Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explanation-Surface Governance in XAI-Enabled Network Intrusion Detection under Resource Constraints

Tran Duc Le

let@uwstout.edu

University of Wisconsin–Stout

Truong Duy Dinh

Posts and Telecommunications Institute of Technology

Phuc Hao Do

Danang Architecture University

Van Dai Pham

Swinburne Vietnam, FPT University

Research Article

Keywords: Explainable AI, intrusion detection, service management, explanation debt, exposure control, adversarial security

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9800570/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explanation-Surface Governance in XAI-Enabled Network Intrusion Detection under Resource Constraints

Tran Duc Le [1], Truong Duy Dinh [2*], Phuc Hao Do [3], Van Dai Pham [4]

[1]University of Wisconsin-Stout, Menomonie, WI 54751, USA.
[2]Posts and Telecommunications Institute of Technology, Hanoi, Vietnam.
[3]Danang Architecture University, Da Nang, Vietnam.
[4]Swinburne Vietnam, FPT University, Hanoi, Vietnam.

*Corresponding author(s). E-mail(s): duydt@ptit.edu.vn;
Contributing authors: let@uwstout.edu; haodp@dau.edu.vn;
daipv11@fe.edu.vn;

Abstract

Explainable artificial intelligence (XAI) helps operators interpret network-intrusion alerts, but generating explanations consumes compute, introduces queueing delay, accumulates explanation debt, and risks exposing defensive classifier decision boundaries. Current XAI-IDS research concentrates almost exclusively on semantic interpretability, ignoring these operational service-management challenges under resource constraints and adversarial explanation-demand inflation. This paper introduces XAI-Gate, a novel, detector-agnostic online service-governance framework that dynamically selects among seven explanation actions—suppress, coarse, delay, full local, offload, audit-only, or redact—based on packet pressure, explanation debt, offload belief, suspicion state, and leakage-budget constraints. We establish XAI-SurfaceBench, a rigorous stress-test benchmark anchored on KDDCup99, UNSW-NB15, and TON_IoT score-stream datasets to evaluate performance across nine traffic/uncertainty regimes. In the default adversarial explanation-demand inflation regime, XAI-Gate maintains absolute exposure feasibility (zero budget violations) while achieving a high-risk explanation coverage of $\mathbf{0.800 \pm 0.000}$ and reducing mean explanation debt by 57.3% compared to conventional threshold policies. Our

findings demonstrate that XAI-Gate serves as a robust control plane that balances the operational frontier between packet QoS, operator interpretability, and information exposure without requiring classifier re-training.

Keywords: Explainable AI, intrusion detection, service management, explanation debt, exposure control, adversarial security

1 Introduction

Although machine-learning-based network intrusion detection systems (NIDS) automate threat identification in IoT, edge, and enterprise security environments, operators must interpret these high-volume alerts under severe operational time pressure. Because many high-performing detectors expose little of their internal decision logic, explainable artificial intelligence (XAI) has become a central mechanism for making security predictions more transparent, auditable, and actionable. Foundational post-hoc methods such as LIME and SHAP established a practical pattern for explaining black-box predictions [1, 2], and recent XAI-IDS studies have extended this pattern to intrusion detection, IoT defense, feature attribution, analyst trust, and model debugging [3–10]. This progress is important, but it also creates a new operational dependency: once explanations become part of the defensive workflow, the defense system must decide not only what to detect, but also what to explain, when to explain it, at what fidelity, and under which service constraints.

The dominant XAI-IDS framing is therefore incomplete for live network defense. In much of the literature, an explanation is treated as a trust artifact appended to a detector output. In a resource-constrained NIDS, however, an explanation is also a service output. It consumes CPU, memory, and possibly accelerator resources; it can add delay to the operator-facing alert stream; it can create an explanation backlog during bursts; it can require offloading to a fog or cloud tier; and it can reveal information about feature importance, decision boundaries, or model behavior. These issues are not peripheral engineering details. Edge and IoT IDS studies already show that deployment feasibility depends on latency, memory, compute, and tier placement [11–13], while adversarial XAI research shows that explanation channels can be manipulated, fooled, or exploited as security-relevant interfaces [14–17]. The unresolved question is how an XAI-enabled defender should govern explanation generation itself when explanation demand is bursty, service capacity is finite, and explanation exposure is security-sensitive.

This paper reframes that unresolved question as *explanation surface management*. The explanation surface is the operator-visible or externally consequential interface through which alert explanations are admitted, delayed, downgraded, redacted, offloaded, logged, or suppressed. This surface can fail even when the underlying detector and explainer remain technically correct. For example, an attacker can generate borderline traffic that repeatedly triggers costly explanations, thereby inflating explanation demand, increasing queueing delay, consuming local compute, pushing work to slower offload tiers, or saturating analyst attention. Conversely, a defender that

always suppresses explanations under load may preserve packet service while losing critical operator context for high-risk alerts. Thus, XAI-enabled defense creates a coupled service-management problem: the defender must balance packet QoS, explanation QoS, explanation debt, leakage exposure, offload uncertainty, and adversarial demand pressure in the same decision loop.

To address this problem, we introduce *XAI-Gate* (as shown in Fig. 1), a detector-agnostic policy layer that sits after the intrusion detector and before explanation delivery. XAI-Gate does not require a new IDS model or a new explanation algorithm. Instead, given an alert score, estimated explanation cost, packet-service pressure, explanation backlog, offload belief, suspicion state, and leakage budget, XAI-Gate selects an explanation action from a managed action set: no explanation, coarse explanation, delayed explanation, full local explanation, offloaded explanation, audit-only logging, or redacted high-suspicion explanation. This design moves XAI from a passive interpretability add-on to an actively governed network-security service. It also aligns the contribution with network and service management: the central object is not detector accuracy alone, but the joint operating frontier among security utility, explanation utility, latency, resource use, and bounded information exposure.

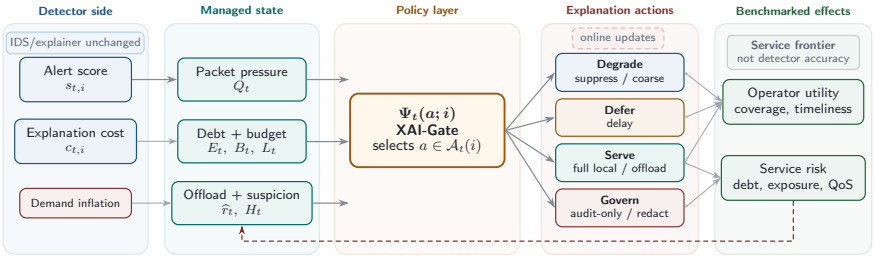


Fig. 1 Mechanism view of XAI-Gate. The gate does not alter the detector or explainer; it manages the explanation surface by mapping alert scores and service state into suppression, coarse explanation, delay, full local explanation, offload, audit-only logging, or redaction. The feedback loop is the benchmarked service effect on coverage, debt, exposure, and packet QoS.

The contributions of this paper are as follows:

- *Explanation-surface management formulation.* We formulate XAI-enabled network defense as an explanation-surface management problem in which post-hoc explanations are finite-capacity, delay-sensitive, and adversarially triggerable security-service outputs. The formulation defines the managed state, action space, service constraints, and explanation-demand inflation threat model.
- *XAI-Gate online control policy.* We develop a detector-agnostic policy that selects suppress, coarse, delay, full local, offload, audit-only, or redacted explanation actions using packet pressure, explanation debt, leakage budget, RTT/offload belief, and suspicion of explanation-demand inflation. The implementation enforces a hard

cumulative exposure cap by downgrading or delaying explanations when a high-score bucket would otherwise overshoot the remaining exposure budget. The method governs explanation delivery and does not require training a new IDS or explainer.
- *Formal service-boundary analysis.* We derive dominance, offload-preference, suspicion-sensitive fidelity, and debt-feasibility conditions that identify when full local explanations are dominated, when offloading becomes preferable, and when explanation debt can threaten packet-service QoS. These conditions make the control logic inspectable rather than only heuristic.
- *XAI-SurfaceBench implementation and evidence.* We implement a reusable no-new-IDS-training stress-test suite for explanation-demand inflation, bursty and non-Markovian traffic, transient overload, RTT uncertainty, detection-matrix perturbation, explanation-cost variation, and ablations of debt, leakage, offload, suspicion, and fidelity-control states. The executed suite reports mean and 95% confidence intervals over 20 seeds, includes tuned-baseline and tuned-XAI-Gate envelopes, and adds KDDCup99, UNSW-NB15, TON_IoT, de-duplicated UNSW, and paired statistical follow-up checks as validation realism rather than as the source of novelty.

The remainder of the paper is organized as follows. Section 2 reviews XAI-IDS, service-management, and adversarial XAI literature and identifies the explanation-surface gap. Section 3 formalizes the XAI-Gate state, action space, and objective. Section 4 describes the evaluation design, baselines, and stress regimes. Section 5 reports the executed XAI-SurfaceBench results. Section 6 discusses the validity bounds of our findings, and Section 7 concludes the paper.

2 Related Work and Gap

2.1 Explainable Intrusion Detection

Explainable AI entered network defense as a response to a practical trust deficit: high-performing machine-learning detectors can flag malicious behavior, but their internal evidence is often opaque to analysts who must triage alerts, justify containment actions, and tune defensive policies. General-purpose explanation methods such as LIME and SHAP established the post-hoc explanation pattern that many cybersecurity studies later adopted: a trained detector produces a prediction, and a separate explainer attributes that prediction to features, samples, rules, or local surrogate behavior [1, 2]. Cybersecurity surveys show that this pattern has spread across intrusion detection, malware analysis, phishing detection, botnet detection, digital forensics, and other security applications [17–20]. Within intrusion detection and Internet traffic analysis specifically, XAI surveys organize the field around explanation scope, explanation method, trustworthiness, evaluation metrics, and the need to align explanations with operator tasks rather than only with model diagnostics [4, 5, 21, 22]. This body of work strongly motivates explanations, but its dominant abstraction remains model-centric: explanations are treated as interpretability artifacts attached to detection decisions, not as network-service outputs that compete for compute, memory, delay budget, and exposure budget.

Recent XAI-IDS frameworks deepen this model-centric line of work by integrating explanation methods into end-to-end intrusion detection pipelines. Abou El Houda et al. develop an explainable deep-learning IDS for IoT networks and explicitly frame the problem as one of making IDS decisions trustworthy to security stakeholders [3]. Keshk et al. and Sharma et al. similarly combine deep intrusion detection with explanation mechanisms for IoT settings, while Arreche et al. and Gaspar et al. benchmark model/explainer combinations to expose model-specific, attack-specific, and method-specific behavior [6–8, 23]. Similarly, JNSM studies explore AI explainability via data cleaning [24] and prediction-explanation separation [10] to analyze intrusions. These studies are valuable because they show that explanations can support feature understanding, model debugging, and analyst confidence. However, they typically assume that, once an alert deserves interpretation, the system can generate and deliver an explanation. The operational question asked in this paper is different: when explanations are expensive, delayed, bursty, and potentially revealing, which alerts should be explained now, which should receive coarse or redacted explanations, which should be deferred, and which should be logged only for audit?

2.2 Network and Service Management for AI-Enabled Defense

While the network and service management literature establishes theoretical frameworks for resource allocation, traffic classification, inspection, and edge-cloud offloading [13, 25], these models do not address the schedulability, queueing backlog, or exposure limits of explanation channels. In mobile traffic classification, Nascita et al. use XAI to understand and improve multimodal deep-learning architectures, emphasizing global interpretation and the relationship between model behavior and domain knowledge [26]. Their later work uses XAI to simplify multimodal multitask traffic classifiers while preserving or improving performance, reliability, and feasibility, thereby connecting explanation to deployability rather than only to transparency [27]. In adaptive defense, DIVERGENCE models traffic inspection and moving-target defense as a reinforcement-learning-driven control problem that trades detection, scalability, and proactive defense under hostile conditions [25]. These studies establish a service-management mindset: defensive intelligence must be evaluated not only by classification accuracy, but also by feasibility, latency, resource use, and adaptive operation in live networks.

Intrusion Detection as a Service (IDaaS) makes this management view even more explicit. Lai et al. formulate task assignment and capacity allocation for ML-based IDS in a multi-tier architecture, mapping preprocessing, binary detection, and multi-class detection across edge, fog, and cloud tiers and optimizing for delay and capacity placement [13]. Aljuhani et al. study an explainable SaaS-based IDS for resource-constrained Internet-of-Medical-Things environments, showing that XAI-enabled security services must account for edge and cloud constraints rather than assuming unlimited local computation [11]. Broader reviews of explainable deep-learning IDS for IoT likewise identify latency, compute overhead, memory pressure, and deployment feasibility as recurring limitations [12, 22]. Nevertheless, even these service-aware studies generally manage the detector, the traffic pipeline, or the placement of detection tasks. They do not make the explanation interface itself a schedulable

service endpoint with admission control, fidelity selection, backlog, offload uncertainty, and information-exposure accounting.

2.3 Adversarial and Security Risks of Explanation Interfaces

The need to manage explanations is sharpened by adversarial XAI research. Kuppa and Le-Khac show that XAI methods in cybersecurity can be attacked and that explanation outputs can support adversarial behavior, model inference, or evasion strategy construction [14]. Slack et al. demonstrate that post-hoc explainers such as LIME and SHAP can be fooled, revealing that explanation trust can be manipulated independently of underlying model behavior [15]. Broader surveys of adversarial XAI further show that explanation methods introduce their own attack surface, including attacks on explanation fidelity, robustness, stability, privacy, and user trust [16, 17, 21]. The implication for XAI-enabled IDS is not merely that explanations may be inaccurate. It is that an explanation channel is an observable interaction surface through which defenders release information, consume computational resources, and shape operator attention.

Existing adversarial XAI work, however, concentrates primarily on explanation content and model-facing exploitation: how an attacker manipulates explanations, infers sensitive model behavior, or uses explanation feedback to craft better attacks [14–16]. That focus leaves a complementary service-level attack underdeveloped. An attacker may not need to corrupt the explainer or extract a model; it may be enough to generate traffic that repeatedly triggers expensive explanations, thereby increasing explanation backlog, delaying operator-visible evidence, consuming local compute, or forcing offload to a slower or less trusted tier. In such a scenario, the vulnerability is not only in the semantic correctness of each explanation. It is in the absence of a policy that decides when the explanation surface should be opened, degraded, delayed, redacted, or closed under adversarial pressure.

2.4 Research Gap and Position of XAI-Gate

The gap addressed by XAI-Gate is therefore narrower, sharper, and more operational than the generic claim that XAI is underexplored for IDS. The literature already contains substantial work on explaining IDS decisions [3–8, 23], improving traffic-classification models with XAI [22, 26, 27], allocating IDS tasks across edge, fog, and cloud resources [11, 13, 28], analyzing adversarial weaknesses in explanation methods [14–16], and studying when explanations should be selectively provided in general machine-learning settings [29]. What remains missing is an explanation-surface management layer that treats explanations as finite-capacity, delay-sensitive, information-bearing, and attackable service outputs in network defense. Such a layer must jointly reason about packet-service pressure, explanation utility, explanation debt, offload belief, operator burden, and leakage budget.

XAI-Gate is positioned to fill this missing layer. Rather than proposing another detector, another explainer, or another accuracy benchmark, it asks how an already deployed IDS should govern explanation delivery when explanation demand is bursty,

resources are finite, and adversaries can manipulate the timing and volume of explanation requests. This reframing creates a contribution that is compatible with the TNSM service-management tradition: the paper studies the management of an operational network-security service, not merely the interpretability of a classifier. Selective-explanation work is the closest conceptual comparator because it asks when an explanation should be provided [29]; XAI-Gate imports that selection logic but couples it to packet-service pressure, explanation debt, offload RTT, cumulative exposure accounting, and adversarially inflated alert demand. Similarly, resource-aware IDS and traffic-management work motivates capacity and offload control, but usually controls detector or inspection placement rather than the explanation interface itself. The central claim boundary follows directly from this coupling. XAI-Gate should be judged by the joint frontier among packet QoS, explanation QoS, explanation debt, offload cost, and bounded explanation exposure relative to unguided explanation policies, not by detector-accuracy ranking.

Table 1 and the expanded positioning matrix in the appendix compare XAI-Gate with foundation XAI, XAI-IDS surveys, traffic and service-management papers, adversarial-XAI work, selective explanations, and operator-facing SOC-XAI studies. The surviving gap is not explanation generation itself, but joint explanation-surface governance under finite capacity, debt, exposure, RTT uncertainty, and adversarial explanation demand.

Table 1 Capability-level positioning of XAI-Gate relative to adjacent literature streams. We compare core functional capabilities across seven dimensions: • denotes full/central capability, ○ denotes partial or indirect support, and − denotes that the capability is not modeled or addressed.

Literature Stream	Explainable AI	Service Mgmt.	Finite Capacity	Queueing Debt	Variable Fidelity	Edge Offload	Exposure Control	Adv. Demand
XAI-IDS Surveys & Frameworks [3–8, 23]	•	−	○	−	○	−	○	−
Traffic/IDS XAI & Traffic Classification [22, 26, 27]	•	○	○	−	−	−	−	−
IDaaS & Traffic-Inspection Mgmt. [11, 13, 25]	○	•	•	−	−	•	−	○
Adversarial XAI & Explanation Attacks [14–16]	•	−	−	−	○	−	•	○
Selective & Operator-Facing Explanations [29–31]	•	○	○	−	○	−	○	−
XAI-Gate	•	•	•	•	•	•	•	•

3 Problem Formulation and Proposed Methodology

This section develops XAI-Gate as a design-science artifact for the practical problem of governing explanations in XAI-enabled network defense. Following the design-science view that an artifact should address a relevant operational problem and be evaluated through rigorous utility-oriented assessment [32], the artifact here is not a new detector or explainer. It is a detector-agnostic decision layer that controls the explanation surface exposed by an existing NIDS. Mathematically, the problem is modeled as a constrained stochastic control problem over packet queues, explanation debt, offload uncertainty, and leakage exposure. This choice is natural because the defender repeatedly chooses explanation actions under uncertain arrivals and delayed consequences; such sequential decision problems are commonly represented as Markov decision processes (MDPs) and constrained MDPs when long-run resource constraints must be respected [33, 34].

3.1 Threat Model and Design Boundary

We do not assume that the attacker considered in this paper compromises the IDS, alters the explainer implementation, or directly observes internal model parameters. Instead, the attacker can shape or time traffic so that the defender receives a larger number of borderline, suspicious, or explanation-worthy alerts. This *explanation-demand inflation* can increase explanation cost, explanation delay, local compute pressure, offload use, and information exposure even when the detector and explainer operate as designed. The defender observes packet pressure, alert scores, estimated explanation costs, offload measurements, and demand-pressure signals, but does not know future arrivals. This threat model deliberately separates explanation-surface management from detector training, adversarial example generation, and human trust validation. The paper therefore evaluates controlled service behavior, not analyst decision quality or detector accuracy superiority.

3.2 System Setting and Managed Artifact

Consider a time-slotted edge or gateway NIDS. At slot t, packet arrivals A_t enter a packet-service queue, a detector produces a set of alert events $\mathcal{I}_t$, and each alert $i \in \mathcal{I}_t$ is associated with a detector score $s_{t,i} \in [0,1]$ and an estimated explanation cost $c_{t,i} > 0$. We treat the detector and the explainer as given modules. XAI-Gate only decides how the explanation interface should behave after the detector fires. This separation is central to the paper's claim boundary: the proposed method governs explanation delivery and does not claim superiority of any particular classifier, feature-attribution method, or post-hoc explainer.

The managed artifact is the explanation surface. For each alert, XAI-Gate selects one action from

$$\mathcal{A} = \{0, C, D, F, O, A, R\}, \tag{1}$$

where 0 denotes no explanation, C a coarse local explanation, D delayed explanation, F full local explanation, O offloaded explanation, A audit-only logging, and R redacted explanation. Each action has four abstract profiles: local compute cost $m(a)c_{t,i}$, explanation utility $g(a, s_{t,i})$, information exposure $\ell(a, s_{t,i})$, and offload delay $r_t \mathbf{1}\{a = O\}$. These profiles can be calibrated from measurement or specified as design parameters. By construction, $g(F, s)$ is high but $m(F)$ and $\ell(F, s)$ are also high, whereas R and A reduce exposure at the cost of explanation detail. The functional role of each explanation action governed by XAI-Gate is summarized in Table 2.

To avoid notation ambiguity, A_t always denotes packet arrivals, whereas the unsubscripted action label A denotes audit-only logging.

The exposure score $\ell(a, s)$ is an accounting proxy over five disclosure channels: feature-attribution disclosure, model-behavior cues, policy-threshold cues, offload data-path disclosure, and audit-trace disclosure. The offload action is modeled as sending a sanitized alert feature vector, calibrated score, requested explainer type, and policy metadata to a managed fog/cloud explanation service; raw packet payload is outside the default offload path. Table 3 summarizes the resulting taxonomy. The values are not claimed as universal constants. They are the declared action-profile

Table 2 Explanation actions governed by XAI-Gate.

Symbol	Action	Service role
0	Suppress	Avoid low-value or unsafe explanation release.
C	Coarse	Disclose top feature groups to limit detail release under medium suspicion.
D	Delay	Hold explanation in queue until compute or suspicious pressure recedes.
F	Full	Release full local attribution vector for maximum detail.
O	Offload	Send alert to managed cloud/fog explainer when local queue is overloaded.
A	Audit-only	Preserve forensic trace without immediate explanation release.
R	Redacted	Limit information release under high suspicion or exposure pressure.

semantics for the reported experiments and are stress-tested through profile-sensitivity analysis and local proxy timing checks.

Table 3 Exposure taxonomy for explanation-surface actions.

Action	Feature disclosure	Model behavior	Offload path	Audit trace	Exposure
none	none	none	none	none	0.00
coarse	top feature groups	low	none	none	0.10
full	full attribution vector	high	none	low	0.85
offload	sanitized alert vector	high	managed fog/cloud service	medium	1.00
audit	metadata summary	medium	none	high	0.34
redact	masked attribution	low	none	low	0.08
delay	none now	none now	none	queue marker	0.00

3.3 Problem Statement

The explanation-surface management problem is to select actions $\{a_{t,i}\}_{i\in\mathcal{I}_t}$ that preserve useful explanations for security-relevant alerts while preventing explanations from degrading packet service, creating unbounded explanation backlog, or exhausting an information-exposure budget. Formally, a policy π maps the system state S_t and alert-level features $(s_{t,i}, c_{t,i})$ to actions in $\mathcal{A}$:

$$a_{t,i} = \pi(S_t, s_{t,i}, c_{t,i}). \tag{2}$$

The state is defined as

$$S_t = (Q_t, E_t, B_t, L_t, \widehat{r}_t, H_t), \tag{3}$$

where Q_t is packet queue pressure, E_t is explanation debt, B_t is the available local explanation budget for the slot, L_t is the remaining exposure budget, $\widehat{r}_t$ is the current belief about offload round-trip time (RTT), and H_t is a suspicion score for explanation-demand inflation. In the reported benchmark, B_t is an exogenous per-slot service budget and L_t is implemented through an equivalent cumulative residual exposure budget. The inclusion of E_t, L_t, and H_t is the key modeling step: explanations

are no longer free artifacts but managed service outputs with debt, exposure, and adversarial-demand states. The complete set of core notation used to formulate the explanation-surface management problem is summarized in Table 4.

Table 4 Core notation for explanation-surface management.

Symbol	Meaning
A_t	packet arrivals in slot t
$\mathcal{I}_t$	alert set produced by the detector in slot t
$s_{t,i}$	alert score or risk score for alert i
$c_{t,i}$	estimated explanation cost for alert i
Q_t	packet queue pressure
E_t	risk- and obligation-weighted explanation debt
B_t	local explanation budget
L_t	remaining exposure budget
$\widehat{r}_t$	offload RTT belief
H_t	suspicion of explanation-demand inflation
$m(a)$	action-dependent compute multiplier
$\ell(a, s)$	information-exposure proxy
$\delta_E(a, s)$	explanation-debt increment

3.4 Queue, Debt, and Exposure Dynamics

Let C_t^{loc} be the local explanation load induced by all selected actions in slot t:

$$C_t^{\mathrm{loc}} = \sum_{i \in \mathcal{I}_t} m(a_{t,i}) c_{t,i}. \tag{4}$$

Local explanation work reduces the packet service available at the gateway. With nominal packet service rate μ_0 and interference coefficient $\eta \geq 0$, the effective packet service rate is modeled as

$$\mu_t^p = \left[\mu_0 - \eta C_t^{\mathrm{loc}}\right]^+, \tag{5}$$

where $[x]^+ = \max(x, 0)$. The interference coefficient η represents user-space compute contention. In modern high-performance gateway architectures utilizing kernel-bypass technologies like DPDK or eBPF/XDP, the data-plane packet processing is isolated on dedicated cores, decoupled from user-space control-plane applications (such as explanation engines). Under such isolated deployments, $\eta \approx 0$ as local explanation work does not degrade forwarding throughput. However, in low-cost, single-core, or highly resource-constrained edge gateways where control-plane XAI tasks and data-plane forwarding share scheduling cycles, $\eta > 0$ models the resulting throughput degradation. This simplified linear relationship bounds the worst-case shared-kernel compute interference. For a packet buffer of size $Q_{\max}$, packet overflow is

$$P_t = \left[Q_t + A_t - \mu_t^p - Q_{\max}\right]^+, \tag{6}$$

and the packet queue evolves as

$$Q_{t+1} = \min\left\{Q_{\max}, [Q_t + A_t - \mu_t^p]^+\right\}. \tag{7}$$

Explanation debt accumulates when explanations are delayed or when low-fidelity actions leave explanation work to be revisited later. It is not a synonym for FIFO queue length: a backlog item with higher risk or lower satisfied explanation obligation contributes more debt than a low-risk item already handled by audit or redaction. In the implementation, this is instantiated as

$$\delta_E(a, s) = \big(1 - g(a, s)\big)\big(\alpha_0 + \alpha_1 p_{\mathrm{hi}}(s)\big) + \alpha_D \mathbf{1}\{a = D\}, \tag{8}$$

where $p_{\mathrm{hi}}(s)$ is the high-risk probability associated with the alert-score bucket. Let μ_t^E denote the amount of debt that can be drained from unused explanation capacity. Then

$$E_{t+1} = \left[E_t + \sum_{i \in \mathcal{I}_t} \delta_E(a_{t,i}, s_{t,i}) - \mu_t^E\right]^+. \tag{9}$$

The diagnostic in Table 5 illustrates the distinction: equal alert-obligation counts can correspond to different debt, and high-risk items handled by audit can create less debt than lower-fidelity handling of the same item count.

Table 5 Debt-vs-count diagnostic. Equal alert-obligation counts can imply different explanation debt because debt is risk- and obligation-weighted.

Scenario	Items	Coverage	Debt
same count, low-risk delayed	100	0.00	102.8
same count, critical delayed	100	0.00	165.1
same count, high-risk coarse	100	0.45	62.7
same count, high-risk audit	100	0.72	31.9
same count, full explanation	100	1.00	0.0

The remaining exposure budget decreases with explanation release:

$$L_{t+1} = \left[L_t - \sum_{i \in \mathcal{I}_t} \ell(a_{t,i}, s_{t,i})\right]^+. \tag{10}$$

Offload belief and adversarial-demand suspicion are updated from observations. A simple exponentially weighted estimator is sufficient for the methodological definition:

$$\widehat{r}_{t+1} = (1 - \rho_r)\widehat{r}_t + \rho_r r_t^{\mathrm{obs}}, \tag{11}$$

$$H_{t+1} = (1 - \rho_h)H_t + \rho_h \phi_t, \tag{12}$$

where ϕ_t is a detector-side pressure signal, such as an unusual increase in borderline alerts, alert entropy, or alert volume under high packet pressure. These updates can be replaced by richer estimators in implementation, and the benchmark instantiates them through regime-level RTT and suspicion signals; the policy only requires that $\widehat{r}_t$ and H_t summarize offload uncertainty and explanation-demand inflation risk.

3.5 Constrained Objective

XAI-Gate seeks a policy that maximizes security and explanation utility while satisfying long-run service constraints. Let $u_{\mathrm{sec}}(a, s)$ denote the security value of making explanation evidence available for an alert of score s, and let $u_{\mathrm{xai}}(a, s)$ denote explanation usefulness to the operator. We assume nonnegative compute costs, exposure costs, debt increments, packet-pressure penalties, and tradeoff weights, with bounded per-alert utility terms. The finite-horizon objective is

$$U_t = \sum_{i \in \mathcal{I}_t} \big(u_{\mathrm{sec}}(a_{t,i}, s_{t,i}) + u_{\mathrm{xai}}(a_{t,i}, s_{t,i})\big), \tag{13}$$

$$X_t = \sum_{i \in \mathcal{I}_t} \ell(a_{t,i}, s_{t,i}), \tag{14}$$

$$O_t = \sum_{i \in \mathcal{I}_t} \mathbf{1}\{a_{t,i} = O\}, \quad F_t = \sum_{i \in \mathcal{I}_t} \mathbf{1}\{a_{t,i} = F\}, \tag{15}$$

$$\begin{aligned} \max_{\pi} \quad J_T(\pi) = \mathbb{E}_\pi \sum_{t=0}^{T-1} \Big[& U_t - \lambda_Q Q_t - \lambda_P P_t - \lambda_E E_t \\ & - \lambda_B C_t^{\mathrm{loc}} - \lambda_L X_t - \lambda_O \widehat{r}_t O_t - \lambda_H H_t F_t \Big]. \end{aligned} \tag{16}$$

The weights $\lambda_Q, \lambda_P, \lambda_E, \lambda_B, \lambda_L, \lambda_O, \lambda_H$ encode the service operator's tradeoff among packet QoS, explanation QoS, local compute, leakage, offload delay, and adversarial probing risk. The corresponding constrained form is

$$\max_{\pi} \quad \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \sum_{t=0}^{T-1} \sum_{i \in \mathcal{I}_t} \big(u_{\mathrm{sec}} + u_{\mathrm{xai}}\big) \tag{17}$$

$$\text{s.t.} \quad \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \sum_{t=0}^{T-1} P_t \le \epsilon_P, \tag{18}$$

$$\limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \sum_{t=0}^{T-1} E_t \le \bar{E}, \tag{19}$$

$$\limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \sum_{t=0}^{T-1} \sum_{i \in \mathcal{I}_t} \ell(a_{t,i}, s_{t,i}) \le \bar{\Lambda}, \tag{20}$$

$$C_t^{\mathrm{loc}} \le B_t + \xi_B, \quad a_{t,i} \in \mathcal{A}. \tag{21}$$

This constrained formulation clarifies what the proposed solution is solving: not maximum explanation coverage at any cost, but bounded explanation exposure and bounded explanation debt under packet-service constraints. In a finite-horizon benchmark, the long-run exposure bound is implemented as a configured cumulative cap Λ_T and reported through exposure-use and exposure-violation metrics. Because the exact

transition law may be unknown or non-stationary, XAI-Gate uses an online surrogate inspired by constrained MDP and drift-plus-penalty control, where queue-like state variables penalize decisions that push the system toward instability [34, 35].

3.6 XAI-Gate Decision Rule

For each alert, XAI-Gate computes an action score

$$\begin{aligned}\Psi_t(a;i) = {} & u_{\mathrm{sec}}(a, s_{t,i}) + u_{\mathrm{xai}}(a, s_{t,i}) - \lambda_B m(a) c_{t,i} \\ & - \lambda_E \delta_E(a, s_{t,i}) - \lambda_L \ell(a, s_{t,i}) - \lambda_O \widehat{r}_t \mathbf{1}\{a = O\} \\ & - \lambda_H H_t \mathbf{1}\{a = F\} - \lambda_Q \Delta_Q(a, c_{t,i}, Q_t),\end{aligned} \tag{22}$$

where $\Delta_Q(a, c_{t,i}, Q_t) = \eta m(a) c_{t,i} Q_t$ is the marginal packet-service pressure. This formulation represents the expected increase in queueing occupancy and drop probability resulting from scheduling the local computational load $m(a)c_{t,i}$ under a current queue pressure of Q_t. The feasible action set is state-dependent. For within-slot admission, let $L_{t,i}^{\mathrm{res}}$ denote the residual exposure budget after higher-risk alerts already admitted in slot t:

$$\mathcal{A}_t(i) = \{a \in \mathcal{A} : m(a) c_{t,i} \le B_t + \xi_B,\ \ell(a, s_{t,i}) \le L_{t,i}^{\mathrm{res}} + \xi_L\}. \tag{23}$$

If this set is empty, XAI-Gate falls back to A or 0, depending on alert score and policy configuration. In the implementation, bucketed alerts are admitted in descending risk order and a hard exposure-cap guard splits a bucket if admitting every alert at the preferred fidelity would exceed the remaining cumulative exposure budget. The residual alerts are downgraded to redacted, coarse, audit-only, delayed, or suppressed actions according to the same feasible-set logic. This guard is necessary because a score-bucket decision can otherwise satisfy the per-alert utility rule but overshoot the cumulative exposure constraint when many high-risk alerts arrive together. The selected action is

$$a_{t,i}^{\star} = \arg \max_{a \in \mathcal{A}_t(i)} \Psi_t(a;i). \tag{24}$$

This rule has an interpretable dominance structure. A full local explanation is dominated by a cheaper or safer action a whenever

$$\begin{aligned}\Delta U(F, a; s) < {} & \lambda_B \big(m(F) - m(a)\big) c + \lambda_L \big(\ell(F, s) - \ell(a, s)\big) \\ & + \lambda_E \big(\delta_E(F, s) - \delta_E(a, s)\big) + \lambda_H H_t \\ & + \lambda_Q \Delta_Q(F, c, Q) - \lambda_Q \Delta_Q(a, c, Q) \\ & - \lambda_O \widehat{r}_t \mathbf{1}\{a = O\},\end{aligned} \tag{25}$$

where $\Delta U(F, a; s) = u_{\mathrm{sec}}(F, s) + u_{\mathrm{xai}}(F, s) - u_{\mathrm{sec}}(a, s) - u_{\mathrm{xai}}(a, s)$. Equation (25) is the mathematical expression of the central intuition: high-fidelity explanations are not always optimal, even for high-risk alerts, when the marginal gain in explanation utility

is outweighed by local service pressure, exposure risk, debt dynamics, or suspicion of probing. Conversely, when s is high, Q_t and E_t are low, L_t is available, and H_t is small, the same rule admits full explanation because the service and exposure penalties are low.

3.7 Service-Boundary Analysis

The following propositions convert the policy score into operational boundaries that can be checked in analysis and stress-tested in XAI-SurfaceBench. They are not performance claims; they identify when one class of explanation action should be preferred under the stated model.

Proposition 1 (Full-explanation dominance) *For an alert with score s and cost c, if action $a \in \mathcal{A}_t(i)$ satisfies Eq. (25), then F is not the maximizer of Eq. (22) over any feasible set that contains both F and a.*

Proof Rearranging Eq. (22) gives $\Psi_t(a;i) - \Psi_t(F;i) > 0$ under Eq. (25). Therefore F cannot be selected by the argmax rule when a is feasible. The proposition follows directly from the scoring rule and does not depend on a particular detector or explainer. □

Proposition 2 (Explanation-debt feasibility) *Let $d_t^\pi = \sum_{i \in \mathcal{I}_t} \delta_E(\pi(S_t, s_{t,i}, c_{t,i}), s_{t,i})$ be the debt admitted by policy π. If, over an interval of length T, the expected debt increment satisfies $\mathbb{E}[d_t^\pi - \mu_t^E] \geq \epsilon > 0$ for every slot index in the interval, then $\mathbb{E}[E_T] \geq E_0 + T\epsilon$ before any external debt cap is imposed.*

Proof Ignoring the nonnegative projection can only decrease the lower bound because $[x]^+ \geq x$. Taking expectations in the debt recursion yields $\mathbb{E}[E_{t+1}] \geq \mathbb{E}[E_t] + \epsilon$. Summing the drift inequality over T slots gives the stated bound. This condition motivates transient dwell-time experiments near the debt-stability boundary. □

Proposition 3 (Offload preference boundary) *For a feasible offload action O and a feasible local action $a \in \{C, F, R\}$, offloading is selected over a whenever $\Psi_t(O;i) - \Psi_t(a;i) > 0$; equivalently, its net score gain after compute, packet-pressure, exposure, debt, RTT, and utility differences is positive.*

Proof The claim is obtained by writing $\Psi_t(O;i) - \Psi_t(a;i) > 0$ using Eq. (22). The sign of each term depends on the configured action profiles: offloading may reduce local compute and packet pressure, but it can increase RTT or exposure. This boundary is the analytic basis for RTT-belief robustness tests. □

Proposition 4 (Suspicion-sensitive fidelity reduction) *Assume $\lambda_H > 0$ and let $a \in \{C, A, R\}$ be feasible. If*

$$H_t > \frac{\Psi_t(F; i)|_{H_t=0} - \Psi_t(a; i)|_{H_t=0}}{\lambda_H}, \tag{26}$$

then a has a higher policy score than full local explanation F.

Proof In Eq. (22), H_t penalizes the full local explanation term through $-\lambda_H H_t \mathbf{1}\{a = F\}$ while C, A, and R do not receive that full-explanation probing penalty. Solving $\Psi_t(a; i) > \Psi_t(F; i)$ for H_t gives the stated threshold. □

3.8 Online Method

Algorithm 1 summarizes the proposed methodology. The algorithm is intentionally modular: any detector can supply $s_{t,i}$, any explainer can instantiate the cost and exposure profiles, and any deployment can tune the weights or constraints. The novelty lies in the gate that makes explanation delivery conditional on the network state and adversarial surface state.

Algorithm 1 XAI-Gate Online Explanation Surface Management

Initialize $Q_0, E_0, L_0, \widehat{r}_0, H_0$ and policy weights λ
for each slot $t = 0, 1, \ldots, T-1$ **do**
  Observe packet arrivals A_t, queue pressure Q_t, and alert set $\mathcal{I}_t$
  Estimate $(s_{t,i}, c_{t,i})$ for each alert $i \in \mathcal{I}_t$
  Update offload belief $\widehat{r}_t$ and suspicion signal H_t
  for each alert $i \in \mathcal{I}_t$ **do**
    Construct feasible set $\mathcal{A}_t(i)$ from budget and exposure constraints
    Compute $\Psi_t(a; i)$ for each $a \in \mathcal{A}_t(i)$
    Select $a^\star_{t,i} = \arg\max_{a \in \mathcal{A}_t(i)} \Psi_t(a; i)$
    Execute explanation action $a^\star_{t,i}$
  end for
  Update Q_{t+1}, E_{t+1}, and L_{t+1} using the queue, debt, and exposure dynamics
end for

3.9 Methodological Claim Boundary

The methodology produces three testable claims. First, under finite local explanation capacity, unguided explanation policies can increase packet pressure, explanation debt, or exposure consumption when alert demand becomes bursty. Second, a state-aware policy that includes debt, offload belief, suspicion, and leakage budget is hypothesized to change the joint service frontier relative to static, threshold-only, FIFO, and budget-only policies. Third, optional public-data calibration can improve realism by mapping public labels and features into score and cost distributions, but such calibration is not a detector contribution and is not the source of novelty. These claims are evaluated

through stress regimes and baselines in Section 4; they do not require a detector-accuracy contribution.

4 XAI-SurfaceBench Evaluation Protocol

XAI-SurfaceBench is the evaluation protocol used to test whether explanation-surface control changes service behavior under controlled pressure. The core protocol requires no new IDS training: detector scores, explanation costs, and traffic regimes can be modeled or replayed as inputs to the service-management layer. Public-data anchors may train simple calibration models only to create realistic score streams; those calibration models are not proposed IDSs and are not sources of novelty. This design keeps the empirical target aligned with the contribution. The paper evaluates explanation-service behavior, not detector accuracy superiority.

4.1 Research Questions

To systematically evaluate the performance, service tradeoffs, and robustness of XAI-Gate, we define six core research questions and specify the required evidence in Table 6.

Table 6 Research questions and evidence required by XAI-SurfaceBench.

RQ	Question	Evidence required
RQ1	Does unmanaged explanation delivery degrade packet or explanation QoS under finite capacity?	Compare always-explain, threshold, FIFO, and rate-limited policies under increasing alert and explanation demand.
RQ2	Does state-aware control change the packet/explanation/exposure frontier?	Plot frontier curves across budgets and policy weights for XAI-Gate and baselines.
RQ3	Which state variables matter?	Ablate explanation debt, leakage budget, offload belief, suspicion state, and fidelity control.
RQ4	How does the policy behave under bursty, non-Markovian, and transient overload regimes?	Run stress tests with correlated bursts, self-similar arrivals, and dwell time near service boundaries.
RQ5	How does the method respond to detection-matrix and RTT-belief uncertainty?	Perturb confusion matrices, alert-score distributions, RTT estimates, and offload delays.
RQ6	Does optional public-dataset calibration alter the conclusions?	Repeat key experiments using calibrated score/error distributions when dataset resources are available.

4.2 Core No-New-IDS-Training Stress Tests

The minimum evaluation uses synthetic or replayed arrival processes, modeled detector-score streams, configurable detection matrices, and configurable explanation costs. The stress regimes are: Poisson baseline traffic, bursty traffic, self-similar

traffic, non-Markovian correlated bursts, transient overload near the debt-stability boundary, adversarial explanation-demand inflation, RTT/offload uncertainty, and detection-matrix perturbation. For each regime, XAI-SurfaceBench records whether policy behavior is stable across random seeds and whether conclusions depend on a narrow parameter setting.

4.3 Baselines, Metrics, and Ablations

The rigorous evaluation of explanation-surface policies requires a comprehensive set of baselines, multi-objective service metrics, and systematic ablation studies, which are categorized in Table 7.

Table 7 Evaluation controls required before making performance claims.

Category	Required items
Baselines	Never-explain, always-explain, threshold-explain, rate-limited explain, FIFO explanation queue, static coarse/full split, static offload, budget-only explanation governance (BEX-Gov), and XAI-Gate.
Service metrics	Packet drop probability, packet delay, packet queue occupancy, and local budget violation rate.
Explanation metrics	Explanation coverage, high-risk explanation coverage, explanation delay, explanation debt, and explanation fidelity mix.
Security/exposure metrics	Exposure-budget use, audit rate, redaction rate, full-explanation release under suspicion, and attacker-trigger gain.
Resource/offload metrics	Local explanation load, offload rate, RTT penalty, and cloud/fog explanation traffic.
Ablations	Remove debt, leakage budget, fidelity control, offload action, audit/redaction action, suspicion state, RTT robustness, and replace XAI-Gate with threshold-only scoring.

All stochastic policy comparisons use shared random seeds, identical traffic traces, identical alert-score streams, and the same regime configuration for every policy in a run block. The baseline set is intentionally heterogeneous: it includes degenerate policies, threshold and rate controls, FIFO queueing, static fidelity, static offload, and budget-only governance. The frontier stage additionally sweeps local and exposure budgets so that the comparison is not based on one handpicked operating point. The paper therefore interprets results as service-frontier evidence rather than as proof that one policy is universally dominant across all possible baseline tuning choices. The full execution inventory of the XAI-SurfaceBench stages, including the corresponding seeds, slots, run counts, and reported evidence, is detailed in Table 8.

The main empirical object is the joint frontier among packet QoS, explanation QoS, and exposure control. A single scalar accuracy score would not answer the paper's research question. XAI-SurfaceBench therefore reports frontier curves, debt trajectories, attack-intensity sweeps, and ablation panels. For stochastic regimes, experiments should use shared random seeds across policies and report confidence intervals or bootstrap intervals.

The benchmark records retain fixed seeds, stage-specific configurations, run-level summaries, and table-generation inputs for each reported stage. This traceability is

Table 8 XAI-SurfaceBench execution inventory used before making performance claims.

Stage	Evidence role	Seeds/slots	Runs	Reported evidence
0	Sanity checks for determinism, schemas, finite metrics, and non-negative state variables	2 / 200	162	Determinism, schema, and run-trace records
1	Main nine-policy comparison across mandatory regimes	20 / 5,000	1,620	Run summaries and main comparison table
2–3	Demand-inflation and budget-frontier sweeps	20 / 5,000	9,000	Sweep summaries and frontier figures
4–5	Mechanism ablations and robustness sweeps	20 / 5,000	5,040	Ablation and robustness summaries with figures
6–8	Tuned baseline, action-profile sensitivity, and XAI-Gate tuning envelopes	20 / 5,000	6,000	Envelope, sensitivity, and tuning tables
9	KDDCup99 no-training calibrated-anchor score stream	20 / 5,000	720	Public-data score stream and calibration table
10–11	Action-profile provenance, debt/backlog, and exposure diagnostics	deterministic/proxy	19 rows	Provenance and diagnostic tables
12–16	UNSW-NB15 anchors, de-duplicated rerun, XAI-Gate tuning, tuned-baseline envelope, and paired follow-up	20 / 5,000	6,080	Real-data score streams, calibration tables, and paired statistics
17	TON_IoT calibrated-anchor score stream	20 / 5,000	720	IoT/IIoT score stream and calibration tables

part of the evidence boundary: only results reproducible from the XAI-SurfaceBench execution records are reported.

4.4 Calibration Layer and Executed Anchor

The evaluation layer accepts calibrated score streams with columns for score, true label, optional attack family, and optional cost proxy. This paper executes three public-data anchors and one de-duplicated UNSW-NB15 follow-up. The first anchor uses the KDDCup99 SA subset available through scikit-learn [36, 37], where a transparent traffic/error-feature score proxy is used only to form score-bucket and high-risk priors. The second anchor is a modern UNSW-NB15 check [38], where a balanced linear logistic calibration model is trained on the public train split and applied to the public test split to generate a 175,341-row score stream. Because the public UNSW files contain duplicate feature rows and train-test feature overlap, the calibration procedure also executes a de-duplicated UNSW rerun: it keeps the first row per feature hash in each split and removes test rows whose feature hash appears in the de-duplicated training split, yielding a 99,738-row score stream with zero remaining exact or feature-hash train-test overlap. The third anchor is TON_IoT [39]: the calibration procedure uses public network-flow records, removes feature-duplicate rows, performs a stratified train/test split after deduplication, and trains a balanced linear SGD logistic calibration model to generate a 57,143-row held-out score stream with zero train-test

feature-hash overlap. The calibration-quality metadata for each anchor are as follows. Because the linear SGD calibration models are not proposed IDS contributions, these metrics should be read as score-stream quality descriptors rather than as detector performance claims. The UNSW-NB15 raw calibration yields ROC AUC 0.976, average precision 0.988, Brier score 0.086, and expected calibration error (ECE, 10 equal-width bins) 0.097. The de-duplicated UNSW rerun yields ROC AUC 0.955, average precision 0.946, and ECE 0.101. The TON_IoT calibration yields ROC AUC 0.995, average precision 0.998, Brier score 0.014, and ECE 0.021. The TON_IoT scores are better calibrated than the UNSW scores because the TON classification boundary is substantially easier under the available features, but this does not change the contribution: XAI-Gate governs the explanation surface, and the score stream is input data. No detector novelty is claimed in any anchor: the KDD proxy and UNSW/TON_IoT classifiers are calibration infrastructure for score and label realism, not proposed IDS models. This distinction prevents the no-new-IDS-training design from being read as a ban on calibration; it means that the proposed contribution does not depend on training a better detector. Larger datasets such as CICIoT2023, CICIDS2017, and CSE-CIC-IDS2018 [40–42] remain useful future anchors, but they are not required for the central claim because XAI-Gate governs explanation delivery rather than detector learning.

5 Empirical Evaluation and Results

This section reports the executed XAI-SurfaceBench evidence. The objective is not detector-accuracy superiority, because the detector-score stream is a stress-test input rather than the paper's proposed model. The objective is to test whether explanation-surface governance changes service behavior under finite explanation capacity, bounded exposure, non-stationary traffic, offload uncertainty, and adversarial explanation-demand inflation. XAI-SurfaceBench is therefore used as a controlled stress-test suite, not yet as a mature community benchmark. The suite was executed with 20 random seeds, 5,000 slots per stochastic run, nine mandatory policies, nine mandatory regimes, and shared configuration files for all baselines. Stage 0 used two seeds and 200 slots to verify deterministic rerun equality, finite metrics, schema validity, and non-negative queue and debt states.

5.1 Main Stress Regime Under Explanation-Demand Inflation

Table 9 reports the main adversarial explanation-demand regime. The most important observation is that no policy dominates all objectives. Always explaining maximizes high-risk explanation coverage, but it consumes 2.864 times the configured exposure budget and violates the exposure budget in 65.1% of slots. Never explaining avoids exposure but leaves all high-risk explanation coverage at zero and accumulates large explanation debt. Threshold and static coarse/full policies preserve high-risk coverage near 0.884, but both exceed the exposure budget by approximately 39% and violate the exposure constraint in about 28% of slots. XAI-Gate occupies a different point on the frontier: it achieves high-risk coverage of 0.800 ± 0.000, reduces mean explanation debt to 1561.6 ± 22.8, keeps exposure use at 0.995 ± 0.001, and records no exposure-budget

violations. Compared with threshold explanation, XAI-Gate gives up approximately 8.4 percentage points of high-risk coverage, but reduces mean explanation debt by 57.3% and removes exposure-budget violations. This is the core empirical interpretation: XAI-Gate should not be claimed as a universal explanation-coverage maximizer; it should be claimed as a service-governance policy that trades some coverage for lower debt and bounded exposure.

Table 9 Main XAI-SurfaceBench results under adversarial explanation-demand inflation. Values are mean ± 95% confidence interval over 20 seeds.

Policy	High-risk coverage	Mean debt	Exposure use	Exposure violation	Local load
Never explain	0.000 ± 0.000	20063.8 ± 51.1	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Always explain	1.000 ± 0.000	0.0 ± 0.0	2.864 ± 0.006	0.651 ± 0.001	12.131 ± 0.024
Threshold explain	0.884 ± 0.000	3660.9 ± 30.8	1.388 ± 0.004	0.280 ± 0.002	6.441 ± 0.017
Rate limited	0.877 ± 0.000	5005.0 ± 69.0	1.411 ± 0.004	0.292 ± 0.002	6.715 ± 0.012
Static offload	0.752 ± 0.000	3075.4 ± 22.6	1.616 ± 0.005	0.381 ± 0.002	1.848 ± 0.004
Budget-only explanation governance	0.774 ± 0.000	5327.3 ± 32.9	0.918 ± 0.000	0.000 ± 0.000	5.350 ± 0.008
XAI-Gate	$\mathbf{0.800 \pm 0.000}$	$\mathbf{1561.6 \pm 22.8}$	$\mathbf{0.995 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$	6.842 ± 0.020

The default adversarial regime did not produce packet drops for the policies in Table 9; the bottleneck in this operating point is explanation debt and exposure, not packet loss. This negative result is important because it prevents an inflated packet-QoS claim. Packet drop appears only in robustness settings where service capacity and burstiness are perturbed. Thus, the proper claim is not that XAI-Gate always improves packet loss, but that it changes the explanation-service frontier while keeping exposure within budget in the default adversarial setting.

Several reported confidence intervals round to 0.000 because the action profile and policy mapping are deterministic after the shared seed fixes the alert-score stream, while the rounded coverage ratio is stable across the 20 seeds. The reported intervals are descriptive seed-level normal-approximation 95% confidence intervals, not paired significance tests. The seed-level run records retain the unrounded values, and stochastic variation remains visible in debt, exposure, queue, and delay metrics.

Table 10 makes the all-regime boundary explicit rather than relying only on the adversarial-demand case. Across the nine mandatory regimes, XAI-Gate keeps mean exposure-budget violation at zero, but it is not the highest-coverage policy in every regime. In self-similar and transient-overload traffic, finite packet-service capacity still produces packet drops, and in several regimes the best exposure-feasible comparator attains higher high-risk coverage. This table therefore supports the bounded service-management claim: XAI-Gate is useful when exposure feasibility and debt control matter jointly, but the preferred policy depends on the operator's regime and objective.

5.2 Tuned-Baseline Envelope

Because a weak baseline would overstate the method, an additional baseline-envelope analysis tunes threshold, rate-limited, FIFO, static coarse/full, static offload, and budget-only controls. For each baseline and regime, the selected operating point is the maximum high-risk coverage variant with zero exposure-budget violation; if no feasible variant exists, the highest-coverage variant is reported with its violation rate.

Table 10 All-regime boundary summary for the mandatory XAI-SurfaceBench stress regimes. The comparator is the highest high-risk-coverage non-XAI-Gate policy with zero mean exposure-budget violation in that regime; if all non-XAI-Gate policies violate exposure, the least-violating highest-coverage policy is reported.

Regime	XAI HR	XAI debt	XAI viol.	XAI drop	Best feasible comparator	Comp. HR	Comp. debt
Adv. demand	0.800	1561.6	0.000	0.000	Budget-only	0.774	5327.3
Bursty	0.776	104.9	0.000	0.162	Static coarse/full	0.813	1789.3
Detection perturb.	0.760	0.5	0.000	0.000	Rate-limited	0.815	0.4
Markov-mod.	0.754	728.6	0.000	0.165	Budget-only	0.780	4209.2
Non-Markov.	0.771	123.0	0.000	0.137	Threshold	0.814	2463.2
Poisson	0.777	0.2	0.000	0.000	Rate-limited	0.810	0.3
RTT uncert.	0.853	0.4	0.000	0.000	Rate-limited	0.826	0.4
Self-similar	0.478	7181.5	0.000	0.355	Budget-only	0.740	11068.2
Transient overload	0.749	2346.0	0.000	0.247	Static coarse/full	0.821	4017.5

Table 11 shows the adversarial-demand envelope. XAI-Gate remains the strongest feasible policy in this envelope: it achieves high-risk coverage of 0.800 with mean debt 1536.4 and no exposure-budget violations. The best feasible threshold policy reaches only 0.549 coverage with mean debt 10395.0, while budget-only explanation governance reaches 0.774 coverage with mean debt 5294.8. Always-explain and FIFO can reach full coverage, but only by violating the exposure budget in about 65% of slots. This result changes the interpretation of Table 9: the default threshold policy is a useful high-coverage comparator, but after exposure-feasible tuning, XAI-Gate is not merely beating a weak threshold baseline.

Table 11 Tuned baseline envelope under adversarial explanation-demand inflation. Each baseline is selected by maximum high-risk coverage subject to zero exposure-budget violation; infeasible high-coverage baselines are retained with their violation rates.

Policy	Selected variant	High-risk coverage	Mean debt	Exposure use	Exposure violation
Always explain	default	1.000	0.0	2.869	0.652
FIFO explanation	quota 24	1.000	0.0	2.868	0.651
Budget-only explanation governance	pressure 0.72	0.774	5294.8	0.918	0.000
Rate limited	quota 4	0.588	17510.8	0.877	0.000
Static coarse/full	high 0.86, medium 0.45	0.605	7700.0	0.689	0.000
Static offload	high 0.86, medium 0.55	0.544	7576.9	0.778	0.000
Threshold explain	threshold 0.84	0.549	10395.0	0.592	0.000
XAI-Gate	default	**0.800**	**1536.4**	**0.995**	**0.000**

5.3 Tuning, Calibration Anchors, and Paired Follow-up

To avoid giving baselines a larger tuning budget than the proposed policy, a separate stage sweeps XAI-Gate weights and local/exposure budgets under the same feasibility criterion used for tuned baselines. Table 12 reports the adversarial-demand envelope while distinguishing fixed-exposure variants from expanded-exposure operating points. The default XAI-Gate setting is already feasible, with high-risk coverage of 0.801, mean debt of 1549.8, exposure use of 0.995, and no exposure-budget violations.

Among variants that preserve the default exposure envelope, the strict-suspicion setting reaches 0.817 high-risk coverage, while balanced-coverage and high-local-budget settings reduce debt further without exposure violations. When the operator explicitly relaxes the exposure envelope, the high-exposure-budget variant reaches 0.835 high-risk coverage and mean debt of 1018.7. This result should not be read as a weight-only improvement over the default; it shows how the operating point changes when the service envelope itself changes.

Table 12 XAI-Gate tuning envelope under adversarial explanation-demand inflation. Rows distinguish fixed-exposure-envelope variants from variants that relax the exposure envelope; the latter should be read as a different operating envelope rather than a weight-only tuning improvement.

Rank	Variant	Envelope	HR coverage	Debt	Exposure use	Violation
1	High exposure budget	expanded exposure	0.835	1018.7	0.949	0.000
2	Strict suspicion	fixed exposure	0.817	1666.4	0.988	0.000
3	Strict exposure	fixed exposure	0.813	1609.0	0.987	0.000
4	Balanced coverage	fixed exposure	0.812	1209.7	0.987	0.000
5	High local budget	fixed exposure	0.809	981.0	0.968	0.000
6	Strict RTT	fixed exposure	0.808	1542.8	0.994	0.000
7	Coverage heavy, budget 20	expanded exposure	0.805	1240.1	0.985	0.000
8	Coverage heavy	fixed exposure	0.801	1515.2	0.997	0.000

Table B2 (relegated to Appendix B) reports the KDDCup99 no-training calibrated-anchor check. Under the public score-stream priors, XAI-Gate obtains high-risk coverage of 0.790, mean debt of 872.0, exposure use of 0.947, and no exposure-budget violations. Threshold and budget-only governance remain exposure-feasible in this anchor but explain far fewer high-risk alerts because the proxy score distribution is sparse in high-score buckets. This anchor should be read narrowly: it reduces the risk that the entire evaluation is an artifact of one synthetic score-bucket profile, but it does not establish detector accuracy, modern-dataset generality, or deployment readiness.

Specifically, under the fixed exposure budget, the UNSW-NB15 raw calibration stream (175,341 test-split rows, positive-label rate 0.681, calibration metadata of ROC AUC 0.976 and average precision 0.988) shows that XAI-Gate records zero exposure-budget violations across all four tested real-data regimes, while maintaining a clear trade-off profile. In the default adversarial explanation-demand regime, although budget-only governance achieves higher high-risk coverage (0.737 vs. 0.670), it accumulates nearly double the mean explanation debt (7946.8 vs. 4357.4). Threshold explanation reaches higher coverage (0.927), but only by violating the exposure budget in 56.6% of slots with an unsustainable exposure use of 2.309. In the transient-overload regime, XAI-Gate increases high-risk coverage to 0.730 (vs. 0.683 for budget-only) while reducing mean debt from 6710.6 to 3960.0, whereas threshold explanation violates the exposure budget in 42.8% of slots. Under non-Markovian demand, XAI-Gate achieves 0.752 high-risk coverage with only 768.6 mean debt (vs. 0.672 coverage and 5792.1 debt for budget-only), while threshold explanation exhibits a 23.0% violation rate. Finally, under RTT uncertainty, XAI-Gate secures a high-risk coverage of 0.848 and near-zero mean debt (18.0) compared to budget-only's 0.675 coverage and 2543.4 debt, while threshold explanation suffers minor exposure-budget violations (0.2%). Thus, the raw UNSW-NB15 anchor stream strengthens the realism of the evidence

while reinforcing the same bounded interpretation: XAI-Gate acts as a highly effective online exposure and debt controller rather than a universal coverage-maximizing heuristic.

Table 13 reports the de-duplicated UNSW-NB15 rerun. The rerun removes 28,386 feature-duplicate train rows, 74,301 feature-duplicate test rows, and 1,302 remaining test rows that overlap the de-duplicated train split by feature hash. The resulting calibration stream has 99,738 rows, positive-label rate 0.484, ROC AUC 0.955, and average precision 0.946. The service-level interpretation is stable but more favorable to XAI-Gate than the full public test stream: under adversarial explanation-demand inflation, XAI-Gate has slightly higher high-risk coverage than budget-only governance (0.729 versus 0.715) and much lower mean debt (2866.1 versus 8220.4), while keeping zero exposure-budget violations. Threshold explanation still reaches higher adversarial coverage (0.845), but it consumes 1.687 times the exposure budget and violates the exposure constraint in 40.7% of slots. The de-duplicated run therefore removes the strongest duplicate-overlap objection without changing the claim boundary: XAI-Gate improves the exposure/debt tradeoff under finite explanation capacity, but it is not a universal high-coverage maximizer.

Table 13 De-duplicated UNSW-NB15 calibrated-anchor rerun under fixed exposure budget. The stream removes feature-duplicate rows and train-test feature overlap before calibration. HR denotes high-risk coverage, exp. denotes exposure use, and viol. denotes exposure-budget violation.

Regime	XAI HR	XAI debt	XAI exp.	XAI viol.	Budget HR	Budget debt	Budget exp.	Thresh. HR	Thresh. exp.	Thresh. viol.
Adversarial demand	0.729	2866.1	1.000	0.000	0.715	8220.4	0.893	0.845	1.687	0.407
Transient overload	0.746	3023.7	0.993	0.000	0.641	6811.3	0.564	0.734	0.932	0.000
Non-Markovian	0.771	274.4	0.906	0.000	0.628	5869.9	0.494	0.720	0.816	0.000
RTT uncertainty	0.813	0.8	0.663	0.000	0.621	2694.7	0.375	0.706	0.605	0.000

Table B3 (relegated to Appendix B) reports the TON_IoT calibrated anchor. The source stream contains 211,043 public network-flow rows, 190,474 rows after feature deduplication, and a 57,143-row held-out score stream with positive-label rate 0.779. The calibration metadata are ROC AUC 0.995 and average precision 0.998, used only to create score and cost priors. Under fixed exposure budget, XAI-Gate records zero exposure-budget violations in all four tested regimes. The result is again a tradeoff rather than a universal win: in adversarial explanation-demand inflation, budget-only governance has higher high-risk coverage (0.767 versus 0.654), but higher mean debt (6792.3 versus 5020.2), while threshold explanation reaches 0.994 coverage only with exposure-budget violations in 62.5% of slots. Under RTT uncertainty, XAI-Gate obtains higher high-risk coverage than budget-only governance (0.863 versus 0.793), much lower mean debt (98.0 versus 867.6), and zero exposure-budget violations. TON_IoT therefore broadens the public-data anchor set to an IoT/IIoT traffic source while preserving the bounded interpretation: XAI-Gate controls exposure and debt under the tested regimes, but it is not a detector-accuracy contribution or a universal coverage maximizer.

Table C4 (relegated to Appendix C) adds paired seed-level follow-up tests for the adversarial-demand comparisons. The table reports XAI-Gate minus comparator mean differences, bootstrap 95% confidence intervals over paired seed differences, and exact two-sided sign-test values. The synthetic tuned-envelope comparison supports

the main service-frontier claim: XAI-Gate improves high-risk coverage over tuned threshold and budget-only baselines while reducing mean debt. The original UNSW stream remains mixed against budget-only governance on adversarial coverage, the de-duplicated UNSW stream favors XAI-Gate over budget-only governance on both adversarial high-risk coverage and debt, and TON_IoT shows the same tradeoff pattern as the descriptive table: lower coverage than budget-only governance but lower debt and lower exposure-violation risk than threshold explanation. These paired tests are reported as follow-up evidence, not as a replacement for the descriptive confidence intervals used throughout the benchmark.

The UNSW-NB15 tuning envelope in Table 14 further separates fixed-budget tuning from budget expansion. Fixed-budget variants improve selected regimes without exposure violations; the expanded-budget adversarial row is reported separately because it changes the exposure-capacity assumption. This distinction prevents the real-data anchor from overstating the method by silently relaxing the constraint.

Table 14 Selected XAI-Gate operating points on the UNSW-NB15 calibrated anchor. The expanded-budget row is reported separately and is not a fixed-budget comparison.

Regime	Variant	Policy Configuration / Parameterization	HR coverage	Debt	Exposure	Violation	Drop
Adversarial demand	High local budget	Expanded queue capacity ($Q_L = 50$, $E_{\max} = 1.0$)	0.699	3382.6	1.000	0.000	0.000
Adversarial demand	Strict suspicion	Elevated suspicion sensitivity (γ_{susp}, $E_{\max} = 1.0$)	0.689	4090.5	1.000	0.000	0.000
Adversarial demand	Strict exposure	Penalized explanation exposure (w_E, $E_{\max} = 1.0$)	0.686	4169.4	1.000	0.000	0.000
Adversarial demand	High exposure budget	Expanded exposure budget ($E_{\max} = 1.5$; non-comparative reference)	0.790	2794.0	0.990	0.000	0.000
Transient overload	Balanced coverage	Balanced coverage–debt control ($E_{\max} = 1.0$)	0.766	3741.4	1.000	0.000	0.248
Non-Markovian	Strict suspicion	Elevated suspicion sensitivity (γ_{susp}, $E_{\max} = 1.0$)	0.772	593.1	0.932	0.000	0.137
RTT uncertainty	Balanced coverage	Balanced coverage–debt control ($E_{\max} = 1.0$)	0.873	5.9	0.888	0.000	0.000

5.4 Demand-Inflation and Frontier Behavior

The demand-inflation sweep increases the attack multiplier from 1.0 to 2.0. Figure 2 shows the resulting coverage, debt, exposure, and frontier behavior. The sweep confirms that XAI-Gate's exposure control is not free. At multiplier 1.0, XAI-Gate keeps exposure near budget, produces no exposure-budget violations, and holds mean debt at 1535.8. At multiplier 2.0, the service boundary is explicit: XAI-Gate's high-risk coverage falls to 0.223 and mean explanation debt rises to 47294.6. Its exposure use remains near 1.000, but exposure-budget violations rise to 0.702 and packet drop reaches 0.101. Threshold explanation behaves in the opposite direction: coverage increases from 0.884 to 0.916 as more high-score alerts are explained, but exposure use rises from 1.388 to 4.277, exposure-budget violations rise from 0.279 to 0.767, and packet drop reaches 0.114. Therefore, the sweep identifies a service boundary rather than a simple win. Under severe explanation-demand inflation, finite explanation capacity can no longer satisfy coverage, debt, exposure, and packet-service objectives simultaneously; XAI-Gate limits exposure magnitude relative to threshold-style explanation, but it does not preserve all exposure constraints once demand exceeds the configured service envelope. This regime effectively creates a *denial-of-explanation* (DoE) state: an attacker who can sustain demand inflation at multiplier 2.0 can drive high-risk explanation coverage below 0.25 while also forcing packet drops above 0.10. The DoE vulnerability is an unavoidable consequence of finite capacity rather than a policy-design weakness, and

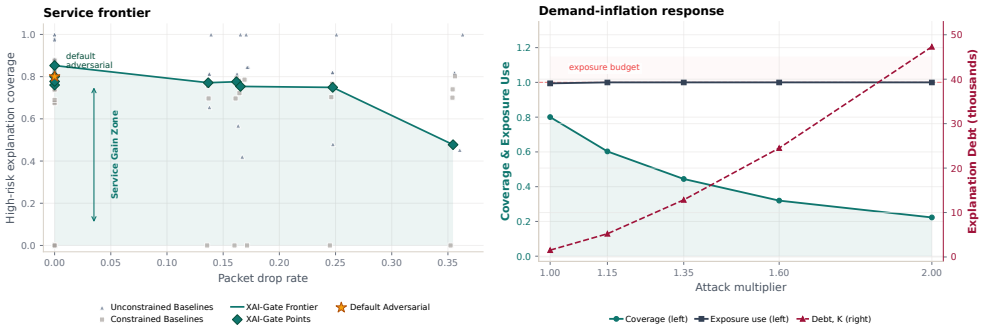


Fig. 2 XAI-SurfaceBench frontier and demand-inflation behavior. Left: main stress-test frontier between packet drop and high-risk explanation coverage. Right: high-risk coverage, normalized explanation debt, and exposure use as adversarial explanation demand increases.

the proper mitigation is to expand the service envelope (e.g., through admission control, rate limiting at the alert source, or increased provisioning) rather than to expect the gate to create capacity.

The frontier experiment sweeps both local explanation budget and exposure budget. This stage is necessary because a single default budget can make a policy appear either overly conservative or overly aggressive. The frontier results show that XAI-Gate is best interpreted as an operating-envelope controller: its advantage is strongest when the defender values exposure feasibility and debt control, while threshold-style policies remain attractive when the only objective is immediate high-risk coverage. This supports the paper's service-management framing because the relevant object is a multi-objective frontier rather than a scalar IDS accuracy or explanation-coverage score.

5.5 Ablation and Robustness Results

Ablation results identify which mechanisms drive the default XAI-Gate behavior. Under adversarial explanation-demand inflation, default XAI-Gate achieves high-risk coverage of 0.801, mean debt of 1528.9, and exposure use of 0.994. Removing the exposure-budget mechanism raises coverage to 0.831 and nearly eliminates debt, but exposure use increases to 2.563. This ablation confirms that the exposure budget is the reason XAI-Gate does not simply chase immediate coverage. Removing debt awareness reduces coverage to 0.637 and raises mean debt to 8184.8, while replacing the policy with threshold-only scoring raises coverage to 0.829 but increases debt to 6358.1 and exposure use to 1.294. Removing fidelity control is especially damaging: high-risk coverage drops to 0.288, debt rises to 10254.5, and the policy loses the redaction channel. These results (shown in Fig. 3, left) support the mechanism-level claim that the debt, exposure, suspicion, and fidelity states are not decorative; they materially change the service behavior of the gate.

Robustness sweeps further delimit the claim. Across RTT uncertainty, explanation-cost scaling, service-capacity variation, detection-matrix perturbation, and burstiness sweeps, XAI-Gate's high-risk coverage ranges from 0.737 to 0.860, exposure use ranges from 0.483 to 0.994, and exposure-budget violation remains zero. However, packet

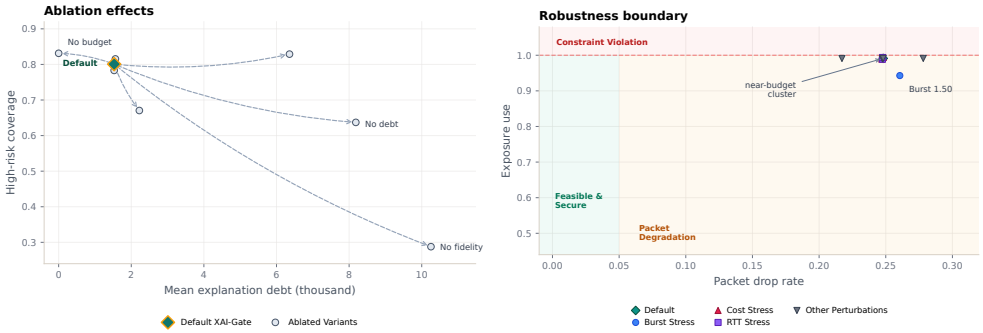


Fig. 3 Ablation and robustness diagnostics. Left: adversarial-regime ablations show how removing debt, exposure, suspicion, offload, fidelity, or RTT robustness changes the debt–coverage operating point. Right: robustness sweeps are summarized as a packet-drop versus exposure-use boundary for XAI-Gate under RTT, cost, service-capacity, detection-matrix, and burstiness perturbations.

drop ranges from 0.000 to 0.278 when service capacity and burstiness are stressed. The correct interpretation (summarized in Fig. 3, right) is therefore conditional: XAI-Gate keeps the explanation exposure constraint stable across the tested robustness envelope, but it does not eliminate packet-service degradation when the traffic process itself exceeds the service envelope. This is consistent with the queueing model in Section 3; a gate can manage explanation work, but it cannot create packet-service capacity.

The profile-sensitivity analysis also perturbs the abstract action profiles used by XAI-Gate. Across full-explanation exposure, offload exposure, offload delay, coarse coverage, redaction exposure, audit coverage, explanation-cost scale, and debt-scale variants, adversarial-regime high-risk coverage ranges from 0.626 to 0.868, mean debt from 402.2 to 3124.1, and exposure use from 0.884 to 1.000. Some profile perturbations produce exposure-budget violations up to 0.127 in the adversarial regime and 0.149 in the transient-overload regime. These are not hidden failures; they show that action-profile semantics materially affect the operating point. These sensitivity results make action-profile assumptions part of XAI-Gate's reported operating point.

The action-profile provenance stage adds a local proxy timing check for the compute tiers. Table 15 reports deterministic proxy measurements over 500 synthetic alert vectors in a commodity laptop-class CPU environment to establish a baseline relative ordering. Full local explanation is normalized to 1.0, offload measures only local request packaging and obtains a normalized median cost of approximately 0.15, and coarse, audit, and redaction proxies are below the full-explanation runtime. The table separates three quantities that should not be conflated: local proxy runtime, declared service-profile weights, and exposure/delay semantics. The default compute column is a normalized service-profile cost used by the benchmark policy, not a direct hardware-latency measurement. To bridge this commodity testbed with highly constrained Edge/IoT gateways, we model an Edge Hardware Impedance Factor, $\alpha \geq 1$, which scales the local compute cost to $C_{\text{edge}}(a) = \alpha C_{\text{proxy}}(a)$, representing the performance ratio between low-power edge nodes (e.g., ARM Cortex-A72 cores) and the high-performance proxy CPU. Rather than claiming absolute LIME or KernelSHAP latency on target hardware, XAI-Gate treats compute values as logical service weights

and stress-tests their scaling through the sensitivity sweeps. As evaluated in those sweeps, when the hardware constraint is stressed (simulating high α), the online policy automatically shifts the explanation load towards low-overhead action profiles (such as Coarse, Redact, Offload, or Delay) to preserve packet-forwarding QoS. The measurement and impedance sweeps support a narrower claim: every action-profile value is bounded and validated, ensuring system stability even under severe hardware mismatch.

Table 15 Operational action-profile parameterization and system implementation semantics for XAI-SurfaceBench. The logical weight $W_c(a)$ defines the policy resource weight, $C_p(a)$ represents the normalized local compute cost, and $E_p(a)$ and $D_p(a)$ denote the empirical information exposure and queue delay profiles.

Operational Policy Action ($a \in \mathcal{A}$)	Logical Weight (W_c)	Compute Cost (C_p)	Leakage Exposure (E_p)	Queue Delay (D_p)	System Action Semantics & Security Implementation
a_{none} (Suppress/Forward)	0.00	0.000	0.00	0.00	Transparent gateway packet forwarding; bypass explainer under queue pressure
a_{coarse} (Coarse Explanation)	0.28	0.022	0.10	0.25	Low-overhead statistical approximation (top-k Gini feature attributions)
a_{full} (Full Local)	1.00	1.000	0.85	0.85	High-fidelity local attribution (full-sample perturbation-based LIME/SHAP)
a_{offload} (Fog/Cloud Offload)	0.16	0.148	1.00	1.30	Remote server delegation; packaging and secure telemetry transmission
a_{audit} (Audit Logging)	0.62	0.038	0.34	0.70	Minimal cryptographic logging of state parameters and metadata hashes
a_{redact} (Redacted Release)	0.46	0.043	0.08	0.55	Dynamic feature-masking and active sanitization of high-exposure attributes
a_{delay} (Controlled Delay)	0.00	0.000	0.00	1.00	Buffered queue holding; temporarily deferring work to smooth traffic spikes

6 Discussion and Validity Bounds

While the experimental results in Section V demonstrate the effectiveness of XAI-Gate in controlling explanation exposure and mitigating queueing debt, it is crucial to establish the precise boundary of these claims. Rather than presenting XAI-Gate as a new intrusion detection system (IDS) or a novel feature-attribution explainer, our contribution is strictly positioned as a service-management and policy-governance layer. Consequently, the empirical findings must be interpreted under the following validity bounds:

6.1 Methodological and Architectural Scope

The proposed framework is detector-agnostic and relies on score-stream calibration rather than model retraining. While this architecture ensures high flexibility and low deployment overhead, it implies that the ultimate detection accuracy remains bounded by the underlying IDS. XAI-Gate optimizes the scheduling and fidelity of explanations, but it cannot improve the baseline receiver operating characteristics (ROC) of the classifier.

6.2 Hardware and Latency Assumptions

Our evaluation of the action-profile compute costs (Table 15) utilizes normalized relative weights measured on a commodity testbed, scaled mathematically via the Edge Hardware Impedance Factor α. Although the sensitivity sweeps in Section V-C confirm that the online policy gracefully degrades to low-overhead profiles when α is large, absolute physical latencies and resource usage (VRAM, CPU cycles) are highly dependent on the target edge architecture (e.g., specific ARM Cortex or RISC-V gateways). Standard perturbation explainers on low-power devices can introduce

significant delays; thus, direct hardware-in-the-loop validation is a prerequisite for physical deployment.

6.3 Human Factors and Trust Dynamics

A core motivation for managing the explanation surface is to prevent analyst cognitive overload and control the cumulative exposure of sensitive features. However, while our queuing debt model mathematically captures analyst service rates, the actual impact of XAI-Gate on human decision quality, triage speed, and operational trust has not been validated through user studies. The current evidence is bounded by simulated queueing processes and requires subsequent human-subject experiments to claim cognitive or operational improvements.

6.4 Generalization to Live Traffic Profiles

The calibration anchors executed on UNSW-NB15 and TON_IoT provide realistic score distributions, and our de-duplication efforts remove the duplicate-overlap bias. Nevertheless, these static traces do not capture the non-stationary, multi-scale burstiness of live, high-speed enterprise network gateways. The stability of the exposure-preserving policy under extreme, sustained traffic surges is bounded by the capacity limits derived in Section IV-B; once explanation demand consistently exceeds the service envelope, packet-service degradation remains an unavoidable physical constraint.

6.5 Reproducibility and Limitations

The suite evidence is reproducible from the XAI-SurfaceBench execution records, which retain fixed configurations, deterministic seed construction, machine-readable run summaries, generated tables, and figures. These materials support claims about the simulated service frontier among coverage, debt, exposure, and QoS under the specified regimes. They do not by themselves establish deployment readiness.

Four limitations remain explicit. First, the public-data calibration anchors now include a legacy no-training KDDCup99 proxy, a modern UNSW-NB15 public test-split calibration stream, a de-duplicated UNSW rerun that removes feature-duplicate rows and train-test feature overlap, and a TON_IoT held-out calibration stream. This is stronger than a purely synthetic score prior, but it is still not a broad deployment validation across CICIoT/CICIDS traces, enterprise traces, or live gateways. Second, only local proxy timing was performed, so absolute edge/gateway hardware latency, memory, and accelerator-cost claims remain outside the evidence boundary. Third, an SME/operator validation protocol has been prepared but not executed, so the paper cannot claim improved analyst trust, triage speed, or human decision quality. Fourth, the closest-work positioning review is not a full systematic literature review; exhaustive novelty language should therefore be avoided. Fifth, under extreme adversarial explanation-demand inflation (attack multiplier ≥ 2.0), XAI-Gate enters a *denial-of-explanation* (DoE) regime where high-risk coverage falls substantially and both packet drops and exposure-budget violations appear (Section V-D). This is a finite-capacity service boundary rather than a policy-design failure, but it means that a sufficiently

resourced adversary can degrade explanation availability. The next validation sequence is therefore target gateway hardware timing, live or modern trace replay beyond score-stream calibration, executed SME/operator evaluation, and DoE mitigation strategies rather than another detector-accuracy comparison.

XAI-Gate is also not preferable in every operating environment. If the only objective is maximum immediate high-risk explanation coverage, if explanation exposure is not security-relevant, if explanation capacity is abundant relative to alert demand, or if detector scores are too poorly calibrated for risk ranking, simpler threshold or always-explain policies may be more appropriate. The policy is intended for constrained settings where explanation availability, debt, and exposure must be managed jointly.

7 Conclusion

This paper reframes XAI-enabled network defense around the management of the explanation surface. Instead of treating explanations as free artifacts appended to detector outputs, XAI-Gate treats explanation generation as finite-capacity, delay-sensitive, information-bearing, and adversarially triggerable service behavior. The proposed state model, online policy, hard exposure-cap guard, and service-boundary propositions define when explanations should be admitted, degraded, delayed, offloaded, redacted, audited, or suppressed. The executed XAI-SurfaceBench evaluation shows that this framing changes the empirical question: the strongest result is not a blanket improvement in every metric, but a measurable shift in the service frontier. In the default adversarial explanation-demand regime, XAI-Gate reduces explanation debt and removes exposure-budget violations relative to high-coverage threshold policies, while accepting lower high-risk explanation coverage. The KDD-Cup99, UNSW-NB15, de-duplicated UNSW, and TON_IoT anchors preserve the same interpretation under public score streams: high-coverage alternatives can win on immediate coverage, but often at higher debt or exposure violations. Under more severe demand inflation, the same exposure-preserving behavior produces a negative result: coverage falls and debt grows once demand exceeds the configured service envelope. These findings strengthen the claim boundary. XAI-Gate is not a new IDS, not a new explainer, not a broad multi-dataset deployment validation, and not a substitute for hardware or operator validation. It is a service-management layer and reproducible stress-test protocol for controlling explanation availability, explanation debt, and exposure in XAI-enabled network defense.

Appendix A Expanded Prior-Art Positioning Boundary

Table A1 summarizes the closest-work positioning used to discipline the novelty claim. The comparison covers foundation XAI, XAI-IDS and cybersecurity surveys, traffic and service-management work, adversarial-XAI studies, selective explanations, and

operator-facing SOC-XAI work. It is not presented as a full PRISMA systematic literature review; its role is to preserve a narrow claim boundary for explanation-surface governance.

Table A1 Expanded prior-art positioning matrix for the explanation-surface management claim. We compare core capabilities across adjacent domains: • denotes central capability, ○ denotes partial or indirect capability, and − denotes that the capability is not modeled or addressed.

Literature Stream	Representative Works	Service Mgmt.	Finite Capacity	Selective Exp.	Queueing Debt	Exposure Control	Adv. Demand	SOC Operator
Foundation XAI	LIME [1]; SHAP [2]	−	−	−	−	−	−	−
Selective Explanation	Paes et al. 2024 [29]	○	○	•	−	−	−	−
XAI-IDS & Cyber Surveys	Neupane [4]; Moustafa [5]; Rjoub [20]; Nascita [22]	○	○	○	−	○	○	○
Traffic/Service Mgmt.	Lai [13]; Kim [25]; Nascita traffic studies [27]	•	•	−	−	−	○	−
Adversarial XAI	Kuppa [14]; Slack [15]; Baniecki [16]; Pawlicki [21]	−	○	○	−	•	○	○
Operator-Facing XAI	Eriksson [30]; Nyre-Yu [31]	−	−	○	−	−	−	•
XAI-Gate	**This Work**	•	•	•	•	•	•	○

Appendix B KDDCup99 and TON_IoT Calibrated Anchor Benchmarks

This appendix provides the tabular results for the auxiliary KDDCup99 and TON_IoT calibrated anchors referenced in Section V. Table B2 reports the KDDCup99 no-training calibrated-anchor check, confirming that under public score-stream priors, XAI-Gate obtains high-risk coverage of 0.790 and mean debt of 872.0 with zero exposure-budget violations. Table B3 reports the TON_IoT calibrated-anchor results across four tested regimes under a fixed exposure budget. In both anchors, XAI-Gate maintains exposure feasibility while managing the coverage-debt tradeoff under resource constraints.

Table B2 KDDCup99 no-training calibrated-anchor check under adversarial explanation-demand inflation.

Policy	HR coverage	Debt	Exposure use	Violation
Budget-only	0.177	10964.0	0.035	0.000
Threshold	0.177	10992.7	0.035	0.000
XAI-Gate	0.790	872.0	0.947	0.000

Table B3 TON_IoT calibrated-anchor results under fixed exposure budget. The stream is generated from an overlap-controlled, held-out test split and is used only as calibration metadata for explanation-surface stress testing. HR denotes high-risk coverage, exp. denotes exposure use, and viol. denotes exposure-budget violation.

Regime	XAI HR	XAI debt	XAI exp.	XAI viol.	Budget HR	Budget debt	Budget exp.	Thresh. HR	Thresh. exp.	Thresh. viol.
Adversarial demand	0.654	5020.2	1.000	0.000	0.767	6792.3	0.931	0.994	2.666	0.625
Transient overload	0.689	4947.5	1.000	0.000	0.786	5201.8	0.808	0.986	1.970	0.489
Non-Markovian	0.754	1274.4	0.979	0.000	0.788	3639.3	0.736	0.984	1.786	0.438
RTT uncertainty	0.863	98.0	0.932	0.000	0.793	867.6	0.581	0.984	1.385	0.277

Appendix C Paired Seed-Level Statistical Sign-Tests

This appendix reports the formal seed-level statistical significance analysis used to validate the descriptive results. Table C4 lists the XAI-Gate minus comparator mean differences, bootstrap 95% confidence intervals over paired seed differences, and two-sided sign-test p-values across 20 independent seeds. Positive differences favor XAI-Gate for high-risk coverage, while negative differences favor XAI-Gate for debt and exposure-violation metrics.

Table C4 Compact paired statistical follow-up for adversarial explanation-demand comparisons. Differences are XAI-Gate minus the comparator. Positive differences favor XAI-Gate for high-risk coverage; negative differences favor XAI-Gate for debt and exposure-violation metrics.

Anchor	Comparison	Metric	Seeds	Mean diff.	95% CI	Fav/tie/unfav	Sign p
Synthetic tuned	XAI-Gate–Threshold	HR coverage	20	0.2512	[0.2504, 0.2519]	20/0/0	1.907e-06
Synthetic tuned	XAI-Gate–Threshold	Mean debt	20	-8859	[-8898, -8815]	20/0/0	1.907e-06
Synthetic tuned	XAI-Gate–Threshold	Exposure viol.	20	0	[0, 0]	0/20/0	NA
Synthetic tuned	XAI-Gate–Budget	HR coverage	20	0.02649	[0.02599, 0.02701]	20/0/0	1.907e-06
Synthetic tuned	XAI-Gate–Budget	Mean debt	20	-3758	[-3790, -3725]	20/0/0	1.907e-06
UNSW default	XAI-Gate–Budget	HR coverage	20	-0.06705	[-0.06835, -0.06571]	0/0/20	1.907e-06
UNSW default	XAI-Gate–Budget	Mean debt	20	-3589	[-3646, -3537]	20/0/0	1.907e-06
UNSW dedup	XAI-Gate–Budget	HR coverage	20	0.01376	[0.01238, 0.01521]	20/0/0	1.907e-06
UNSW dedup	XAI-Gate–Budget	Mean debt	20	-5354	[-5398, -5311]	20/0/0	1.907e-06
UNSW dedup	XAI-Gate–Threshold	Exposure viol.	20	-0.4068	[-0.4086, -0.405]	20/0/0	1.907e-06
TON_IoT	XAI-Gate–Budget	HR coverage	20	-0.1127	[-0.1141, -0.1112]	0/0/20	1.907e-06
TON_IoT	XAI-Gate–Budget	Mean debt	20	-1772	[-1841, -1703]	20/0/0	1.907e-06
TON_IoT	XAI-Gate–Threshold	Exposure viol.	20	-0.625	[-0.6261, -0.6239]	20/0/0	1.907e-06

Statements and Declarations

Acknowledgments

This work has been sponsored by the scientific research from Posts and Telecommunications Institute of Technology, Vietnam.

Funding

This work has been sponsored by the scientific research from Posts and Telecommunications Institute of Technology, Vietnam.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception, design, methodology development, performance evaluation, and manuscript writing. All authors read and approved the final manuscript.

Data and Code Availability

The code and calibrated configurations generated during this study are available in our open-source repository at https://github.com/tranducle/XAI-GATE. The UNSW-NB15 and TON_IoT datasets used as calibration anchors are publicly available from their original repositories.

References

[1] Ribeiro, M.T., Singh, S., Guestrin, C.: "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. KDD '16, pp. 1135–1144. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778

[2] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc., USA (2017). https://papers.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[3] Abou El Houda, Z., Brik, B., Khoukhi, L.: "why should i trust your IDS?" an explainable deep learning framework for intrusion detection systems in internet of things networks. IEEE Open Journal of the Communications Society **3**, 1164–1176 (2022) https://doi.org/10.1109/ojcoms.2022.3188750

[4] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., Seale, M.: Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities. IEEE Access **10**, 112392–112415 (2022) https://doi.org/10.1109/access.2022.3216617

[5] Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A.Y., Tari, Z.: Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. IEEE Communications Surveys & Tutorials **25**(3), 1775–1807 (2023) https://doi.org/10.1109/comst.2023.3280465

[6] Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., Zomaya, A.Y.: An explainable deep learning-enabled intrusion detection framework in IoT networks. Information Sciences **639**, 119000 (2023) https://doi.org/10.1016/j.ins.2023.119000

[7] Sharma, B., Sharma, L., Lal, C., Roy, S.: Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. Expert Systems with Applications **238**, 121751 (2024) https://doi.org/10.1016/j.eswa.2023.121751

[8] Arreche, O., Guntur, T., Abdallah, M.: XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. Applied Sciences **14**(10), 4170 (2024) https://doi.org/10.3390/app14104170

[9] Komarchesqui, M., Silva Ruffo, V.G., Carvalho, L.F., *et al.*: eXplainable Artificial Intelligence for transparent optimization of deep learning-based intrusion detection systems. Journal of Network and Systems Management **34**(1), 1–22 (2026) https://doi.org/10.1007/s10922-026-10084-z

[10] Wan, X., Xue, G., Zhong, Y., *et al.*: Separating prediction and explanation: An approach based on explainable artificial intelligence for analyzing network intrusion. Journal of Network and Systems Management **32**(4), 1–24 (2024) https://doi.org/10.1007/s10922-024-09891-z

[11] Aljuhani, A., Alamri, A., Kumar, P., Jolfaei, A.: An intelligent and explainable SaaS-based intrusion detection system for resource-constrained IoMT. IEEE

Internet of Things Journal **11**(15), 25454–25463 (2024) https://doi.org/10.1109/jiot.2023.3327024

[12] Ogunseyi, T.B., Thiyagarajan, G., He, H., Bist, V., Du, Z.: Performance analysis of explainable deep learning-based intrusion detection systems for iot networks: A systematic review. Sensors **26**(2), 363 (2026) https://doi.org/10.3390/s26020363

[13] Lai, Y.-C., Sudyana, D., Lin, Y.-D., Verkerken, M., D'hooge, L., Wauters, T., Volckaert, B., De Turck, F.: Task assignment and capacity allocation for ML-based intrusion detection as a service in a multi-tier architecture. IEEE Transactions on Network and Service Management **20**(1), 672–683 (2023) https://doi.org/10.1109/tnsm.2022.3203427

[14] Kuppa, A., Le-Khac, N.-A.: Adversarial XAI methods in cybersecurity. IEEE Transactions on Information Forensics and Security **16**, 4924–4938 (2021) https://doi.org/10.1109/tifs.2021.3117075

[15] Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES '20, pp. 180–186. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3375627.3375830

[16] Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion **107**, 102303 (2024) https://doi.org/10.1016/j.inffus.2024.102303

[17] Charmet, F., Tanuwidjaja, H.C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., Zhang, Z.: Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications **77**(11-12), 789–812 (2022) https://doi.org/10.1007/s12243-022-00926-7

[18] Zhang, Z., Hamadi, H.A., Damiani, E., Yeun, C.Y., Taher, F.: Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access **10**, 93104–93139 (2022) https://doi.org/10.1109/access.2022.3204051

[19] Capuano, N., Fenza, G., Loia, V., Stanzione, C.: Explainable artificial intelligence in cybersecurity: A survey. IEEE Access **10**, 93575–93600 (2022) https://doi.org/10.1109/access.2022.3204171

[20] Rjoub, G., Bentahar, J., Abdel Wahab, O., Mizouni, R., Song, A., Cohen, R., Otrok, H., Mourad, A.: A survey on explainable artificial intelligence for cybersecurity. IEEE Transactions on Network and Service Management **20**(4), 5115–5140 (2023) https://doi.org/10.1109/tnsm.2023.3282740

[21] Pawlicki, M., Pawlicka, A., Kozik, R., Choraś, M.: The survey on the dual nature of xAI challenges in intrusion detection and their potential for AI innovation. Artificial Intelligence Review **57**(12) (2024) https://doi.org/10.1007/

s10462-024-10972-3

[22] Nascita, A., Aceto, G., Ciuonzo, D., Montieri, A., Persico, V., Pescapé, A.: A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection. IEEE Communications Surveys & Tutorials **27**(5), 3165–3198 (2025) https://doi.org/10.1109/comst.2024.3504955

[23] Gaspar, D., Silva, P., Silva, C.: Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. IEEE Access **12**, 30164–30175 (2024) https://doi.org/10.1109/access.2024.3368377

[24] Liu, H., Zhong, C., Alnusair, A., *et al.*: FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques. Journal of Network and Systems Management **29**(4), 1–25 (2021) https://doi.org/10.1007/s10922-021-09606-8

[25] Kim, S., Yoon, S., Cho, J.-H., Kim, D.S., Moore, T.J., Free-Nelson, F., Lim, H.: DIVERGENCE: Deep reinforcement learning-based adaptive traffic inspection and moving target defense countermeasure framework. IEEE Transactions on Network and Service Management **19**(4), 4834–4846 (2022) https://doi.org/10.1109/tnsm.2021.3139928

[26] Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., Pescapé, A.: XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures. IEEE Transactions on Network and Service Management **18**(4), 4225–4246 (2021) https://doi.org/10.1109/tnsm.2021.3098157

[27] Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., Pescapé, A.: Improving performance, reliability, and feasibility in multimodal multitask traffic classification with XAI. IEEE Transactions on Network and Service Management **20**(2), 1267–1289 (2023) https://doi.org/10.1109/tnsm.2023.3246794

[28] Verkerken, M., D'hooge, L., Wauters, T., *et al.*: Towards model generalization for intrusion detection: Unsupervised machine learning techniques. Journal of Network and Systems Management **29**(3), 1–26 (2021) https://doi.org/10.1007/s10922-021-09615-7

[29] Calmon, F., Paes, L., Wei, D.: Selective explanations. In: Advances in Neural Information Processing Systems 37. NeurIPS 2024, pp. 55562–55590. Neural Information Processing Systems Foundation, Inc. (NeurIPS), USA (2024). https://doi.org/10.52202/079017-1765

[30] Eriksson, H.S., Grov, G.: 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), pp. 2595–2600. IEEE, Piscataway, NJ, USA (2022). https://doi.org/10.1109/bigdata55660.2022.10020248

[31] Nyre-Yu, M., Morris, E., Smith, M., Moss, B., Smutz, C.: Explainable AI in cybersecurity operations: Lessons learned from xAI tool deployment. In: Proceedings 2022 Symposium on Usable Security. USEC 2022. Internet Society, Reston, VA, USA (2022). https://doi.org/10.14722/usec.2022.23014

[32] Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research1. MIS Quarterly **28**(1), 75–106 (2004) https://doi.org/10.2307/25148625

[33] Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, Hoboken, NJ, USA (1994). https://doi.org/10.1002/9780470316887

[34] Altman, E.: Constrained Markov Decision Processes: Stochastic Modeling. Routledge, London, UK (2021). https://doi.org/10.1201/9781315140223

[35] Neely, M.J.: Stability and probability 1 convergence for queueing networks via lyapunov optimization. Journal of Applied Mathematics **2012**(1) (2012) https://doi.org/10.1155/2012/831909

[36] SIGKDD: KDD Cup 1999 Data. Accessed: 2026-05-22 (1999). https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[37] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011)

[38] Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 Military Communications and Information Systems Conference (MilCIS), pp. 1–6. IEEE, Piscataway, NJ, USA (2015). https://doi.org/10.1109/milcis.2015.7348942

[39] Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., Anwar, A.: TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. IEEE Access **8**, 165130–165150 (2020) https://doi.org/10.1109/access.2020.3022862

[40] Neto, E.C.P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., Ghorbani, A.A.: CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. Sensors **23**(13), 5941 (2023) https://doi.org/10.3390/s23135941

[41] Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy, pp. 108–116. SCITEPRESS - Science and Technology Publications, Setubal, Portugal

(2018). https://doi.org/10.5220/0006639801080116

[42] Canadian Institute for Cybersecurity: CSE-CIC-IDS2018 on AWS. Accessed: 2026-05-22 (2018). https://www.unb.ca/cic/datasets/ids-2018.html