Let  $\mu_w$  and  $\sigma_w$  denote the mean and standard deviation of anomaly scores over a sliding window of  $w$  recent events. Drift is flagged when:

$$|\mu_w - \mu_{\text{baseline}}| > k \cdot \sigma_{\text{baseline}} \quad (2)$$

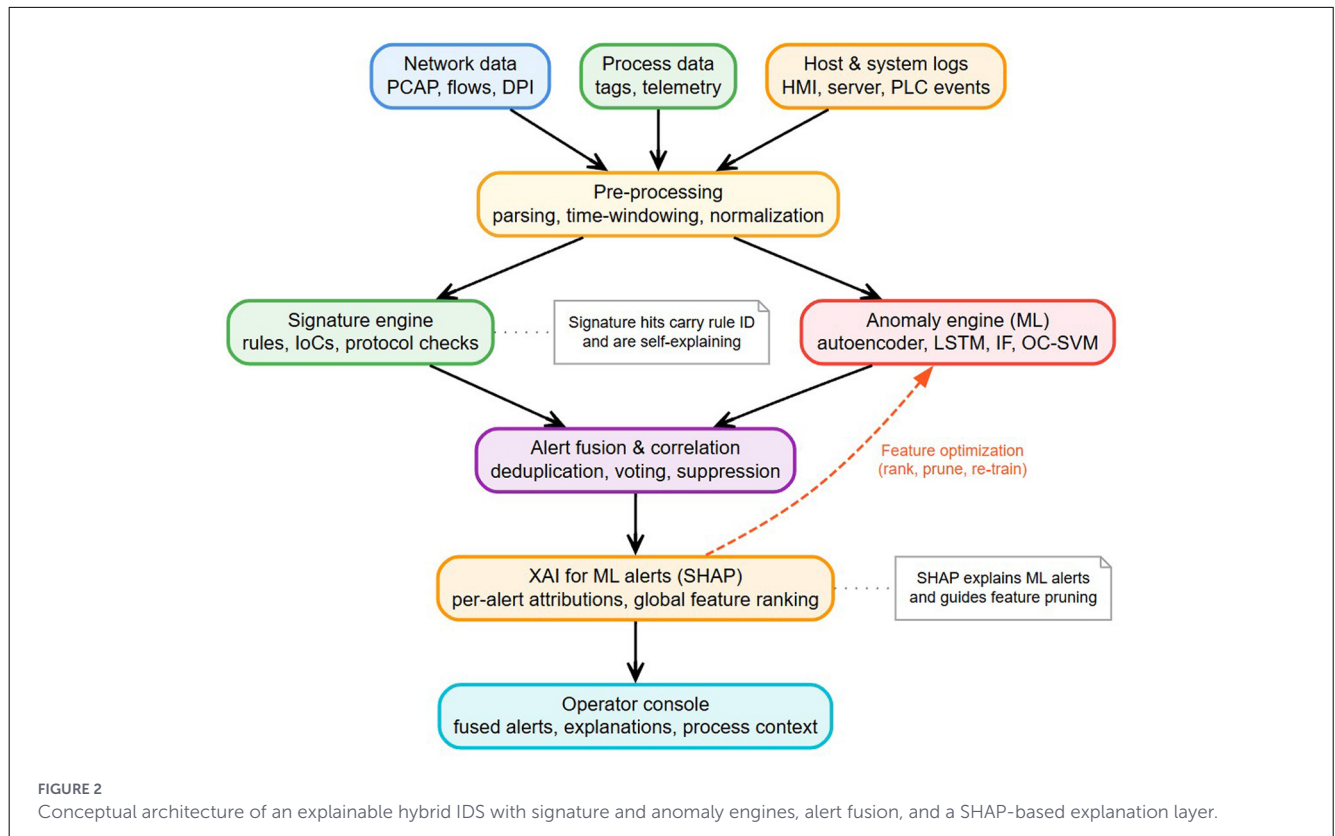
where  $\mu_{\text{baseline}}$  and  $\sigma_{\text{baseline}}$  are computed from initial deployment data, and  $k$  (e.g.,  $k = 3$ ) controls sensitivity. The window length  $w$  should be set to span multiple PLC scan cycles (e.g.,  $w = 500\text{--}5,000$  events, corresponding to tens of seconds to several minutes depending on polling rates) so that transient process fluctuations are averaged out while genuine distribution shifts are detected promptly.

Upon drift detection, the threshold  $\tau_{\text{alert}}$  is recalibrated:

$$\tau_{\text{alert}}^{\text{new}} = \mu_w + \lambda \cdot \sigma_w \quad (3)$$

where  $\lambda$  is a multiplier (e.g.,  $\lambda = 2$ ) balancing sensitivity and false-positive rate. Optional warm-start retraining of anomaly models may be triggered.

The simple mean/variance-based drift sketch in Equations 2, 3 is intended as an illustrative abstraction. In practice, the drift monitor can be instantiated with established streaming



detectors such as KS-test-based methods and ADWIN-style adaptive windows (Bharani et al., 2024b), KL-divergence monitors that adjust ROC thresholds (Shetty et al., 2011), knowledge-distillation schemes that reweight models under distribution shift (Xu et al., 2023), and adaptive sliding-window thresholding on anomaly scores (Clark et al., 2018). A key requirement in the ICS context is security-aware drift handling: updates to models and thresholds must be conservative under suspected attack or poisoning, rather than purely driven by recent data.

### 6.2.3 SHAP-based explanation pipeline

Algorithm 1 describes the SHAP-based explanation pipeline for anomaly alerts.

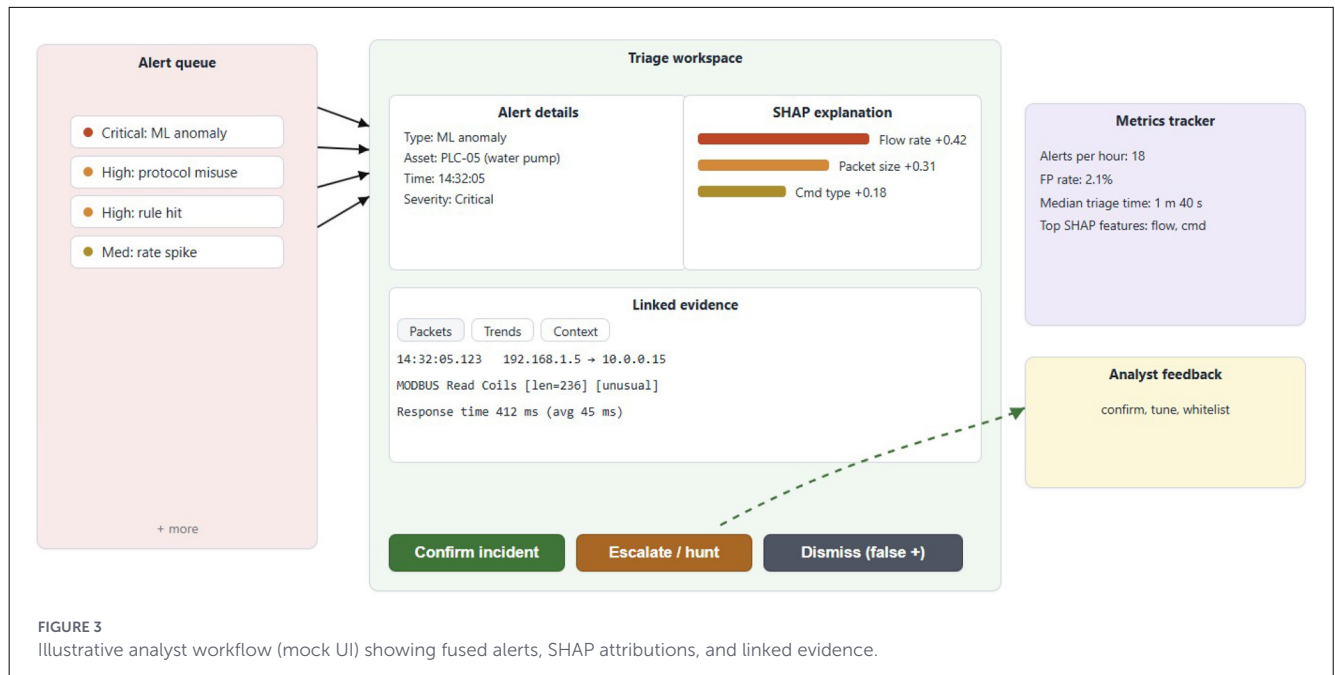
The `GenerateNarrative` function maps feature attributions to process context [e.g., “Elevated Modbus write coil rate (+0.42 contribution) on PLC-05 during idle state”]. To support this mapping, the implementation maintains a lightweight lookup table that associates each model feature index with its corresponding process variable (e.g., sensor tag, actuator name) and protocol field (e.g., Modbus function code, DNP3 data object); this table is populated during feature engineering and allows the narrative generator to produce human-readable explanations without runtime inference. Feasibility is supported by ICS-focused XAI studies demonstrating SHAP computation overhead of 10–50 ms with GPU acceleration (Yang, 2021; Jethani et al., 2021; Oyedotun et al., 2025).

## 6.3 Design choices

**Temporal fusion.** Adaptive time windows aligned to process phases and queue-graph correlation link host/service events to reduce alert duplication and chain related alerts (Bateni and Baraani, 2013; Koucham et al., 2022).

**Explainers.** Tree-based models (Random Forest, XGBoost) use GPU-accelerated TreeSHAP (Yang, 2021); deep models (autoencoders, LSTMs) use FastSHAP (Jethani et al., 2021) to meet sub-second latency budgets.

In practice, explainers form a design space with trade-offs between fidelity, stability, human interpretability, and computational cost. Comparative studies show that SHAP tends to be a stable, high-fidelity explainer across models, but is not always the most understandable to humans (Nayebi and Heer, 2022; Schlegel et al., 2019). Other methods such as gradient-based saliency or rule-based surrogates may be faster or more intuitive for certain architectures. Empirical work in security and SOC settings further indicates that explanations can be computed with millisecond-scale overheads [around 4.8 ms in recent SOC studies (Kwubeghari et al., 2025)] and can significantly change analyst behavior, but do not uniformly improve decisions across all tasks (Jesus et al., 2021; Wali et al., 2021). SHAP is therefore treated as the default explainer for tabular protocol/process features, but the architecture explicitly allows cheaper gradient-based or approximate methods for strict real-time deployments and more expensive Shapley-style explanations for offline triage.



**Require:** Anomaly model  $M$ , event features  $x = (x_1, \dots, x_d)$ , top- $k$  parameter

**Ensure:** Explanation  $E$  with feature attributions and narrative

```

1:  $a \leftarrow M(x)$ 
2: if  $a > \tau_{\text{anomaly}}$  then
3:    $\phi \leftarrow \text{TreeSHAP}(M, x)$ 
4:    $\text{top\_features} \leftarrow \text{argsort}(|\phi|)[-k:]$ 
5:    $\text{attributions} \leftarrow \{(x_i, \phi_i) : i \in \text{top\_features}\}$ 
6:    $\text{narrative} \leftarrow \text{GenerateNarrative}(\text{attributions}, \text{process\_context})$ 
7:    $E \leftarrow (\text{attributions}, \text{narrative}, a)$ 
8:   return  $E$ 
9: else
10:  return null
11: end if

```

Algorithm 1. SHAP-based explanation pipeline.

## 6.4 Reference implementation blueprint with Malcolm

Malcolm is selected as the reference platform for several reasons:

- Integration:** Malcolm bundles Arkime (PCAP sessionization), Zeek, Suricata, and OpenSearch/Dashboards in a single containerized stack (CISA, 2025), eliminating integration overhead.
- ICS protocol support:** Zeek ICSNPP protocol parsers (Modbus, DNP3, S7, IEC 60870-5-104, OPC UA) install via Malcolm's plugin mechanism, addressing G3.

- Open-source availability:** Malcolm is maintained by CISA under Apache-2.0 license, supporting reproducibility and community adoption.
- Extensibility:** custom sidecars can be added for ML/SHAP computation without modifying core components.

Prior ICS IDS research commonly relies on Zeek, Suricata, and Elastic/Kibana individually (Joy et al., 2024; Waagsnes and Ulltveit-Moe, 2018; Haas et al., 2020). However, no widely adopted standard reference implementation bundles these components for ICS research. Malcolm provides this integration as a pragmatic baseline for prototyping and evaluation.

In this article, Malcolm is used only as a reference platform/blueprint. No new experimental results on Malcolm are presented. Future work will implement and evaluate the proposed architecture on Malcolm or a similar stack.

The proposed extensions to Malcolm include:

- ICS parsing:** Zeek ICSNPP parsers for Modbus, DNP3, S7, IEC 60870-5-104, OPC UA.
- Signature channel:** Suricata with ET Open/Pro rulesets plus ICS-specific decoder events.
- ML+SHAP sidecar:** a container that computes temporal features, scores events, fuses alerts, and writes `ids-fusion-*` indices with SHAP attributions.
- Dashboards:** OpenSearch Dashboards panels joining fusion indices with Arkime sessions for drill-down.

Table 6 maps components to architecture roles.

## 6.5 Runtime and operator impact

Target latency objectives: detection within  $< 400$  ms and alert-plus-explain within  $< 800$  ms at the 95th percentile (p95) under sustained load. These targets are informed by accelerated

TABLE 6 Mapping of implementation components to architecture roles.

Role	Concrete component(s)	Notes
Signature engine	Suricata (ET Open/Pro; ICS decoder events)	EVE JSON $\rightarrow$ suricata-*
Anomaly engine	Sidecar models (temporal)	Reads zeek-*/suricata-*; optional OpenSearch AD baseline
Protocol parsing	Zeek + ICSNPP (Modbus, DNP3, S7, IEC-104, OPC UA)	Installed via Malcolm Zeek plugin mechanism
Sessionization	Arkime	PCAP pivot and carving
Storage/search	OpenSearch	Schema for ids-fusion-* with SHAP fields
Dashboards	OpenSearch Dashboards	Fusion view + SHAP summaries
Sensors (optional)	Hedgehog Linux	Passive baseline and live capture

SHAP methods (Yang, 2021; Jethani et al., 2021) and prior IDS latency measurements (Waagsnes and Ulltveit-Moe, 2018). These latency targets should be interpreted as budgets for DMZ/SCADA monitoring and SOC triage rather than hard real-time PLC control. In high-load settings, detection and alerting can occur immediately, with explanations computed either using a lightweight online explainer or asynchronously in batches, balancing the millisecond-scale XAI overheads reported in recent SOC studies (Kwubeghari et al., 2025) against the constraints of the underlying streaming pipeline.

SOC studies show that SHAP/LIME attributions help analysts understand alert rationales and can improve trust, though objective accuracy gains are mixed (Eriksson and Grov, 2022; Gaspar et al., 2024; Weerts et al., 2019). The proposed design combines feature attributions with concise narratives, counterfactuals, and process context mappings to maximize actionability.

## 7 Research agenda

This section outlines a research agenda for progressing from conceptual architecture to validated prototype and pilot deployment. Each agenda item explicitly addresses one or more of the five practical gaps (G1–G5).

- 1. Prototype integration** [G1, G2, G3]: implement parallel signature and anomaly engines with the fusion layer described in Section 6.2. Integrate ICSNPP protocol parsers for Modbus, DNP3, IEC 60870-5-104, and OPC UA to address protocol blindness (G3). Validate fusion weight  $\alpha$  and threshold  $\tau_{\text{alert}}$  calibration against baseline traffic.
- 2. Benchmark evaluation** [G1, G2]: compare hybrid configurations against signature-only and anomaly-only

baselines on multiple ICS datasets (SWaT, WADI, HAI, IEC-104 corpus). Report gains and trade-offs for zero-day coverage (G1) and false-positive control (G2) using standardized metrics.

- 3. Explainability UX** [G4]: design operator-facing narratives mapping SHAP attributions to process context. Conduct usability studies with ICS operators and SOC analysts to assess triage speed, decision accuracy, and trust (Khediri et al., 2024).
- 4. SHAP-guided optimization** [G4, G5]: automate feature selection and model retraining using accumulated SHAP importance statistics. Implement drift detection (Equation 2) and threshold update (Equation 3) mechanisms to address deployment drift (G5) (Wadinger and Kvasnica, 2024).
- 5. Adversarial robustness** [G1, G5]: assess evasion and poisoning risks against the hybrid IDS. Adopt ensemble defenses and attribution-ensemble methods to improve robustness (Chen et al., 2022; Awad et al., 2025; Fung et al., 2024). Evaluations should also treat the explanation interface as a security-relevant component: adversarial examples can evade ML detectors, and interpretations and *post-hoc* explainers (including LIME/SHAP-style perturbation approaches) have been shown to be fragile and manipulable (Ghorbani et al., 2019; Slack et al., 2020). Minimal mitigations that can be assessed without redesign include access control and rate limiting for explanation endpoints, plus monitoring explanation stability under adversarial probing (Ghorbani et al., 2019).
- 6. SOC integration and ATT&CK for ICS** [G4, G5]: map detections and explanations to MITRE ATT&CK for ICS techniques to improve triage workflow integration and situational awareness (Bhosale et al., 2024; Kaya et al., 2025).
- 7. Pilot deployment** [G5]: validate performance and integration in a high-fidelity testbed or partner operational environment. Address protocol coverage gaps, latency constraints, and real-world drift patterns. Document lessons learned for field deployment.
- 8. Specification and stateful model fusion** [G1, G3]: integrate specification-based and stateful protocol models as first-class fusion inputs. Rather than treating specification-based detectors and protocol state machines as separate baselines, future work should ingest their outputs (e.g., invariant violation scores, automaton state anomalies) as additional channels in the hybrid fusion layer, alongside signature and anomaly model outputs. This would enable systematic evaluation of whether process invariants and protocol state models (Feng et al., 2019; Song et al., 2024; Faisal and Sitnikova, 2016; Matoušek et al., 2021) improve zero-day detection and operator-level explanations when combined with protocol-aware features in a single architecture.

Table 7 summarizes the mapping between agenda items and gaps.

## 8 Evaluation plan

This section describes a planned evaluation methodology; no experiments have yet been conducted. These are evaluation pipelines for future work, not completed experiments.

TABLE 7 Mapping of research agenda items to practical gaps (G1–G5).

Agenda item	Gaps addressed
1. Prototype integration	G1 (zero-day), G2 (FP), G3 (protocol)
2. Benchmark evaluation	G1 (zero-day), G2 (FP)
3. Explainability UX	G4 (explainability)
4. SHAP-guided optimization	G4 (explainability), G5 (drift)
5. Adversarial robustness	G1 (zero-day), G5 (deployment)
6. SOC integration/ATT&CK	G4 (explainability), G5 (deployment)
7. Pilot deployment	G5 (deployment)
8. Specification/stateful model fusion	G1 (zero-day), G3 (protocol)

## 8.1 Datasets and testbeds

The evaluation plan uses multiple datasets exercising protocol coverage and process-aware features. SWaT and HAI are selected primarily for evaluating process awareness and concept-drift handling, while IEC-104 and DNP3 corpora target protocol-semantic detection:

- **Secure water treatment (SWaT)** (Goh et al., 2016): water treatment plant traces with labeled attack scenarios. Used for training/validation/test splits (60/20/20) and cross-validation for zero-day scenarios.
- **IEC 60870-5-104 corpus** (Radoglou-Grammatikis et al., 2019): power grid dataset with labeled IEC-104 protocol events. Addresses protocol coverage gaps for IEC-104 semantics.
- **DNP3 dataset** (Rodofile et al., 2017): substation traffic with DNP3 protocol attacks. Exercises DNP3 function code detection.
- **Hardware-in-the-Loop (HAI)** (Oyedotun et al., 2025): realistic ICS testbed data for multimodal evaluation.

**Real-world baseline:** a sustained, passive baseline capture from a production SCADA/ICS environment (no attack injections) will provide a clean reference distribution for: (i) calibrating false-positive budgets under operational noise; (ii) validating SHAP attribution stability across natural concept drift; and (iii) stress-testing generalization beyond laboratory conditions (Skrodelis H. K. et al., 2024; Mitseva et al., 2022).

## 8.2 Metrics and scenarios

Standard detection metrics: precision, recall,  $F_1$ , AUROC, false-positive rate (FPR), false-negative rate (FNR), time-to-detect, and alert volume. Explanation metrics: SHAP computation latency, attribution stability under drift, and operator-assessed interpretability.

Beyond standard classification metrics, the evaluation includes time-series-aware and operational metrics that better reflect ICS monitoring requirements. Following critiques of Average Run

Length (ARL), which may fail to control the instantaneous false alarm rate and can even be infinite in some regimes (Kuhn et al., 2018; Mei, 2008), the evaluation reports: (i) distributions of detection delay (time from attack onset to first alert), (ii) false alarms per hour/day and average run length between false alarms, and (iii) redundancy and triage-related measures, such as the fraction of low-risk or redundant alerts filtered by post-processing (Kiruki et al., 2023; Zhang et al., 2024). These choices are motivated by SOC studies highlighting rule management burden, alert volume, and alert fatigue as central operational constraints (Vermeer and McGovern, 2023; Yang et al., 2021; Tariq et al., 2025).

Scenarios include: (1) known attack detection (signature channel); (2) zero-day/novel attack detection (anomaly channel); (3) cross-dataset generalization (train on SWaT, test on WADI); and (4) concept drift injection (simulate equipment changes).

## 8.3 Concept drift handling

Drift detection methods—Adaptive Windowing (ADWIN), Fast Hoeffding Drift Detection Method (FHDDM), Kolmogorov–Smirnov Windowing (KSWIN)—trigger threshold recalibration per Equation 3. Pre/post-drift AUROC/ $F_1$  deltas and SHAP attribution stability are logged (Wadinger and Kvasnica, 2024; Pesaranghader and Viktor, 2016; Bharani et al., 2024a).

## 8.4 Operator-in-the-loop study

A within-subjects study with experienced analysts compares baseline (anomaly-only, no explanations) against the hybrid + XAI system. Measured outcomes: time-to-triage, classification correctness, workload (alerts per session), perceived trust (Likert scale), and qualitative feedback (Eriksson and Grov, 2022; Weerts et al., 2019).

## 8.5 Real-time feasibility

End-to-end latency profiling under load targets <400 ms detection and <800 ms alert-plus-explain at p95. GPU-accelerated TreeSHAP/FastSHAP and batching are employed to meet latency budgets (Yang, 2021; Jethani et al., 2021).

For real-time feasibility, the evaluation measures end-to-end latency and throughput of the full streaming pipeline—from packet capture and protocol parsing through feature extraction, model scoring, and explanation generation—rather than model latency in isolation. Prior benchmarking of stream processing frameworks and IDS engines shows substantial variability in achievable latency/throughput across different deployments (Waleed et al., 2022; Gallenmüller et al., 2022). The evaluation therefore reports both per-event and tail latencies (p95/p99) under realistic load, and explicitly relates them to detection and explanation budgets to demonstrate where the proposed architecture is viable (e.g., DMZ monitoring, historian mirroring) and where tighter control-loop constraints would require simplified models or offline explanations.

## 8.6 Reproducibility and threats to validity

Code, configurations, random seeds, and evaluation scripts will be released on a public GitHub repository under an open-source license. To address dataset bias, all metrics will be reported per attack type and per operating mode (e.g., normal, transient, maintenance) rather than as single aggregate figures. Cross-dataset tests and ablations on protocol subsets address external validity. Documented threats include: dataset biases (laboratory vs. field), limited attack diversity, and potential overfitting to testbed characteristics (Mitseva et al., 2022; Wolsing et al., 2024).

## 9 Discussion and conclusion

This systematic review synthesizes a decade of research on hybrid and explainable IDS for SCADA/ICS environments, identifying persistent practical gaps and proposing a conceptual architecture to address them.

The synthesis reveals that hybrid IDS architectures can achieve detection accuracies of 90%–99% with false-positive rates of 0.8%–2.1% in laboratory evaluations. However, zero-day detection remains challenging (approximately 73% for truly novel attacks), and the gap between laboratory and field performance is not well characterized. Protocol-aware features are severely underrepresented, leaving IDS blind to protocol-level attack vectors. Explainability techniques (SHAP, LIME) are implemented in only a small fraction of recent studies despite evidence that they improve analyst trust and triage speed.

The proposed conceptual architecture operationalizes an explainable hybrid IDS under ICS constraints: low latency, determinism, and safety. Mathematical formalization of fusion logic, drift detection, and SHAP-based explanation pipelines provides a foundation for future implementation. Malcolm is justified as a pragmatic reference platform integrating Zeek, Suricata, and OpenSearch with ICS protocol support.

The evidence base is constrained by the laboratory-heavy nature of available datasets, scarce operator studies in real control room settings, and limited protocol diversity in existing evaluations. Questions remain about attribution stability under adversarial drift and long-term maintenance burden.

Looking forward, three emerging directions are particularly relevant to this agenda: (i) foundation/self-supervised representation learning to reduce labeled-data dependence (still largely nascent for OT telemetry) (Shurrab and Duwairi, 2022; Wolf et al., 2020), (ii) digital twin-assisted IDS that enable what-if/counterfactual analyses to support explainability (Krishnaveni et al., 2024), and (iii) human-in-the-loop adaptive IDS workflows that incorporate analyst feedback and safe online updates prior to field rollout (Kim et al., 2024). These directions should be evaluated under the same ICS constraints emphasized in this review (latency, safety, and change control).

Immediate next steps include prototype implementation, preregistered operator studies, and field pilot validation. Releasing reproducible pipelines, reporting negative findings, and testing under concept drift will accelerate shared learning across the ICS security community.

Practitioners gain a blueprint, gap matrix, and evaluation checklist to deploy explainable hybrid IDS with improved zero-day coverage, controlled false-positive rates, and higher analyst trust in SCADA/ICS networks.

## Author contributions

HS: Data curation, Visualization, Resources, Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Validation, Investigation, Formal analysis. AR: Supervision, Writing – review & editing.

## Funding

The author(s) declared that financial support was not received for this work and/or its publication.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomp.2026.1724245/full#supplementary-material>

## References

- Abbas, S. G., Ozmen, M. O., Alsaheel, A., Khan, A., Celik, Z. B., Xu, D., et al. (2024). "SAIN: improving ICS attack detection sensitivity via state-aware invariants," in *USENIX Security Symposium* (Philadelphia, PA: USENIX Association).
- Abdelaty, M. F., Doriguzzi Corin, R., and Siracusa, D. (2021). DAICS: a deep learning solution for anomaly detection in industrial control systems. *IEEE Trans. Emerg. Topics Comput.* 10, 1117–1129. doi: 10.1109/TETC.2021.3073017
- Almalawi, A., Hassan, S., Fahad, A., Iqbal, A., and Khan, A. I. (2025). Hybrid cybersecurity for asymmetric threats: intrusion detection and SCADA system protection innovations. *Symmetry* 17:616. doi: 10.3390/sym17040616
- Altunay, H., Albayrak, Z., Özalp, A., and Çakmak, M. (2021). "Analysis of anomaly detection approaches performed through deep learning methods in SCADA systems," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (Ankara: IEEE), 1–6. doi: 10.1109/HORA52670.2021.9461273
- Awad, Z., Zakaria, M., and Hassan, R. (2025). An enhanced ensemble defense framework for boosting adversarial robustness of intrusion detection systems. *Sci. Rep.* 15:14177. doi: 10.1038/s41598-025-94023-z
- Batani, M., and Baraani, A. (2013). Time window management for alert correlation using context information and classification. *Int. J. Comput. Netw. Inf. Secur.* 5, 9–16. doi: 10.5815/ijcnis.2013.11.02
- Bharani, D., Lakshmi Priya, V., and Saravanan, S. (2024a). "Adaptive real-time malware detection for IoT traffic streams: a comparative study of concept drift detection techniques," in *ICICNIS* (Bengaluru: IEEE).
- Bharani, D., Lakshmi Priya, V., and Saravanan, S. (2024b). "Adaptive real-time malware detection for IOT devices: a comparative study of concept drift detection techniques," in *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)* (Bengaluru: IEEE), 172–179.
- Bhatia, S., Kush, N., Djamaludin, C., Akande, J., and Foo, E. (2014). "Practical Modbus flooding attack and detection," in *Proceedings of the Twelfth Australasian Information Security Conference (AISC 2014)* (Auckland).
- Bhosale, P., Kastner, W., and Sauter, T. (2024). "Mapping ICS vulnerabilities with MITRE ATT&CK and Bayesian networks," in *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)* (Padova: IEEE).
- Burgetova, I., Ahmed, M., Judah, J., and Khorshed, T. (2021). Anomaly detection using supervised learning and temporal features in industrial control systems. *IEEE Access* 9, 24292–24305. doi: 10.1109/CNSM52442.2021.9615542
- Cali, U., Çatak, F. O., and Halden, U. (2024). Trustworthy cyber-physical power systems using AI: dueling algorithms for PMU anomaly detection and cybersecurity. *Artif. Intell. Rev.* 57:183. doi: 10.1007/s10462-024-10827-x
- Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: a survey on methods and metrics. *Electronics* 8:832. doi: 10.3390/electronics8080832
- Chen, J., Gao, X., Deng, R., He, Y., Fang, C., and Cheng, P. (2022). Generating adversarial examples against ML-based intrusion detectors in industrial control systems. *IEEE Trans. Dependable Secure Comput.* 19, 1810–1825. doi: 10.1109/TDSC.2020.3037500
- CISA (2025). *Malcolm Documentation*. Available online at: <https://malcolm.fyi/> (Accessed November 10, 2025).
- Clark, J., Zhang, S., and Hayajneh, T. (2018). "Adaptive sliding window for anomaly-based intrusion detection systems," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (Vancouver, BC: IEEE), 1080–1087.
- Conti, M., Donadel, D., and Turrin, F. (2021). A survey on industrial control system testbeds and datasets for security research. *IEEE Commun. Surv. Tutorials* 23, 2243–2276. doi: 10.1109/COMST.2021.3094360
- Dakhil, Y., and Çakmak, M. (2025). XAI-XGBoost: an innovative explainable intrusion detection approach for securing internet of medical things systems. *Sci. Rep.* 15:22278. doi: 10.1038/s41598-025-07790-0
- Dutta, P., Josan, P. K., Wong, R. K. W., Dunbar, B. J., Diaz-Artiles, A., Selva, D., et al. (2025). Effects of explanations and accuracy on human performance and trust in AI-assisted anomaly diagnosis tasks. *J. Cogn. Eng. Decis. Mak.* 19, 453–473. doi: 10.1177/15553434251338433
- Eißler, J., Schuba, M., Höner, T., Hack, S., and Neugebauer, G. (2023). *Human-Centric Introduction to a Complex Cybersecurity Standard*. Istanbul: AHFE International. doi: 10.54941/ahfe1004249
- Eriksson, H. S., and Grov, G. (2022). "Towards XAI in the SOC—a user centric study of explainable alerts with SHAP and LIME," in *2022 IEEE International Conference on Big Data (Big Data)* (Osaka: IEEE). doi: 10.1109/BigData55660.2022.10020248
- Esquivel-Vargas, H., Frank, J., Niakanlahiji, A., and Zonouz, S. A. (2017). "Automatic specification extraction from industrial control system device documentation," in *Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and Privacy* (New York, NY: ACM), 31–42.
- Faisal, M., and Sitnikova, E. (2016). "Modeling stateful intrusion detection systems for industrial control systems using finite state machine," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)* (Hefei: IEEE), 1395–1400.
- Feng, C., Li, T., and Chana, D. (2017). "Multi-level anomaly detection in industrial control systems via package signatures and LSTM networks," in *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (Denver, CO: IEEE), 261–272. doi: 10.1109/DSN.2017.34
- Feng, C., Li, T., Zhu, Q., and Zhang, H. (2019). Systematic study of distributed anomaly detection for cyber security in power grids. *IEEE Trans. Smart Grid* 11, 1648–1659. doi: 10.1109/TSG.2019.2906316
- Fung, C., Zeng, E., and Bauer, L. (2024). "Attributions for ML-based ICS anomaly detection: from theory to practice," in *NDSS* (San Diego, CA). doi: 10.14722/ndss.2024.23216
- Gallenmüller, S., Wiedner, F., Naab, J., and Carle, G. (2022). How low can you go? A limbo dance for low-latency network functions. *J. Netw. Syst. Manage.* 31:20. doi: 10.1007/s10922-022-09710-3
- García-Teodoro, P., Díaz-Verdejo, J., Macía-Fernandez, G., and Vazquez, E. (2009). Anomaly-based network intrusion detection: techniques, systems and challenges. *Comput. Secur.* 28, 18–28. doi: 10.1016/j.cose.2008.08.003
- Gaspar, D., Silva, P., and Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on MLP. *IEEE Access* 12, 65650–65660. doi: 10.1109/ACCESS.2024.3368377
- Ghazi, Y., Tabaa, M., Ennaji, M., and Zaz, G. (2025). An explainable Markov chain-machine learning sequential-aware anomaly detection framework for industrial IoT systems based on OPC UA. *Sensors* 25:6122. doi: 10.3390/s25196122
- Ghorbani, A., Abid, A., and Zou, J. Y. (2019). Interpretation of neural networks is fragile. *Proc. AAAI Conf. Artif. Intell.* 33, 3681–3688. doi: 10.1609/aaai.v33i01.33013681
- Goh, J., Adepou, S., Junejo, K. N., and Mathur, A. (2016). "A dataset to support research in the design of secure water treatment systems," in *The 11th International Conference on Critical Information Infrastructures Security* (Cham: Springer). doi: 10.1007/978-3-319-71368-7\_8
- Haas, S., Sommer, R., and Fischer, M. (2020). "Zeek-osquery: host-network correlation for advanced monitoring and intrusion detection," in *ICT Systems Security and Privacy Protection. SEC 2020. IFIP Advances in Information and Communication Technology* (Cham: Springer), 248–262. doi: 10.1007/978-3-030-58201-2\_17
- Ike, M., Phan, K., Badapanda, A., Landen, M., Sadoski, K., Guo, W., et al. (2023). Bridging both worlds in semantics and time: domain knowledge based analysis and correlation of industrial process attacks. *arXiv [preprint]*. arXiv:2311.18539. doi: 10.48550/arXiv.2311.18539
- Jacovi, A., and Goldberg, Y. (2020). "Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. doi: 10.18653/v1/2020.acl-main.386
- Jesus, S., Santos, G., Henriques, P. R., and Antunes, N. (2021). "How much do explanations tell about intrusion detection models?" in *2021 IEEE European Symposium on Security and Privacy Workshops (EuroSec&PW)* (Vienna: IEEE), 188–197.
- Jethani, N., Sudarshan, M., Covert, I. C., Lee, S.-I., and Ranganath, R. (2021). "FastSHAP: real-time Shapley value estimation," in *ICLR*.
- Joy, A., Chandane, M., and Kazi, F. (2024). "An investigative evaluation of open source intrusion detection systems for operational technology networks using MITRE ICS attack simulation on a thermal power plant test bed," in *2024 IEEE 21st India Council International Conference (INDICON)* (Kharagpur: IEEE), 1–6. doi: 10.1109/INDICON63790.2024.10958514
- Katar, C., Debar, H., and Viho, C. (2006). "Combining multiple intrusion detection systems in coalition," in *2006 ACS/IEEE International Conference on Computer Systems and Applications* (Amman: IEEE), 303–310.
- Kaya, S. C., Elbez, G., and Hagenmeyer, V. (2025). "Enhancing situational awareness in smart grids through event correlation for ATT&CK mapping," in *Energy-Efficient Computing and Networking* (Cham: Springer).
- Kayode Saheed, Y., Harazeem Abdulganiyu, O., and Ait Tchakoucht, T. (2023). A novel hybrid ensemble learning for anomaly detection in industrial sensor networks and SCADA systems for smart city infrastructures. *J. King Saud Univ. Comput. Inf. Sci.* 35:101532. doi: 10.1016/j.jksuci.2023.03.010
- Kenmogne, L. A., and Mocanu, S. (2024). "Explainable AI for process-aware attack detection in industrial control systems," in *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)* (Saint Louis, MO: IEEE), 363–368. doi: 10.1109/NetSoft60951.2024.10588940

- Khan, I. A., Pi, D., Khan, Z. U., Hussain, Y., and Nawaz, A. (2019). HML-IDS: a hybrid-multilevel anomaly prediction approach for intrusion detection in SCADA systems. *IEEE Access* 7, 89507–89521. doi: 10.1109/ACCESS.2019.2925838
- Khediri, A., Slimi, H., Yahiaoui, A., Derdour, M., Bendjenna, H., Ghenai, C. E., et al. (2024). “Enhancing machine learning model interpretability in intrusion detection systems through SHAP explanations and LLM-generated descriptions,” in *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)* (EL OUED: IEEE). doi: 10.1109/PAIS62114.2024.10541168
- Khrasat, A., Gondal, I., Vamplew, P., and Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity* 2:20. doi: 10.1186/s42400-019-0038-7
- Kim, Y., Dán, G., and Zhu, Q. (2024). Human-in-the-loop cyber intrusion detection using active learning. *IEEE Trans. Inf. Forensics Secur.* 19, 8658–8672. doi: 10.1109/TIFS.2024.3434647
- Kiruki, J. K., Muketha, G. M., and Kamau, G. (2023). Metrics for evaluating alerts in intrusion detection systems. *Int. J. Netw. Secur. Appl.* 15, 15–37. doi: 10.5121/ijnsa.2023.15102
- Kleinmann, A., and Wool, A. (2016). “Automatic construction of stateful modbus/TCP intrusion detection system using scada data,” in *Proceedings of the 2016 ACM Workshop on Cyber-Physical System Security* (New York, NY: ACM), 23–34.
- Koucham, O., Mocanu, S., Hiet, G., Thiriet, J. C., and Majorczyk, F. (2022). Cross-domain alert correlation methodology for industrial control systems. *Comput. Secur.* 118: 102723. doi: 10.1016/j.cose.2022.102723
- Krishnaveni, S., Sivamohan, S., Jothi, B., Chen, T. M., and Sathiyarayanan, M. (2024). TwinSec-IDS: an enhanced intrusion detection system in SDN-digital-twin-based industrial cyber-physical systems. *Concurr. Comput. Pract. Exp.* 37:e8334. doi: 10.1002/cpe.8334
- Kuhn, J., Mandjes, M., and Taimre, T. (2018). Practical aspects of false alarm control for change point detection: beyond average run length. *Methodol. Comput. Appl. Probab.* 21, 25–42. doi: 10.1007/s11009-018-9636-1
- Kwon, H.-Y., Kim, T., and Lee, M.-K. (2022). Advanced intrusion detection combining signature-based and behavior-based detection methods. *Electronics* 11:867. doi: 10.3390/electronics11060867
- Kwubeghari, A., Gates, C., and Atapour-Abarghouei, A. (2025). “Designing Explainable intrusion detection systems for SOC analysts,” in *2025 IEEE Conference on Communications and Network Security (CNS)* (San Antonio, TX: IEEE), 253–259.
- Lundberg, S. M., and Lee, S.-I. (2017). “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (Long Beach, CA: Curran Associates).
- Matoušek, P., Havlena, V., and Holík, L. (2021). “Efficient modelling of ICS communication for anomaly detection using probabilistic automata,” in *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)* (Bordeaux: IEEE), 81–89.
- Mei, Y. (2008). Is average run length to false alarm always an informative criterion? *Seq. Anal.* 27, 354–376. doi: 10.1080/07474940802445790
- Mitchell, R., and Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Comput. Surv.* 46, 55:1–55:29. doi: 10.1145/2542049
- Mitseva, A., Thierse, P., Hoffmann, H., Er, D., and Panchenko, A. (2022). “Challenges and pitfalls in generating representative ICS datasets in cyber security research,” in *Proc. CyberICPS Workshop (ESORICS)* (Cham: Springer). doi: 10.1007/978-3-031-25460-4\_22
- Nayebi, A., and Heer, J. (2022). “An empirical evaluation of explainable AI measures for model trust,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM).
- Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., et al. (2022). Explainable intrusion detection systems (X-IDS): a survey of current methods, challenges, and opportunities. *IEEE Access* 10, 99187–99231. doi: 10.1109/ACCESS.2022.3216617
- Oyedotun, S. A., Oise, G. P., and Ozobialu, C. E. (2025). Towards intelligent cybersecurity in SCADA and DCS environments: anomaly detection using multimodal deep learning and explainable AI. *J. Sci. Res. Rev.* 2, 20–31. doi: 10.70882/josrar.2025.v2i3.76
- Pan, S., Morris, T., and Adhikari, U. (2015). Developing a hybrid intrusion detection system using data mining for power systems. *IEEE Trans. Smart Grid* 6, 3104–3113. doi: 10.1109/TSG.2015.2409775
- Pesaranghader, A., and Viktor, H. L. (2016). “Fast Hoeffding drift detection method for evolving data streams,” in *MLDM* (Cham: Springer). doi: 10.1007/978-3-319-46227-1\_7
- Radoglou-Grammatikis, P., Sarigiannidis, P., Giannoulakis, I., Kafetzakis, E., and Panaousis, E. (2019). *IEC 60870-5-104 Network Characterization and Intrusion Detection* (Thessaloniki).
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “‘Why should I trust you?’: explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM). doi: 10.1145/2939672.2939778
- Rodofle, N. R., Radke, K., and Foo, E. M. (2017). “Framework for SCADA cyber-attack dataset creation,” in *ACSW* (New York, NY: ACM). doi: 10.1145/3014812.3014883
- Rosa, L., Cruz, T., Freitas, M. B., Quitério, P., Henriques, J., Caldeira, F., et al. (2021). Intrusion and anomaly detection for the next-generation of industrial automation and control systems. *Future Gener. Comput. Syst.* 119, 50–67. doi: 10.1016/j.future.2021.01.033
- Sahu, A., and Davis, K. (2021). “Structural learning techniques for Bayesian attack graphs in cyber physical power systems,” in *2021 IEEE Texas Power and Energy Conference (TPEC)* (College Station, TX: IEEE), 1–6. doi: 10.1109/TPEC51183.2021.9384933
- Sahu, A., and Siano, P. (2021). Inter-domain fusion for enhanced intrusion detection in power systems: an evidence theoretic and meta-heuristic approach. *Sensors* 22:2100. doi: 10.3390/s22062100
- Schlegel, U., Arnout, H., Kleickmann, S., Schmidt, M., and Thom, A. (2019). “Towards a rigorous evaluation of XAI methods on time series,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–12. doi: 10.1109/ICCVW.2019.00516
- Shetty, S., McShane, M., and Panwar, S. (2011). An integrated Kullback Leibler divergence and Bayesian update framework for anomaly detection in wireless networks. *Secur. Commun. Netw.* 4, 1–12. doi: 10.1002/sec.151
- Shurrab, S., and Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Comput. Sci.* 8:e1045. doi: 10.7717/peerj-cs.1045
- Skrodelis, H., Kelle, R., and Romanovs, A. (2024). “Cybersecurity in scada systems with advanced ai and ml techniques,” in *2024 IEEE 65th International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)* (Riga: IEEE), 1–5. doi: 10.1109/ITMS64072.2024.10741936
- Skrodelis, H. K., Blumbergs, B., and Romanovs, A. (2024). “Threat scenario generation for IEC104 cyber defense,” in *2024 IEEE 11th Workshop on Advances in Information, Electronic and Electrical Engineering* (Riga: IEEE), 1–7. doi: 10.1109/AIEEE62837.2024.10586596
- Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). “Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (New York, NY: ACM), 180–186. doi: 10.1145/3375627.3375830
- Sommer, R., and Paxson, V. (2010). “Outside the closed world: on using machine learning for network intrusion detection,” in *IEEE Symposium on Security and Privacy* (Oakland, CA: IEEE). doi: 10.1109/SP.2010.25
- Song, Y., Feng, C., Li, T., and Zhang, H. (2024). “Leveraging physics-model based invariants for industrial control system intrusion detection,” in *IEEE Transactions on Industrial Informatics* (Piscataway, NJ: IEEE).
- Stefanidis, K., and Voyiatzis, D. (2016). “An HMM-based anomaly detection approach for SCADA systems,” in *WISTP* (Cham: Springer). doi: 10.1007/978-3-319-45931-8\_6
- Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M., and Hahn, A. (2015). *Guide to industrial Control Systems (ICS) Security. Special Publication 800-82 Rev. 2*. Gaithersburg, MD: NIST. doi: 10.6028/NIST.SP.800-82r2
- Taormina, R., Finney, K. G., Wicker, S. B., and Grave, R. (2018). Battle of the attack detection algorithms: disclosing cyber attacks on water distribution networks. *J. Water Resour. Plan. Manag.* 144:04018048. doi: 10.1061/(ASCE)WR.1943-5452.0000969
- Tariq, S., Allix, K., and Lalande, J.-F. (2025). “Alert fatigue in security operations centers: causes and mitigation,” in *2025 IEEE European Symposium on Security and Privacy (EuroS&P)* (Venice: IEEE), 174–190.
- Umer, M. A., Mathur, A., Junejo, K. N., and Adepu, S. (2017). “Integrating design and data centric approaches to generate invariants for distributed attack detection,” in *CPS '17: Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and Privacy* (New York, NY: ACM). doi: 10.1145/3140241.3140248
- Ustebay, S., Akgün, B. B., and Gaj, P. (2024). “Securing SCADA systems: a protocol-based intrusion detection approach with shapley analysis,” in *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)* (Ankara: IEEE), 1–7. doi: 10.1109/ASYU62119.2024.10756979
- Vermeer, M. J., and McGovern, G. (2023). *Alert Management and Fatigue in Cybersecurity Operations Centers*. Santa Monica, CA: RAND Corporation Reports.
- Waagsnes, H., and Ulltveit-Moe, N. (2018). “Intrusion detection system test framework for SCADA systems,” in *Proceedings of the 4th International Conference on Information Systems Security and Privacy* (Funchal: SciTe Press). doi: 10.5220/0006588202750285
- Wadinger, M., and Kvasnica, M. (2024). Adaptable and interpretable framework for anomaly detection in SCADA-based industrial systems. *Expert Syst. Appl.* 246:123200. doi: 10.1016/j.eswa.2024.123200
- Waleed, A.-K., Dubois, D. J., Haven, J., Rossi, D., and Ziliotto, G. (2022). “Which network intrusion detection system is right for you? A performance evaluation of Suricata, Snort and Zeek,” in *2022 IEEE 11th International Conference on Cloud Networking (CloudNet)* (Paris: IEEE), 108–114.
- Wali, S. A., Ramli, R. I., Abawajy, J., and Hassan, M. M. (2021). “Explainable Intrusion detection system for imbalanced dataset,” in *2021 International Conference on Decision Aid Sciences and Application (DASA)* (Manama: IEEE), 390–394.

- Wang, Y., and Ghorbani, A. (2004). "Distributed intrusion detection using cooperating agents," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)* (The Hague: IEEE), 4432–4437.
- Weerts, H. J. P., van Ipenburg, W., and Pechenizkiy, M. (2019). A human-grounded evaluation of SHAP for alert processing. *arXiv [preprint]*. arXiv:1907.03324. doi: 10.48550/arXiv.1907.03324
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. doi: 10.18653/v1/2020.emnlp-demos.6
- Wolsing, K., Wagner, E., Basels, F., Wagner, P., and Wehrle, K. (2024). "Deployment challenges of industrial intrusion detection systems," in *ESORICS Workshops* (New York, NY: ACM). doi: 10.1007/978-3-031-82349-7\_29
- Wolsing, K., Wagner, E., Saillard, A., and Henze, M. (2022). "IPAL: breaking up silos of protocol-dependent and domain-specific industrial intrusion detection systems," in *Proceedings of the 25th International Symposium on Research in Attacks, Intrusions and Defenses, RAID '22* (New York, NY: Association for Computing Machinery), 510–525. doi: 10.1145/3545948.3545968
- Xu, L., Huang, X., Li, R., Zhang, Z., Luo, J., Xu, H., et al. (2023). ADTCD: anomaly detection via teacher-student knowledge distillation under concept drift. *Neurocomputing* 560:126514. doi: 10.1016/j.neucom.2023.126514
- Xu, Y., and Eckert, C. (2009). A novel Bayesian method for combining multiple classifiers in network intrusion detection. *Comput. Secur.* 28, 209–220. doi: 10.1016/j.cose.2008.09.005
- Yang, J. (2021). Fast TreeSHAP: accelerating SHAP value computation for trees. *arXiv [preprint]*. arXiv:2109.09847. doi: 10.48550/arXiv.2109.09847
- Yang, S., Zhang, X., Cai, Z., Li, S., and Zhang, D. (2021). "Near-real-time alert correlation method for ISP security operation center," in *2021 IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)* (IEEE), 211–219.
- Zhang, S., Liu, Y., Zhu, X., and Cheng, X. (2024). MNSSA: a multi-level network security situational awareness framework. *Comput. Netw.* 238:109966. doi: 10.1016/j.comnet.2023.110162