

CyTReX: Explainable AI-Based Cybersecurity Threat Reasoning Framework for DER Networks

Damilola Popoola, Souradeep Bhattacharya, and Manimaran Govindarasu

Department of Electrical and Computer Engineering

Iowa State University, Ames, IA 50011

Email: popson@iastate.edu, sbhatta@iastate.edu, gmani@iastate.edu

Abstract—Distributed Energy Resource (DER) environments rely on network communication protocols to coordinate control commands, measurements, and device states across edge assets and cloud systems. Edge anomaly detection systems (ADS) monitor this traffic to identify deviations from normal communication behavior, flagging suspicious flows for further investigation. When the ADS flags abnormal network traffic, a single attack label is often insufficient for operational response: the label reports the detector’s selected class but does not expose alternative threat interpretations that may warrant investigation. This paper presents Cybersecurity Threat Reasoning with Explainable Artificial Intelligence (CyTReX), an evidence-grounded threat reasoning framework for DER security that transforms network-level anomaly alerts into ranked, analyst-facing threat hypotheses designed to support Security Operations Center (SOC) triage and investigation. CyTReX constrains large language model (LLM) reasoning through a structured evidence packet, defined as a consolidated record of detection outputs, model explanations, and cyber threat intelligence (CTI) context. The evidence packet integrates edge-layer anomaly detection evidence, cloud reasoning layer attack interpretation, Shapley Additive Explanations (SHAP) network-feature attributions, surrogate decision rules, and Model Context Protocol (MCP)-enabled CTI enrichment. This ensures that every ranked hypothesis and attack-tree branch is traceable to explicit evidence rather than free-form LLM inference, and that incomplete or conflicting evidence is communicated rather than suppressed. Evaluation across five configurations shows that additional reasoning components improve hypothesis specificity, evidence traceability, and analytical grounding, with the complete pipeline providing the richest evidence-grounded reasoning context.

Index Terms—Distributed energy resources, threat reasoning, explainable artificial intelligence, security operations center, cyber threat intelligence

I. INTRODUCTION

Distributed Energy Resources (DER) environments rely on network communication protocols such as Distributed Network Protocol 3 (DNP3) to coordinate control commands, device states, and measurements across edge assets, aggregators, and cloud systems [1]. Abnormal DNP3 traffic may reflect replay attacks, unauthorized control commands, or denial-of-service (DoS) behavior, and multi-class identification requires distinguishing among several attack types from network-flow features [2], [3]. Critically, a single-class detection label reports only the selected class and does not communicate which alternative threat interpretations remain plausible or what additional evidence would distinguish them.

Detectors based on artificial intelligence (AI) and machine learning (ML) improve coverage but do not resolve this gap [4], [5]. Without understanding which network features drove the alert, how they relate to known adversary techniques, and which alternative behaviors warrant investigation, a Security Operations Center (SOC) analyst has no reliable basis for triage, escalation, or response.

This paper presents CyTReX, a threat reasoning framework for DER security that converts ML-generated network anomaly alerts into ranked, evidence-supported threat hypotheses. CyTReX constrains large language model (LLM) reasoning through a structured evidence packet that integrates edge anomaly evidence, cloud reasoning layer attack interpretation, Shapley Additive Explanations (SHAP) network-feature attributions, surrogate decision rules, and Model Context Protocol (MCP)-enabled cyber threat intelligence (CTI) enrichment. This ensures that every ranked hypothesis and attack-tree branch is traceable to explicit evidence rather than free-form LLM inference, and that incomplete or conflicting evidence is communicated rather than suppressed. The main contributions are: (i) an edge-cloud reasoning workflow that transforms DER network anomalies into ranked threat hypotheses to support SOC analyst triage; (ii) a unified evidence packet consolidating network-feature detection outputs, model explanations, CTI context, and uncertainty as structured LLM input; and (iii) a five-configuration DNP3 case study evaluating each reasoning component’s contribution to hypothesis specificity, evidence traceability, and unsupported-claim reduction.

II. BACKGROUND AND MOTIVATION

A. DER-Integrated Grid Architecture

Modern grid environments increasingly rely on DER assets, edge devices, communication gateways, control applications, and cloud-based analytics for monitoring and operation [1]. DER resources such as solar photovoltaic systems, battery storage units, electric vehicles, and smart inverters exchange measurements, commands, and status information using standardized communication protocols, including DNP3 [6]. As shown in Fig. 1, DER-integrated architectures connect DER clients, edge intelligent devices (EIDs), field-area networks, aggregators, and distribution control systems through multiple communication paths [7].

Because DER communication tightly couples network traffic, control commands, and device state, the same network-

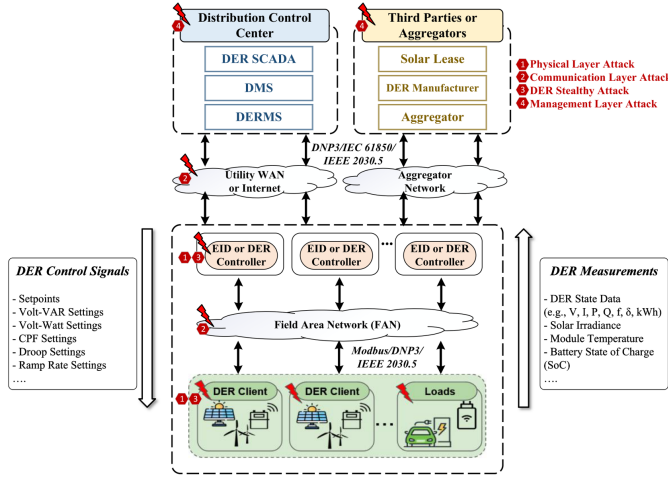


Fig. 1. Representative DER-integrated grid communication architecture and attack surfaces [9].

level anomaly can require different interpretations depending on protocol, device, and operational context. A single-class detector reports its selected class but does not enumerate alternative hypotheses or the evidence needed to distinguish them. Edge devices observe communication close to the asset and forward alerts to cloud-side analytics frameworks for aggregation and interpretation, but neither layer alone supplies the full evidence context needed for such reasoning [8]. This motivates the need for SOC-oriented reasoning frameworks that go beyond the flagged label to examine what else the traffic may indicate.

B. DER Cybersecurity Context and Threat Reasoning

DER cybersecurity protects the communication, computation, and control processes that support reliable operation of distributed energy assets [10], [11]. In DER-integrated environments, security events can affect control actions, telemetry visibility, device coordination, and operator decision-making [12]. As these environments become more digitalized, field devices and edge monitoring systems generate network traffic data that may reveal abnormal communication patterns. However, a single detection label does not show which alternative threat interpretations remain plausible or how the affected device and command context should influence triage.

Cybersecurity threat reasoning addresses this interpretation gap by moving beyond the detection label toward structured explanation. Rather than treating an alert class as the final output, threat reasoning connects network communication behavior, model interpretation, and contextual knowledge to determine what the alert may indicate and what alternative behaviors may remain plausible [13]. This is important in DER environments because the same abnormal traffic pattern may reflect different operational consequences depending on the affected device, command type, and available supporting evidence.

C. Need for Evidence-Grounded Reasoning

Evidence-grounded reasoning is important for DER cybersecurity because operational decisions cannot rely only on

anomaly scores or isolated class labels [14]. Without supporting evidence, a SOC analyst cannot determine whether an alert reflects a confirmed threat, a benign operational condition, or an uncertain behavior that requires further investigation, and cannot assess which secondary threat hypotheses merit further investigation.

By linking each threat interpretation to observable network behavior, model explanations, CTI context, and evidence gaps, evidence-grounded reasoning improves the transparency and usefulness of DER security alerts [5]. It helps analysts understand why a hypothesis is plausible, which alternative explanations remain possible, and what additional evidence is needed before response actions are taken. This paper adopts this evidence-grounded view by constraining LLM reasoning through a structured evidence packet, ensuring that every ranked hypothesis is traceable to explicit evidence and that incomplete or conflicting evidence is communicated rather than suppressed.

III. RELATED WORK

Machine learning-based anomaly detection has been widely studied for networked cyber-physical systems, industrial control systems, and grid communication environments [4], [19], [20]. In DNP3-enabled DER networks, supervised classifiers have also been used to distinguish multiple cyber/physical attack classes [3]. However, anomaly detection mainly identifies abnormal behavior and does not provide sufficient attack semantics. Closed-set classifier accuracy does not by itself establish reliability for attack behavior excluded from training; unknown-attack performance must be evaluated separately [21]. These limitations have motivated DER cybersecurity research that moves beyond detection toward situational intelligence and higher-level alert interpretation [15]. While such work improves the operational context of DER security monitoring, there remains a need for threat reasoning methods that explicitly connect ML evidence to CTI context.

Explainable AI methods, including SHAP [22] and Local Interpretable Model-agnostic Explanations (LIME), have been used to interpret anomaly detection decisions and identify influential features [5], [23]. However, feature explanations alone do not show whether the evidence is consistent with a plausible cyber threat hypothesis or how it should be combined with operational context, rule evidence, and threat intelligence. Recent DER-focused work has addressed part of this issue through self-verifying anomaly detection, where explanation consistency is used to assess the reliability of anomaly evidence [16]. These efforts show the importance of moving from feature-level explanation toward evidence-aware interpretation in grid cybersecurity.

CTI repositories provide external context for interpreting observed cyber behavior. Structured Threat Information Expression (STIX) and Trusted Automated Exchange of Intelligence Information (TAXII) support standardized CTI representation and exchange [24], [25]. MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) for ICS and Common Attack Pattern Enumeration and Classification

TABLE I
COMPARISON OF CYBERSECURITY THREAT REASONING AND ANALYSIS FRAMEWORKS

Framework / Approach	ML Anomaly Detection	Explainable AI (XAI)	CTI Enrichment	LLM-Assisted Reasoning	Ranked Hypotheses	Grid/Industrial Control Systems (ICS) Focus
ML-Based Anomaly Detection System (ML-ADS) [3]	✓	×	×	×	×	✓
CySIDER [15]	✓	×	✓	×	×	✓
XAI-Forensic Intrusion Detection System (IDS) [5]	✓	✓	×	×	×	×
ExCyDER [16]	✓	✓	×	×	×	✓
LLM Cybersecurity Survey [17]	×	×	✓	✓	×	×
SEvenLLM [18]	×	×	✓	✓	×	×
CyTRex	✓	✓	✓	✓	✓	✓

(CAPEC) provide complementary behavior-level knowledge about adversary tactics, techniques, and procedures (TTPs), attack patterns, and potential mitigations [26], [27]. While these sources can enrich alert interpretation, CTI must be grounded in observed evidence to avoid unsupported associations, especially when direct packet-level or operational evidence is incomplete.

LLMs and generative AI have recently been explored for cybersecurity decision support, incident summarization, threat analysis, CTI synthesis, vulnerability interpretation, and analyst-assistance workflows [17], [18], [28], [29]. These models are useful because they can organize heterogeneous security information, summarize complex evidence, and generate human-readable explanations that support investigation. In cyber-physical and grid-security settings, such capabilities are relevant because alerts may involve protocol behavior, device context, operational impact, and external threat knowledge [30]. However, LLM outputs can be sensitive to prompt design and may overgeneralize, introduce unsupported CTI associations, or present uncertain interpretations as confirmed conclusions when evidence constraints are weak [17], [31].

Table I summarizes the landscape of related works across these domains. Taken together, existing methods either focus on isolated detection, provide feature-level explanations without operational context, or apply LLMs without strict evidence grounding. This leaves a critical gap for alert interpretation methods that connect low-level cyber-physical evidence with ranked, evidence-grounded threat hypotheses.

IV. PROPOSED METHODOLOGY

A. Framework Overview

The proposed CyTRex framework provides an explainable and evidence-grounded cyber threat reasoning workflow for DER-integrated grid security. As shown in Fig. 2, CyTRex transforms cyber-physical alerts from communication or telemetry behavior into structured attack trees with ranked threat hypotheses. Rather than treating detection outputs as final decisions, the proposed framework organizes alert evidence through an edge-cloud pipeline, where the edge layer generates early abnormality evidence and the cloud reasoning layer performs threat interpretation, model explanation, CTI enrichment, and LLM-based attack-tree reasoning to support uncertainty-aware analyst investigation.

Operationally, CyTRex informs SOC triage by converting a flagged DER network alert into a structured investigation object. The edge layer first identifies abnormal traffic and forwards the alert evidence to the cloud reasoning layer.

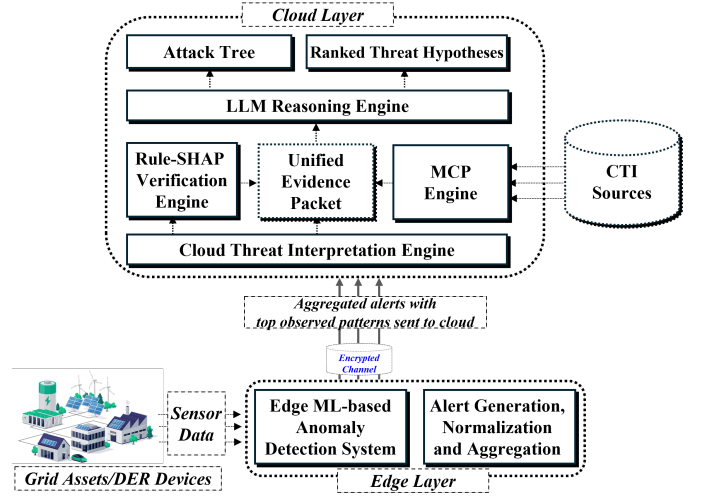


Fig. 2. CyTRex framework showing transmission of edge-generated alert evidence to the cloud for evidence-grounded threat reasoning and hypothesis generation

The cloud layer then adds attack-class interpretation, model explanation evidence, CTI context, and evidence-quality signals before the LLM structures the information into ranked hypotheses and attack-tree branches. This output helps analysts see the most supported threat interpretation, related secondary possibilities, relevant CTI and mitigation context, and remaining evidence gaps before deciding whether to escalate, investigate further, or approve a response action.

B. Edge Evidence Generation Layer

The edge layer provides the first level of evidence for the reasoning workflow. It observes communication or telemetry behavior close to the grid asset and determines whether the behavior deviates from expected operation. Its role is to produce early abnormality evidence, not final attack attribution.

The output of this layer summarizes the observed anomaly, the strength of the deviation, relevant communication context, and the features that contributed to the alert. This evidence is forwarded to the cloud reasoning layer, where it is combined with interpretation, model explanation features, and threat intelligence to support ranked cyber threat hypotheses.

C. Cloud Threat Reasoning Layer

The cloud layer transforms edge-generated abnormality evidence into structured cyber threat interpretation. Its role is to examine whether the alert is consistent with known or plausible attack behavior, enrich the evidence with model explanation features and CTI, and organize the result for attack-tree generation.

Because a single alert may contain indicators of multiple attack behaviors, the cloud reasoning layer does not treat classification as a final decision. Instead, it supports ranked threat hypotheses based on the available evidence, its quality, and the remaining uncertainty.

D. Model Explanation Layer

The model explanation layer extracts explicit reasoning from the cloud classifier’s decisions by combining SHAP feature attributions with surrogate decision rules. Its primary role is to expose the exact behavioral features driving a classification, ensuring the threat hypothesis is grounded in interpretable model behavior rather than opaque scores. While the layer includes a secondary consistency check between SHAP and rule evidence to generate an evidence-quality signal, the main contribution to the unified evidence packet is the detailed feature-level and rule-based explanation. This allows the reasoning process to preserve uncertainty when evidence is weak or incomplete and to assign stronger support when the model explanation is clear and consistent. In this way, robust model explanations help reduce unsupported conclusions and strengthen evidence-grounded threat reasoning.

E. Cyber Threat Intelligence Context Layer

The CTI layer adds external context to the alert by linking observed behavior to known adversary techniques, attack patterns, and official mitigation guidance. In CyTReX, CTI retrieval is behavior-oriented: MITRE ATT&CK for ICS provides candidate techniques, tactics, and associated mitigations, while CAPEC provides related attack-pattern context.

This layer is accessed through an MCP-enabled interface [32] and contributes provenance-aware CTI evidence to the unified evidence packet. This design keeps the reasoning process traceable and prevents unsupported CTI claims when the alert provides only behavior-level evidence.

F. Unified Evidence Packet and Attack-Tree Reasoning

The unified evidence packet serves as the structured input for LLM-based reasoning. It brings together edge evidence, cloud reasoning layer attack interpretation, model explanation features, CTI, evidence-quality signals, and provenance meta-data.

Using this packet, the LLM generates an attack tree that includes the root suspicious event, ranked threat hypotheses, supporting evidence, uncertainty notes, and evidence gaps. The output is intended to support analyst interpretation and decision readiness, not automated response execution.

V. EXPERIMENTAL EVALUATION

A. Experimental Setup

Fig. 3 illustrates the experimental setup, which emulates a DER network environment. CyTReX is implemented in Python 3.13 on an Apple M4 system (16 GB of random-access memory (RAM), macOS 15.5) with a Streamlit-based interface running Ollama model locally. Three virtual DER devices simulate DNP3 traffic in real time to the edge detection layer,

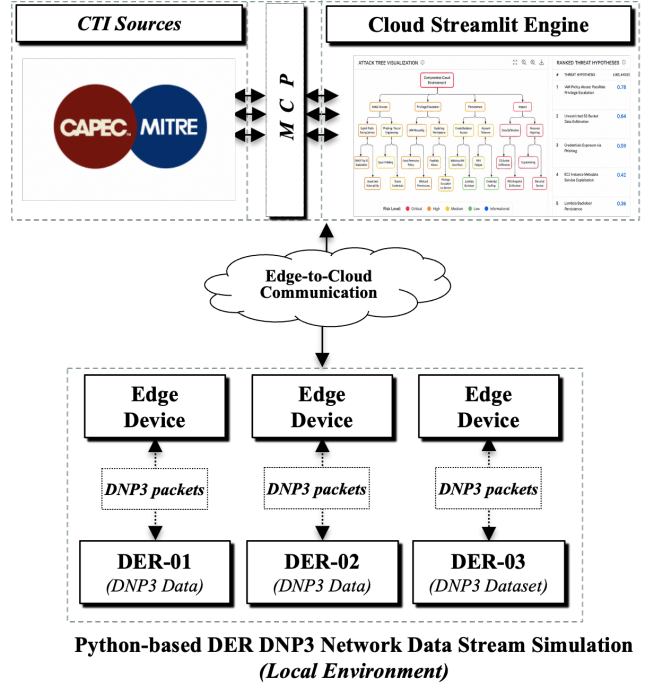


Fig. 3. Experimental setup showing DER devices transmitting DNP3 packets to edge devices, with edge-to-cloud communication, MCP-enabled CTI retrieval, and cloud-based threat reasoning through the Streamlit engine.

where a semi-supervised anomaly detector is used to identify abnormal communication behavior and aggregate alerts into JavaScript Object Notation (JSON) batches for cloud layer processing. The cloud reasoning layer uses a supervised classifier for attack-class interpretation, SHAP for feature-attribution evidence, surrogate rule evidence for explanation verification, and an MCP-enabled CTI interface for external context retrieval. The LLM layer uses llama3.1:8b through a local Ollama backend for all attack-tree generation tasks. The same model is used across all five evaluation configurations, with no configuration-specific LLM changes, so output differences are attributed to the active evidence components rather than to different LLM settings.

B. Dataset and Case-Study Labels

The evaluation uses a DNP3-based communication dataset as the case-study environment [33]. Table II summarizes the dataset labels and their meaning in a DNP3 communication context. The supervised cloud classifier is trained on six attack labels and normal traffic, while DNP3_ENUM and DNP3_INFO are excluded from training and used as held-out out-of-distribution (OOD) cases for reconnaissance-style behavior. A representative edge-device alert batch is used across all configurations so that each setting reasons over the same cyber-physical event. Although DNP3 provides the case-study protocol, CyTReX operates on normalized evidence objects rather than protocol-specific assumptions.

For model selection, the dataset was split into training and testing subsets (70:30) before evaluating the edge and cloud models reported in Table VI. The edge anomaly detector was trained to learn normal communication behavior and tested

on normal and attack traffic, while the cloud classifier was trained and tested using the closed-set labels in Table II. This separation ensures that Table VI reports closed-set model performance, while the held-out OOD cases evaluate reasoning behavior outside the trained attack-label space.

C. Model Explanation and Feature Attribution

The Rule-SHAP consistency formulation is adapted from prior work [16], while the surrogate rule-quality score is introduced here as an additional verification signal. CyTRex extracts the unique feature names from the surrogate rules, denoted as F_r , and uses their count to set the SHAP evidence budget:

$$k = \max(1, |F_r|), \quad (1)$$

where k is the number of top-ranked SHAP features retained for comparison and $F_s^{(k)}$ denotes the corresponding top- k SHAP feature set. The Rule-SHAP consistency score combines feature overlap O , direction agreement D , and rule coverage C :

$$S_{rs} = \alpha O + \beta D + \gamma C, \quad (2)$$

where $\alpha = 0.45$, $\beta = 0.10$, and $\gamma = 0.45$. The larger weights on feature overlap and rule coverage place the score mainly on feature-level agreement: overlap captures whether SHAP and surrogate rules identify similar influential features, while coverage captures whether rule-derived features are represented in the SHAP evidence budget. Direction agreement receives a lower weight because SHAP contribution direction and surrogate branch direction may differ even when both identify the same behaviorally relevant feature.

CyTRex also computes a surrogate rule-quality score:

$$Q_r = \eta F_{sur} + (1 - \eta) R_d, \quad (3)$$

where F_{sur} is surrogate fidelity and R_d is rule density relative to the SHAP budget. In this study, $\eta = 0.65$, giving fidelity a moderate majority weight over rule density because rule explanations are most useful when the surrogate closely approximates the cloud classifier. Rule density retains a secondary weight of 0.35 to reward explanation availability without treating more rules as inherently better explanations. Both S_{rs} and Q_r are normalized scores in the range $[0, 1]$, where values closer to 1 indicate stronger explanation support and values closer to 0 indicate weaker support. Rule-only features that do not overlap with SHAP are retained because they may expose additional surrogate decision patterns.

The verified confidence is computed as:

$$\text{Conf} * v = \text{Conf} * m \times (0.5 + 0.5V), \quad (4)$$

where $\text{Conf} * m$ is the original model confidence and $V = \max(S * rs, 0.5Q_r)$ is the combined verification signal. The weighting values are design parameters, and future work will assess how alternative schemes affect hypothesis ranking, evidence-quality labels, and reasoning consistency.

D. MCP-Enabled CTI Retrieval

The MCP layer adds structured CTI context to the unified evidence packet through a FastMCP-based client-server module. Its scope is limited to CTI enrichment and provenance-aware retrieval, and it does not perform anomaly detection, attack classification, SHAP computation, surrogate-rule extraction, attack-tree generation, or automated mitigation. In the evaluation, MCP enrichment is enabled only in the edge-plus-CTI and complete-pipeline configurations. Rather than allowing the LLM to perform unconstrained threat-intelligence lookup, the MCP client converts alert evidence into structured CTI requests by extracting discriminative behavioral terms from edge alerts, cloud classifications, SHAP-ranked features, surrogate decision-rule features, and flow metadata, while filtering generic domain vocabulary such as DNP3, ICS, and supervisory control and data acquisition (SCADA).

The server retrieves behavior-level CTI from MITRE ATT&CK for ICS and CAPEC. MITRE ATT&CK for ICS provides adversary techniques, tactics, STIX relationships, and official mitigation guidance, while CAPEC provides complementary attack-pattern context. The server queries MITRE ATT&CK for ICS through the TAXII 2.1 endpoint, matches active STIX techniques against evidence-derived terms, and resolves official mitigations through the STIX mitigates relationship. CAPEC patterns are retrieved from the CAPEC XML corpus and scored using the same evidence-derived terms. The retrieved techniques, attack patterns, mitigations, source references, retrieval status, and provenance metadata are packaged into a unified CTI context for LLM-based reasoning. In the local evaluation, live retrieval is used with caching disabled and in-process transport as the default, while standard input/output (stdio), Hypertext Transfer Protocol (HTTP), and Server-Sent Events (SSE) transport modes are also supported.

E. Evaluation Configurations

The evaluation uses five configurations grouped into two tracks. The edge-only track comprises configurations C1 and C2 and evaluates how much reasoning can be supported from edge evidence alone and how CTI changes that edge-only reasoning. The full-pipeline track comprises configurations C3, C4, and C5 and evaluates the added value of cloud reasoning layer attack interpretation, model explanation features, CTI enrichment, and the complete evidence-grounded reasoning workflow. Table IV summarizes the configurations.

For the ranked outputs, H1 denotes the highest-ranked hypothesis, while H2 and H3 denote the second- and third-ranked hypotheses, respectively.

F. Evaluation Criteria

The evaluation considers two aspects: ML detection performance and threat-reasoning behavior. Detection performance is measured using accuracy, precision, recall, the F1-score (the harmonic mean of precision and recall), confusion matrix, and prediction latency to verify the reliability of the edge and cloud model outputs used as evidence. Threat-reasoning

TABLE II
DATASET LABELS IN THE DNP3 CASE STUDY.

Label	Use in this study	Meaning in DNP3 context
NORMAL	Benign baseline	Expected DNP3 communication behavior.
REPLAY	Trained attack	Replay of DNP3 packets previously captured from a legitimate DNP3 master or outstation.
COLD_RESTART	Trained attack	A malicious master-like entity sends the DNP3 cold-restart function code, causing the target outstation to initiate a complete restart.
WARM_RESTART	Trained attack	A malicious master-like entity sends the DNP3 warm-restart function code, causing a partial restart or reinitialization of the DNP3 service.
STOP_APP	Trained attack	A stop-application command associated with DNP3 Function Code 18 asks the slave/outstation to stop a function so it cannot continue receiving messages from the master.
DISABLE_UNSOLICITED	Trained attack	A command associated with DNP3 Function Code 21 disables unsolicited messages from the target outstation.
INIT_DATA	Trained attack	An initialize-data command associated with DNP3 Function Code 15 requests unauthorized reinitialization of device data or configuration values.
DNP3_ENUM	Held-out OOD test	Reconnaissance activity intended to identify which DNP3 services and function codes are supported by the target system.
DNP3_INFO	Held-out OOD test	Reconnaissance activity that collects DNP3 diagnostic or usage information from the target environment.

TABLE III
MCP-ENABLED CTI TOOLS USED IN THE EVALUATION.

Tool	Purpose
extract_evidence_terms	Extracts CTI query terms from edge, cloud, SHAP, rule, and flow evidence.
query_mitre_attack_ics	Retrieves MITRE ATT&CK for ICS techniques, tactics, STIX relationships, and official mitigations.
query_capec_patterns	Retrieves CAPEC attack-pattern context from the CAPEC Extensible Markup Language (XML) corpus.
retrieve_full_cti_context	Orchestrates evidence-term extraction, MITRE retrieval, CAPEC retrieval, normalization, and provenance recording.

TABLE IV
FIVE-CONFIGURATION EVALUATION DESIGN.

Config.	Configuration	Purpose
C1	Edge-only: Edge evidence + LLM	Tests whether edge anomaly evidence alone supports attack-tree generation.
C2	Edge-only: Edge evidence + CTI + LLM	Tests whether CTI enriches edge-only reasoning without cloud reasoning layer attack interpretation.
C3	Full pipeline: Edge evidence + cloud reasoning layer attack interpretation + LLM	Tests the effect of cloud reasoning layer attack interpretation.
C4	Full pipeline: Edge evidence + cloud reasoning layer attack interpretation + Rule-SHAP + LLM	Tests whether model explanation features improves evidence support.
C5	Complete pipeline: Edge evidence + cloud + Rule-SHAP + CTI + LLM	Tests the complete evidence-grounded reasoning workflow.

behavior is evaluated using the primary H1 hypothesis, runtime latency, evidence grounding, and uncertainty handling across the five configurations. The discussion focuses on whether each configuration preserves the dataset label or a close semantic interpretation, whether the generated hypothesis aligns with the DNP3 behavior described in Table II, and whether additional components such as cloud reasoning layer attack interpretation, Rule-SHAP verification, and CTI enrichment improve the traceability of the reasoning output. For held-out DNP3_ENUM and DNP3_INFO cases, the evaluation examines whether the framework treats unseen reconnaissance-like behavior as uncertain rather than forcing unsupported closed-set conclusions.

VI. RESULTS AND DISCUSSION

A. ML Performance and Model Selection

Table VI reports the model-selection results for the edge and cloud layers using accuracy, precision, recall, and F1-score, abbreviated as Acc., Prec., Rec., and F1, respectively. Accuracy measures the overall proportion of correct predictions, precision measures the correctness of predicted attack or anomaly cases, recall measures how many true attack or anomaly cases are detected, and F1-score summarizes the balance between precision and recall, with higher values indicating stronger performance. Based on these results, the Isolation Forest and nearest-neighbor ensemble was selected for edge-layer abnormality detection, while Light Gradient Boosting Machine (LightGBM) was selected for cloud-layer supervised attack-behavior interpretation.

B. Attack-Label Reasoning Results and Latency

Table V reports the primary H1 hypothesis and runtime latency for each trained DNP3 attack label across C1–C5. These outputs are interpreted as threat-reasoning results rather than direct classifier predictions using three levels of interpretation. An exact label match occurs when H1 preserves the dataset label or a direct variant, such as REPLAY being reported as DNP3 replay. A semantic match occurs when H1 does not repeat the dataset label but captures the same attacker role, command action, or operational effect, such as restart behavior being mapped to denial of service or STOP_APP being mapped to command injection. A mismatch occurs when H1 shifts to a weakly related behavior, such as C1 mapping COLD_RESTART to DNP3 replay, indicating that the available evidence is insufficient for precise attack attribution.

The results show that edge-only configurations produce broader hypotheses because they lack supervised attack-class context, while CTI can widen the hypothesis space by mapping protocol-specific DNP3 labels to broader behavior-level concepts. C3 provides the strongest label anchoring by preserving COLD_RESTART, DISABLE_UNSOLICITED, REPLAY, and WARM_RESTART as H1. REPLAY appears as H1 under all five configurations, showing strong behavioral separability, whereas INIT_DATA is not preserved as H1, suggesting overlap with broader command or access-related behavior. C5 provides richer semantic interpretation for STOP_APP

TABLE V
OVERVIEW OF GENERATED HIGHEST-RANKED (H1) HYPOTHESES AND LATENCY ACROSS TRAINED DNP3 ATTACK LABELS.

Attack label	Result	C1	C2	C3	C4	C5
COLD_RESTART	H1	DNP3 replay	Rogue master	Cold restart	Cold restart	Denial of service
	Latency	152.06 s	180.74 s	138.85 s	117.95 s	160.33 s
DISABLE_UNSOLICITED	H1	Command replay	Distributed denial-of-service (DDoS) attack	Disable unsolicited	Cold restart	DNP3 replay
	Latency	135.82 s	182.78 s	122.78 s	151.20 s	134.18 s
INIT_DATA	H1	Rogue master	Denial of service	Rogue master	Cold restart	DNP3 replay
	Latency	150.28 s	157.64 s	142.43 s	176.39 s	133.37 s
REPLAY	H1	DNP3 replay	DNP3 replay	DNP3 replay	DNP3 replay	DNP3 replay
	Latency	158.71 s	172.93 s	132.41 s	145.39 s	172.38 s
STOP_APP	H1	Injection	DNP3 replay	Device restart	Cold restart	Command injection
	Latency	295.72 s	181.80 s	126.10 s	186.58 s	142.61 s
WARM_RESTART	H1	DNP3 manipulation	Denial of service	Warm restart	Cold restart	Denial of service
	Latency	150.15 s	373.03 s	184.95 s	167.56 s	151.29 s

TABLE VI
ML MODEL SELECTION RESULTS FOR EDGE AND CLOUD REASONING LAYERS.

Layer	Model	Acc.	Prec.	Rec.	F1
Edge	Isolation Forest+Nearest Neighbor	98.43	99.04	99.37	99.21
	Isolation Forest	97.93	98.98	98.85	98.91
Edge	Extended Isolation Forest	57.52	98.50	56.26	71.61
Cloud	LightGBM	99.58	99.58	99.58	99.58
Cloud	Extreme Gradient Boosting (XGBoost)	99.08	99.10	99.08	99.08
Cloud	Random Forest	85.79	88.31	85.80	85.32

and restart-related cases, but it may also shift from protocol-specific labels toward broader threat interpretations when CTI and explanation evidence are combined. These H1 mismatches show where the reasoning pipeline needs stronger packet-level, temporal, or open-set support. C5 appears weaker than C3 under exact label matching because C3 is mainly guided by the supervised cloud classifier, while C5 combines classifier evidence with Rule-SHAP and CTI context. This can shift protocol-specific labels toward broader threat effects, such as restart cases mapping to denial of service or STOP_APP mapping to command injection. Therefore, C5 trades some label anchoring for richer evidence-grounded interpretation, mitigation context, and visibility into possible co-occurring behaviors.

Latency across C1–C5 is not strictly monotonic because runtime reflects the combined cost of LLM generation, CTI retrieval, evidence-packet size, and attack-tree complexity. C3 has the lowest average latency, while C2 has the highest mainly because the WARM_RESTART case produced broader CTI matches that increased evidence-packet size and generation time. C5 remains relatively stable despite using the complete pipeline, showing that latency is better interpreted as an end-to-end reasoning cost than as a direct function of the number of active components. This also clarifies the role of the LLM in CyTRex. The LLM is not evaluated as a standalone detector or unconstrained CTI search engine, but as a controlled reasoning layer that converts each configuration’s evidence packet into ranked hypotheses and attack-tree branches. Its contribution is therefore assessed through H1 correctness, semantic alignment, evidence grounding, uncertainty handling, and end-to-

TABLE VII
OOD REASONING RESULTS FOR HELD-OUT DNP3 LABELS.

OOD label	Result	C2	C5
DNP3_ENUM	H1	Parameter manipulation	DNP3 replay
	H2	Data collection	Denial of service
	H3	Process disruption	Rogue master injection
	Latency	153.00 s	136.15 s
DNP3_INFO	H1	Parameter manipulation	DNP3 replay
	H2	Initial access attempt	DNP3 spoofing
	H3	Lateral movement attempt	DNP3 command control
	Latency	202.84 s	135.85 s

end latency.

C. Replay Attack-Tree Analysis

Fig. 4 shows the attack tree generated for the REPLAY case using the C5 complete-pipeline configuration. This case is representative because REPLAY is a well-defined DNP3 threat, appears as H1 across all configurations in Table V, and still requires evidence-grounded interpretation beyond label matching. The generated tree assigns DNP3_REPLAY_ATTACK as the primary hypothesis and DNP3_MANIPULATION_ATTACK as a secondary hypothesis, since a replayed command may manipulate device state, disrupt control flow, or trigger an unauthorized response. The UNKNOWN hypothesis preserves uncertainty where packet-sequence evidence is insufficient to fully confirm replay.

The retrieved MITRE ATT&CK for ICS context frames the alert around replay-related behaviors, including command manipulation, unauthorized control activity, and operational disruption. The CTI and mitigation entries therefore provide behavior-level support and investigation guidance, not direct proof that every retrieved technique or pattern occurred. Mitigation overlap across H1, H2, and H3, including communication authenticity, authentication, network allowlisting, and out-of-band validation, can support action prioritization by highlighting defenses that remain relevant across multiple plausible hypotheses. Although Fig. 4 presents the detailed replay example, Table V also shows similar reasoning for other labels, where STOP_APP maps to command injection and restart-related cases map to denial-of-service-oriented behavior, reflecting their operational effects.

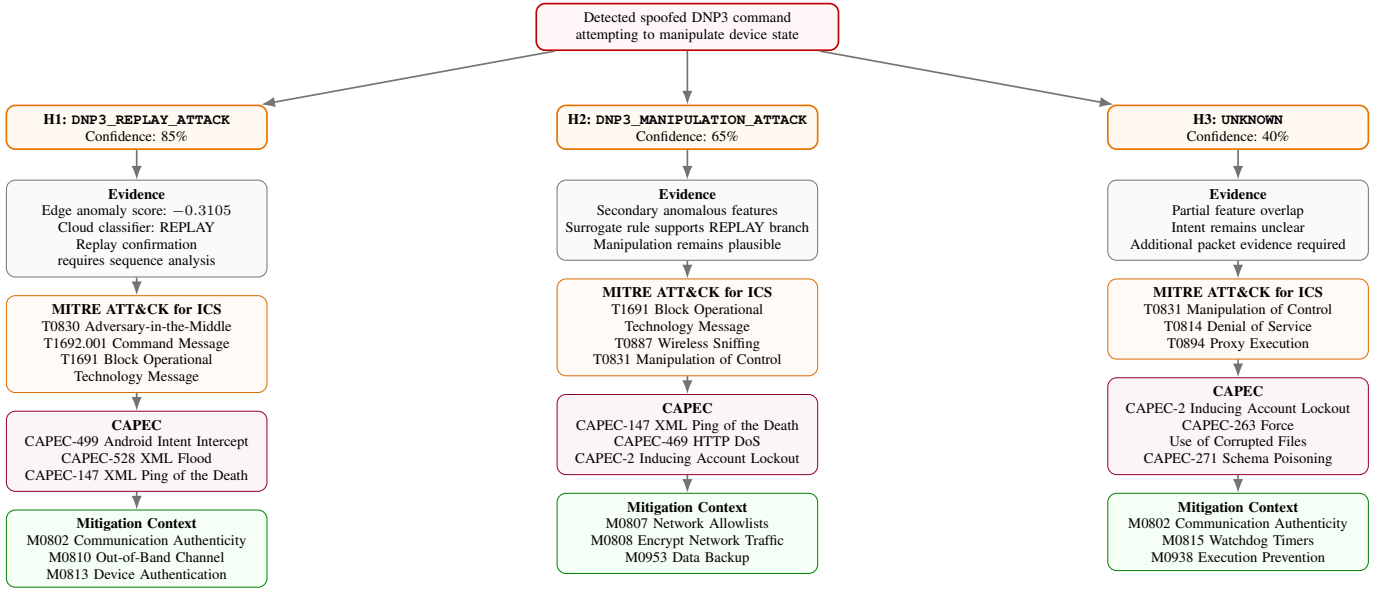


Fig. 4. Compact evidence-grounded attack-tree representation for the REPLAY case, showing the root event, ranked threat hypotheses, supporting evidence, MITRE ATT&CK for ICS techniques, CAPEC patterns, and mitigation context.

D. Out-of-Distribution Behavior

The held-out DNP3_ENUM and DNP3_INFO cases are treated as OOD tests. As shown in Table VII, CyTReX generates security-relevant hypotheses for both unseen reconnaissance-style behaviors. In C2, edge anomaly evidence and CTI context produce broad hypotheses such as parameter manipulation, data collection, initial access, and lateral movement. In C5, cloud interpretation, SHAP attribution, and surrogate rule evidence enrich the output, but the closed-set classifier can also pull unseen behavior toward known classes such as replay. This confirms that CyTReX provides transparent reasoning for OOD cases while highlighting the need for explicit open-set detection to avoid forcing unseen behaviors into known attack categories.

VII. CONCLUSION

This paper presented CyTReX, an evidence-grounded cybersecurity threat reasoning framework for DER networks that addresses a key limitation of single-label anomaly detection: a detection label reports the selected class but does not expose alternative threat interpretations or the evidence needed to distinguish them. CyTReX moves beyond this limitation by generating ranked, traceable threat hypotheses from integrated detection evidence, XAI-based network-feature explanations, MCP-enabled CTI enrichment, and LLM-based attack-tree reasoning, providing SOC analysts with evidence context for informed triage and response. The five-configuration evaluation shows that edge-only reasoning detects abnormal behavior but remains limited for attack attribution, while cloud reasoning, model explanation, and CTI enrichment improve evidence traceability and analytical grounding. The OOD evaluation further shows that CyTReX can generate security-relevant hypotheses for unseen reconnaissance-like behavior, while also highlighting the need for explicit open-set handling.

Future work will evaluate CyTReX on larger datasets and broader DER communication scenarios, and extend it toward human-in-the-loop response support where evidence-grounded reasoning can inform analyst-approved mitigation actions.

ACKNOWLEDGEMENT

This work is funded in part by the U.S. Department of Energy (DOE), Office of Cybersecurity, Energy Security, and Emergency Response (CESER), under Award Number DE-CR0000049.

REFERENCES

- [1] I. Zografopoulos, N. D. Hatziaargyriou, and C. Konstantinou, "Distributed energy resources cybersecurity outlook: Vulnerabilities, attacks, impacts, and mitigations," *IEEE Systems Journal*, vol. 17, no. 4, pp. 6695–6709, 2023.
- [2] G. Dangwal, S. Mittal, M. Wazid, J. Singh, A. K. Das, D. Giri, and M. J. F. Alenazi, "An effective intrusion detection scheme for Distributed Network Protocol 3 (DNP3) applied in SCADA-enabled IoT applications," *Computers & Electrical Engineering*, vol. 120, p. 109828, 2024.
- [3] M. Abdelkhalek and M. Govindarasu, "ML-based anomaly detection system for DER DNP3 communication in smart grid," in *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2022, pp. 209–214.
- [4] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, p. e4150, 2021.
- [5] P. Hermosilla, S. Berrios, and H. Allende-Cid, "Explainable ai for forensic analysis: A comparative study of shap and lime in intrusion detection models," *Applied Sciences*, vol. 15, no. 13, p. 7329, 2025.
- [6] IEEE, "IEEE Standard for Electric Power Systems Communications—Distributed Network Protocol (DNP3)," *IEEE Std 1815-2012*, Oct. 2012.
- [7] Electric Power Research Institute, "EPRI Security Architecture for the Distributed Energy Resources Integration Network: Risk-Based Approach for Network Design," Electric Power Research Institute (EPRI), Palo Alto, CA, USA, Tech. Rep. 3002016781, Oct. 2019, [Online]. Available: <https://www.epri.com/research/products/000000003002016781>. Accessed: Aug. 17, 2026.

- [8] N. Sugunaraj, S. Ram Abayankar Balaji, B. Subash Chandar, P. Rajagopalan, U. Kose, D. Charles Loper, T. Mahfuz, P. Chakraborty, S. Ahmad, T. Kim, G. Apruzzese, A. Dubey, L. V. Strezoski, S. Ghosh, M. Jaya Bharata Reddy, H. Vardhan Padullaparti, B. Blakely, and P. Ranganathan, "Distributed energy resource management system (DERMS) cybersecurity scenarios, trends, and potential technologies: A review," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 224–277, 2026.
- [9] S. Bose, S. Bhattacharya, N. Saqib, and M. Govindarasu, "Zt-mtd: Zero trust-enabled dynamic moving target defense for der networks," in *2025 IEEE Power & Energy Society General Meeting (PESGM)*, 2025, pp. 1–5.
- [10] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber-physical system security for the electric power grid," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 210–224, 2012.
- [11] A. Ashok, M. Govindarasu, and J. Wang, "Cyber-physical attack-resilient wide-area monitoring, protection, and control for the power grid," *Proceedings of the IEEE*, vol. 105, no. 7, pp. 1389–1407, 2017.
- [12] Ö. Sen, A. Mahjoob, M. Neumüller, A. Ulbig, and S. Decker, "On cybersecurity incident response decision support system in smart grids," in *2025 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2025, pp. 1–7.
- [13] Ö. Sen, Y. Aust, M. Neumüller, I. Hacker, and A. Ulbig, "Towards a comprehensive framework for cyber-incident response decision support in smart grids," in *14th Mediterranean Conference on Power Generation Transmission, Distribution and Energy Conversion (MEDPOWER 2024)*, vol. 2024, no. 29. IET, 2025, pp. 573–578.
- [14] V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," *Frontiers in Artificial Intelligence*, vol. 8, p. 1526221, 2025.
- [15] D. Popoola, S. Bhattacharya, and M. Govindarasu, "Cysider: Cybersecurity situational intelligence framework for der networks," in *2025 Resilience Week (RWS)*, 2025, pp. 1–10.
- [16] D. Popoola, S. Bhattacharya, and M. Govindarasu, "Self-verifying anomaly detection using explainable AI for cybersecurity of DER networks," in *2026 IEEE Power & Energy Society General Meeting (PESGM)*, 2026, pp. 1–5.
- [17] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, N. Tihanyi, T. Bisztray, and M. Debbah, "Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities," *Internet of Things and Cyber-Physical Systems*, vol. 5, pp. 1–46, 2025.
- [18] H. Ji, J. Yang, L. Chai, C. Wei, L. Yang, Y. Duan, Y. Wang, T. Sun, H. Guo, T. Li, C. Ren, and Z. Li, "SEvenLLM: Benchmarking, eliciting, and enhancing abilities of large language models in cyber threat intelligence," arXiv:2405.03446 [cs.CR], May 2024, [Online]. Available: <https://arxiv.org/abs/2405.03446>.
- [19] S. A. Elsaid and A. Binbusayyis, "An optimized isolation forest based intrusion detection system for heterogeneous and streaming data in the industrial Internet of Things (IIoT) networks," *Discover Applied Sciences*, vol. 6, no. 9, p. 483, 2024.
- [20] K. Roopraj, N. Pillay, and N. Singh, "Intrusion detection in an industrial control system using unsupervised machine learning algorithms," in *2025 Conference on Information Communications Technology and Society (ICTAS)*, 2025, pp. 1–6.
- [21] E. Ozdogan, "Structured defense model against DNP3-based critical infrastructure attacks," *Arabian Journal for Science and Engineering*, vol. 50, no. 15, pp. 11 611–11 629, 2025.
- [22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Curran Associates, Inc., 2017, pp. 4768–4777.
- [23] A. E. Muhammad, K.-C. Yow, N. Baćanin-Džakula, and M. A. Khan, "L-XAIDS: A LIME-based eXplainable AI framework for intrusion detection systems," *Cluster Computing*, vol. 28, no. 10, p. 654, 2025.
- [24] B. Jordan, R. Piazza, and T. Darley, "STIX Version 2.1," OASIS Standard, Jun. 2021, [Online]. Available: <https://docs.oasis-open.org/cti/stix/v2.1/os/stix-v2.1-os.html>. Accessed: Aug. 17, 2026.
- [25] B. Jordan and D. Varner, "TAXII Version 2.1," OASIS Standard, Jun. 2021, [Online]. Available: <https://docs.oasis-open.org/cti/taxii/v2.1/os/taxii-v2.1-os.html>. Accessed: Aug. 17, 2026.
- [26] MITRE, "MITRE ATT&CK for industrial control systems: ICS Matrix," MITRE ATT&CK, 2026, [Online]. Available: <https://attack.mitre.org/matrices/ics/>. Accessed: Aug. 17, 2026.
- [27] MITRE, "Common Attack Pattern Enumeration and Classification (CAPEC)," MITRE, CAPEC List Version 3.9, 2026, [Online]. Available: <https://capec.mitre.org/>. Accessed: Aug. 17, 2026.
- [28] S. Ren and S. Chen, "Large language models for cybersecurity intelligence, threat hunting, and decision support," *Computer Life*, vol. 13, no. 3, pp. 39–47, 2025.
- [29] S. Srinivas, B. Kirk, J. Zendejas, M. Espino, M. Boskovich, A. Bari, K. Dajani, and N. Alzahrani, "AI-Augmented SOC: A survey of LLMs and agents for security automation," *Journal of Cybersecurity and Privacy*, vol. 5, no. 4, p. 95, 2025.
- [30] Y. Yigit, M. A. Ferrag, M. C. Ghanem, I. H. Sarker, L. A. Maglaras, C. Chrysoulas, N. Moradpoor, N. Tihanyi, and H. Janicke, "Generative AI and LLMs for critical infrastructure protection: Evaluation benchmarks, agentic AI, challenges, and opportunities," *Sensors*, vol. 25, no. 6, p. 1666, 2025.
- [31] A. Ali and M. C. Ghanem, "Beyond detection: Large language models and next-generation cybersecurity," *SHIFRA*, vol. 2025, pp. 81–97, 2025.
- [32] Anthropic, "Model Context Protocol Specification: Architecture," Model Context Protocol, revision 2024-11-05, Nov. 2024, [Online]. Available: <https://modelcontextprotocol.io/specification/2024-11-05/architecture>. Accessed: Aug. 17, 2026.
- [33] P. Radoglou-Grammatikis, V. Kelli, T. Lagkas, V. Argyriou, and P. Sargiannidis, "DNP3 Intrusion Detection Dataset," IEEE DataPort, Nov. 2022, [Online]. Available: <https://ieee-dataport.org/documents/dnp3-int-rusion-detection-dataset>. Accessed: Sept. 2, 2026.