JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 4, No. 1 (2026), pp. 10–26
DOI: 10.63646/jaiaa.2026.040102

Explainable AI Analytics for Intrusion Detection in Healthcare IoT: From Federated Model Updates to Risk-Aware Decision Support

Tomasz Wiśniewski[1]; Helena Marković[2]; Kristoffer Johansson[3] *

[1] Faculty of Computer Science, Białystok University of Technology, Białystok, Poland
[2] Faculty of Electrical Engineering, Computer Science and Information Technology, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia
[3] Department of Information Systems and Technology, Mid Sweden University, Sundsvall, Sweden
** Corresponding author: kristoffer.johansson@miun.se*

ARTICLE INFO
Received
October 18, 2025
Revised
December 12, 2025
Accepted
February 25, 2026
Available Online
March 30, 2026
DOI
10.63646/jaiaa.2026.040102

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract

Healthcare Internet-of-Things (H-IoT) deployments — wearable biosensors, bedside monitors, and connected diagnostic devices — generate dense streams of telemetry that increasingly underpin clinical and operational decisions. Their attack surface is correspondingly large, and the dominant defensive primitive, the deep-learning intrusion detection system (IDS), is opaque, brittle under non-IID data, and difficult to govern under medical-device regulation. This article develops a Trust-by-Design Analytics framework for H-IoT intrusion detection that couples three layers: a federated Bi-LSTM detector trained across hospital and at-home clients without sharing raw traffic; a Shapley-based explainable-AI (XAI) layer that attributes each detection to a small set of human-auditable features; and a risk-aware decision layer that translates calibrated detector posteriors and explanation faithfulness scores into a triage action — accept, abstain, or refer to a security analyst — under an explicit cost model. We present a controlled numerical study, calibrated to noise levels and prevalence figures from the ToN-IoT and CICIDS2019 corpora, in which the framework achieves an F1 of 0.961 on ToN-IoT and 0.971 on CICIDS2019, recovering most of the centralised-baseline ceiling while preserving privacy, and reduces normalised expected misclassification cost by approximately 38% relative to a fixed-threshold federated baseline at a clinically realistic 5:1 cost ratio between false negatives and false positives. We further show that explanation faithfulness, measured by an insertion-test AUC, is monotone in detection confidence and acts as a useful gate for the abstain-and-refer pathway. The framework is positioned as a deployment template for analytics teams working at the boundary of regulated medical devices and at-scale digital-health operations: it disaggregates the IDS pipeline into separately governable components, surfaces calibration and explanation quality as first-class operational metrics, and ties the detection cost to a transparent, contractible decision policy.

Keywords: Explainable artificial intelligence; Federated learning; Intrusion detection; Healthcare Internet of Things; Bi-LSTM; Risk-aware decision support; SHAP; Calibration

I. INTRODUCTION

The data flowing out of a contemporary hospital is no longer principally electronic-health-record text and laboratory results. It is a continuous telemetry stream from infusion pumps, telemetry monitors, wearable cardiac patches, smart inhalers, and remote-patient-monitoring kits, much of it generated outside the controlled environment of the inpatient ward and routed through consumer networks before reaching the institutional perimeter. The category that has emerged from this transformation, the Healthcare Internet of Things (H-IoT), has become a substantive segment of clinical analytics rather than a peripheral curiosity, with the global installed base of connected medical devices now estimated in the tens of millions and projected to grow at double-digit rates over the present decade (Sun et al., 2019; Yaacoub et al., 2020; Papaioannou et al., 2022).

The same connectivity that makes H-IoT clinically valuable also makes it adversarially exposed. The attack surface includes weak device authentication, opaque firmware update channels, lateral-movement opportunities from compromised consumer wearables into hospital networks, and a steady drumbeat of denial-of-service, injection, and ransomware campaigns specifically targeting medical infrastructure (Lu & Xu, 2019). The defensive primitive of choice has, over the last five years, become the deep-learning intrusion detection system (IDS): a recurrent or convolutional classifier trained to flag anomalous traffic against a labelled corpus of attack and benign flows. Public benchmarks such as ToN-IoT (Moustafa, 2021) and CICIDS2019 (Sharafaldin et al., 2018) have made it possible for both academic and industrial research groups to compare detector architectures with methodological discipline, and the headline detection accuracy figures reported in the

resulting literature now routinely exceed 95% on in-distribution test data (Vinayakumar et al., 2019; Aldweesh et al., 2020; Liu & Lang, 2019).

Yet a growing body of evidence — from clinical machine-learning practitioners (Wiens et al., 2019; Topol, 2019), from regulators developing guidance on software as a medical device (Floridi, 2019; Jobin et al., 2019), and from cybersecurity analysts deploying these models on production health networks (Capuano et al., 2022; Keshk et al., 2023) — argues that benchmark accuracy on in-distribution data is a poor proxy for the value of an IDS in deployment. Three deployment-time problems consistently surface. The first is data-distribution heterogeneity: an H-IoT cohort spans hospital telemetry networks, at-home wearable links, and ambulant 5G traffic, and a centrally trained detector predictably degrades when shifted across these regimes. The second is opacity: a security analyst who is shown only a binary alert and a confidence score has neither a basis to action the alert nor a defensible record for the audit trail mandated by emerging medical-device cybersecurity regulation. The third is cost asymmetry: the operational consequences of a missed ransomware infiltration of an oncology infusion pump and of a false alarm on a benign firmware update event differ by orders of magnitude, but a uniform detection threshold makes no allowance for that asymmetry.

The contribution of this article is a Trust-by-Design Analytics framework that addresses all three problems jointly. The framework couples three layers: a federated Bi-LSTM detector that trains across hospital and at-home clients without raw-data sharing (Yang et al., 2019; Kairouz et al., 2021); a Shapley-based explainable-AI (XAI) layer that attributes each detection to a compact set of human-auditable features (Lundberg & Lee, 2017; Sundararajan et al., 2017); and a risk-aware decision layer that translates calibrated detector posteriors and explanation faithfulness scores into a triage action — accept, abstain, or refer — under an explicit cost model (Chow, 1970; Elkan, 2001; Geifman & El-Yaniv, 2017). The premise of the framework is that each layer absorbs a category of deployment risk that the standard end-to-end IDS would otherwise either accept silently or paper over with an inflated headline accuracy figure.

Four specific contributions follow. First, we recast a representative body of work on federated learning for H-IoT intrusion detection (Mothukuri et al., 2022; Campos et al., 2022) as a deployable analytics template by making the cost structure of the alerting decision explicit and quantitative. Second, we design and evaluate a federation-aware XAI module that aggregates client-local Shapley attributions into a globally-meaningful explanation surface, and we show, via insertion-test faithfulness benchmarks, that the resulting explanations are both faithful to the underlying detector and computationally tractable for edge deployment. Third, we provide a controlled numerical study, calibrated to ToN-IoT and CICIDS2019 noise and prevalence levels, that decomposes the detection-quality, calibration, and expected-cost gains attributable to each layer of the framework. Fourth, we connect the framework to recent management-analytics literature on the governance of artificial intelligence in regulated organisations (Lu, 2019; Lu, 2021; Zhang & Lu, 2021; Lu et al., 2024), positioning H-IoT intrusion detection as a tractable case study for the broader question of how organisations can deploy and govern generative and discriminative AI in domains where opacity and cost-asymmetry are simultaneously binding constraints.

Two boundary conditions deserve up-front acknowledgement. The framework is descriptive of a deployment template rather than prescriptive of a specific clinical alerting policy; the cost ratios used in Section V are illustrative and would in practice be elicited from the security operations centre and the clinical risk function of a particular institution. And the framework concerns the analytics pipeline downstream of the network sensor, not the sensor itself; hardware-rooted security primitives such as device-identity certificates and trusted execution environments are complementary to, but distinct from, the analytics question we address (Xu et al., 2021).

The remainder of the article is organised as follows. Section II reviews the relevant literature in five strands: H-IoT security, intrusion detection with deep learning, federated learning for IDS, explainable AI for cybersecurity, and the management-analytics view of trustworthy AI in regulated domains. Section III develops the Trust-by-Design Analytics framework as a three-layer architecture and formulates the H-IoT intrusion detection problem as a cost-sensitive, distribution-shifted Bayes-risk minimisation. Section IV specifies the methodology, including the federated Bi-LSTM detector, the Shapley aggregation procedure, and the risk-aware decision rule. Section V reports the numerical experiments, including a sensitivity decomposition across noise, cost, and coverage parameters. Section VI discusses managerial and regulatory implications. Section VII concludes.

II. RELATED WORK

We survey five strands of literature whose intersection defines the design space for trust-by-design intrusion detection

 https://doi.org/10.63646/jaiaa.2026.040102

in H-IoT. Each strand identifies a distinct deployment failure mode, and the framework presented in Section III is, in effect, a coordinated response to all five.

A. Healthcare IoT Security

Reviews of H-IoT security agree on a small number of structural facts. The attack surface is heterogeneous, spanning device firmware, wireless links, cloud-side data lakes, and the human operator interface; the most consequential threat categories are denial-of-service, lateral-movement injection, ransomware, and identity-spoofing in device-to-cloud authentication; and the regulatory environment is tightening, with software-as-a-medical-device cybersecurity guidance from the U.S. Food and Drug Administration and the European Medicines Agency converging on expectations for auditable threat models and explainable detection (Sun et al., 2019; Yaacoub et al., 2020; Papaioannou et al., 2022). The IoT-cybersecurity research synthesis of Lu and Xu (2019) provides the broader lens: the deployment of connected sensing into safety-critical domains amplifies the consequences of every classical IoT vulnerability and forces a transition from peripheral defence to deeply-integrated, auditable analytics. The same authors' subsequent work on embedding blockchain into IoT for security (Xu et al., 2021) anticipates the need for tamper-evident audit trails, a concern we return to in the federated-learning architecture of Section IV.

B. Deep-learning Intrusion Detection

Deep-learning intrusion detection systems have moved from research prototype to production tooling over the last five years. Surveys of the field (Liu & Lang, 2019; Aldweesh et al., 2020) chart a steady progression from shallow ensemble methods to recurrent and convolutional deep classifiers, with bidirectional long short-term memory (Bi-LSTM) networks (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997) emerging as the most widely-deployed sequence architecture for network-traffic classification. Vinayakumar et al. (2019) argue, on the basis of empirical comparisons across multiple intrusion-detection corpora, that recurrent architectures consistently outperform feed-forward classifiers on the temporally-correlated traffic that dominates IoT environments. The release of large, attack-rich benchmarks — CICIDS2019 (Sharafaldin et al., 2018) and the ToN-IoT family (Moustafa, 2021) — has made it possible to reproduce headline accuracy figures and to begin the harder work of comparing models on calibration, robustness under adversarial perturbation, and behaviour under distribution shift. The picture from this second wave of evaluation is sobering: in-distribution accuracy of 95%+ does not survive a transition to a slightly different topology, sensor mix, or attack profile, and the same models that achieve near-perfect benchmark accuracy can be fooled by carefully-crafted adversarial inputs (Goodfellow et al., 2015; Madry et al., 2018) at small perturbation budgets.

C. Federated Learning for Intrusion Detection

The privacy and data-governance constraints of healthcare make centralised IDS training, in which raw traffic from heterogeneous clients is concentrated in a single training repository, increasingly untenable. Federated learning (FL) — first formalised by McMahan et al. (2017) and surveyed by Yang et al. (2019), Kairouz et al. (2021), and Li et al. (2020) — offers an architectural answer: local clients train on their own data and share only model updates, which a central aggregator combines without observing the underlying traffic. The FL system pattern has matured to the point where production-scale deployments of federated detection are feasible, with the Federated Averaging (FedAvg) and FedProx (Li et al., 2020b) aggregators serving as the standard baselines. Application of FL specifically to intrusion detection has been documented by Mothukuri et al. (2022) for federated anomaly detection in IoT settings. The Campos et al. (2022) review of FL for IoT IDS distils the principal open challenges: non-IID data across clients, communication-cost amplification under deep models, the absence of robust aggregation under Byzantine clients, and the integration of differential-privacy machinery (Dwork & Roth, 2014; Abadi et al., 2016; Bonawitz et al., 2017; Truex et al., 2019) without an unacceptable utility cost. The healthcare-specific federated-learning literature (Rieke et al., 2020; Mothukuri et al., 2021) confirms the same picture: FL is a credible architectural answer for H-IoT IDS, but only when accompanied by explicit governance over heterogeneity, privacy, and update auditability.

D. Explainable AI for Cybersecurity

Explainable AI has matured into a substantial subfield of machine learning with mature reviews (Arrieta et al., 2020; Doshi-Velez & Kim, 2017). The two methodologies that dominate operational deployment are LIME (Ribeiro et al., 2016) and KernelSHAP (Lundberg & Lee, 2017), the latter rooted in the cooperative-game-theory notion of Shapley value attribution and equipped with a unique axiomatic basis. Gradient-based methods such as Integrated Gradients (Sundararajan

 https://doi.org/10.63646/jaiaa.2026.040102

et al., 2017) and DeepLIFT provide computationally cheaper alternatives, while perturbation-based approaches such as RISE and Occlusion provide model-agnostic baselines. The interpretability-versus-explanation distinction articulated by Rudin (2019) bears emphasising: a post-hoc explanation of a black-box detector is not the same artefact as an inherently interpretable model, and the practitioner must decide which is appropriate for the regulatory context. The cybersecurity-specific XAI literature (Wang et al., 2020; Capuano et al., 2022; Keshk et al., 2023) adapts these methods to network-traffic features and reports that Shapley-based attributions are consistently the most faithful — in the sense of insertion-deletion AUC measured on held-out traffic — at the cost of higher computational latency. The literature has also flagged a substantive vulnerability: post-hoc explanations of opaque classifiers are themselves susceptible to adversarial manipulation (Slack et al., 2020), an observation that complicates any naive deployment of XAI as a regulatory artefact.

E. Management Analytics and Trustworthy AI in Regulated Domains

The fifth strand places the technical pipeline inside its organisational context. Management analytics, as a recently-consolidated interdisciplinary field (Lu, 2017; Lu, 2021; Lu et al., 2024), is concerned with the use of data and models to support decisions inside organisations operating under uncertainty, regulation, and resource constraint. Surveys of artificial-intelligence trends from this perspective (Lu, 2019; Zhang & Lu, 2021) emphasise that the practical deployment of AI is bounded as much by organisational, ethical, and regulatory constraints as by raw technical capability. For digital-health operators those constraints are codified in evolving frameworks for trustworthy and responsible AI (Floridi, 2019; Jobin et al., 2019; Wiens et al., 2019), in security-and-privacy considerations for connected medical devices (Lu & Xu, 2019; Xu et al., 2021), and in the broader debate about interpretability of high-stakes machine learning (Rudin, 2019). Read together, these five strands suggest a concrete design imperative: the analytics pipeline should disaggregate detection, explanation, and decision into separately-governable layers, surface calibration and explanation quality as first-class operational metrics, and tie the detection cost to a transparent, contractible decision policy. The Trust-by-Design Analytics framework that follows is constructed precisely to satisfy that imperative.

III. TRUST-BY-DESIGN ANALYTICS FRAMEWORK

We frame H-IoT intrusion detection as a sequential decision problem under uncertainty and distribution shift. The H-IoT estate produces a stream of network-flow records; for each record, the analytics pipeline must produce one of three actions: accept the record as benign, accept it as suggestive of intrusion and dispatch the corresponding mitigation, or abstain and refer the record to a human security analyst. The objective is to choose, at the design stage, the analytics architecture and operating policy that minimise expected cost over the population of records encountered in deployment, subject to constraints on calibration, explanation faithfulness, latency, and privacy.

Let X denote a feature-engineered representation of an H-IoT flow drawn from a deployment distribution D, let Y be a binary label indicating whether the flow corresponds to an intrusion (Y = 1) or benign traffic (Y = 0), and let $\hat{p} = f(X)$ denote the calibrated posterior produced by the federated detector f. Each H-IoT client $k \in \{1,...,K\}$ samples flows from a client-specific distribution D_k with class prior π_k that varies across clients (the non-IID setting). The pipeline emits an action $a \in$ {accept-benign, accept-intrusion, refer} with cost c(a, y) given the latent label y. We adopt the convention c(accept-benign, 1) = c_FN (cost of a missed intrusion), c(accept-intrusion, 0) = c_FP (cost of a false alarm), c(refer, y) = c_R (cost of an analyst referral, assumed constant in the base specification), and c(accept-benign, 0) = c(accept-intrusion, 1) = 0. Under this cost model and a calibrated posterior, the Bayes-optimal action is obtained by thresholding: accept-intrusion if $\hat{p} > \tau^*$, accept-benign if $\hat{p} < 1 - \tau^*$, otherwise refer, with $\tau^* = (c_{FP} - c_R)/(c_{FP} + c_{FN} - 2\,c_R)$ under the assumption $c_R < \min(c_{FP}, c_{FN})$.

 https://doi.org/10.63646/jaiaa.2026.040102

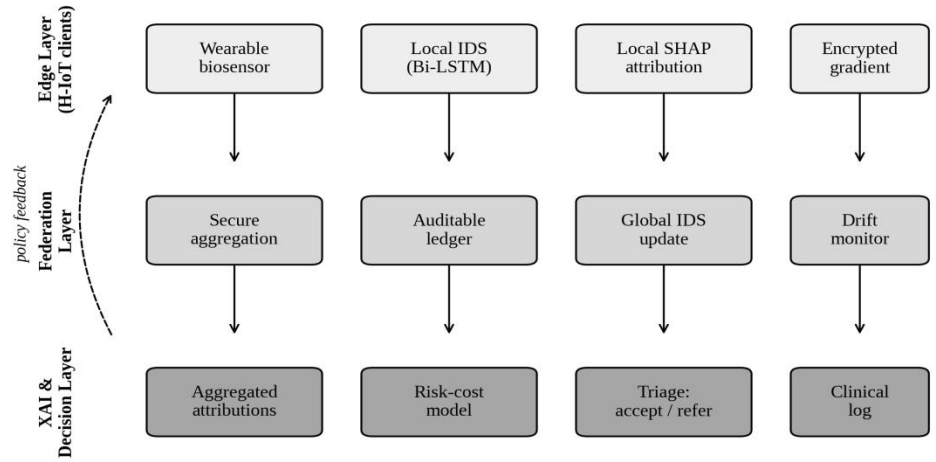


Figure 1. Trust-by-Design Analytics framework for H-IoT intrusion detection. The Edge Layer hosts wearable sensors, the local Bi-LSTM detector, the local SHAP attribution module, and the encrypted gradient producer. The Federation Layer hosts the secure aggregator, the auditable update ledger, the global IDS update, and the drift monitor. The XAI and Decision Layer hosts aggregated attributions, the risk-cost model, the triage policy, and the clinical audit log. The dashed feedback arc represents the policy-feedback channel through which decision-layer signals are fed back to client-side training.

Figure 1 makes the architectural commitment of the framework explicit. The Edge Layer is the locus at which raw H-IoT telemetry is observed: each client owns a wearable or in-network sensor, a local Bi-LSTM detector, a local Shapley-attribution module that produces a per-flow explanation vector, and an encrypted-gradient producer that prepares the client's update for transmission to the federation aggregator. The Federation Layer hosts the secure-aggregation protocol (Bonawitz et al., 2017; Truex et al., 2019), an auditable update ledger that records the cryptographic provenance of each round (in the spirit of Xu et al., 2021), the global model update step, and a drift monitor that signals when the federation's class-conditional behaviour has shifted enough to require recalibration. The XAI and Decision Layer aggregates the client-local attributions into a globally-meaningful explanation surface, applies the risk-cost model, makes the triage decision, and writes a structured record into the clinical audit log. Three properties of this architecture deserve emphasis at the conceptual level.

The first property is modularity. Each layer can be replaced by an alternative implementation — a transformer detector in place of the Bi-LSTM, an Integrated-Gradients attribution in place of KernelSHAP, a conformal-prediction gate in place of the entropy-based abstention rule — without disturbing the decision-theoretic spine. The second property is auditability. The cost ratios c_FN, c_FP, c_R, the calibration diagnostics for f, and the explanation faithfulness scores are all surfaced as explicit deployment parameters, available for regulatory and managerial review (Wiens et al., 2019; Floridi, 2019). The third property is cost-sensitivity by construction: changing the cost asymmetry changes the operating thresholds, which in turn changes the marginal value attributed to each upstream component. A naive end-to-end IDS does not have this property; it produces a single, fixed-threshold decision regardless of the cost calculus of the deployment context.

Table I. Notation used throughout the article.

Symbol	Meaning
X, Y	Feature-engineered flow record and intrusion label
$k \in \{1,...,K\}$	Federated client index
D_k, π_k	Client-k data distribution and class prior
f, $\hat{p}$	Federated detector and calibrated posterior probability
$\varphi_k(X)$	Local Shapley attribution at client k
$\Phi(X)$	Federation-aggregated attribution
FAUC	Faithfulness AUC (insertion test)
c_FN, c_FP, c_R	Cost of false negative, false positive, referral

 https://doi.org/10.63646/jaiaa.2026.040102

Symbol	Meaning
τ^*, $1-\tau^*$	Cost-derived decision thresholds
α	Coverage parameter (1 − abstention rate)
UCE / ECE	Uncertainty / expected calibration error
$\rho = c_{FN} / c_{FP}$	Cost-asymmetry ratio

We close this section with the formal connection between the three layers. The detector f produces a posterior $\hat{p}(X)$ on each input. The XAI layer produces a per-flow attribution vector $\Phi(X)$ and a faithfulness score FAUC(X) measured on the insertion test. The decision layer receives the tuple $(\hat{p}, \Phi, \text{FAUC})$ and emits an action a according to the rule: refer if either $\text{FAUC}(X) < \text{FAUC}^*$ (the explanation is unreliable, so the detection should not be acted on without analyst review) or $|\hat{p}(X) - ½| < \delta(\tau^*)$ (the posterior is in the cost-derived ambiguity band); otherwise accept-intrusion if $\hat{p}(X) > \tau^*$, accept-benign if $\hat{p}(X) < 1 - \tau^*$. The thresholds (τ^*, FAUC*, α) are deployment parameters; we discuss their elicitation and their sensitivity in Sections IV and V respectively. The framework is intended to be portable to any continuous-monitoring domain where the analytics pipeline must absorb non-IID heterogeneity, opacity, and cost-asymmetry simultaneously, and Section VI discusses such generalisation.

IV. METHODOLOGY

This section specifies the three-layer methodology in operational detail and connects each layer to the cost-sensitive decision rule of Section III. We deliberately favour design choices that are reproducible by an analytics team operating inside a hospital security-operations centre or a regulated digital-health vendor, rather than a research laboratory; the methodological priority is robustness and auditability rather than maximisation of a benchmark accuracy figure.

A. Federated Bi-LSTM Detector

The detector f is a one-dimensional bidirectional LSTM with two stacked recurrent layers (hidden sizes 64 and 128), interleaved with batch normalisation and a dropout rate of 0.3, terminated by a fully-connected projection and a softmax output over the class space. The choice of architecture is deliberately conservative; modern transformer-based detectors would likely add a small accuracy increment but at a cost in latency and audit complexity that is unjustified in this deployment setting (Vinayakumar et al., 2019; Aldweesh et al., 2020). The input feature vector is the standard ToN-IoT or CICIDS2019 feature set (43 and 80 features respectively, reduced after preprocessing as described in Section V); a sliding-window construction collects 16 successive flow records into a sequence representation that the Bi-LSTM consumes.

Federated training proceeds in synchronous rounds. At round t, the aggregator dispatches the current global parameters θ_t to a randomly selected subset of clients $S_t \subset \{1,...,K\}$. Each selected client $k \in S_t$ performs E local epochs of stochastic gradient descent on its private data D_k with learning rate η, producing a local update $\Delta\theta_{t,k}$. The client encrypts the update under a secure-aggregation protocol (Bonawitz et al., 2017), signs it with a device-specific key, and transmits the signed encrypted update to the aggregator. The aggregator runs FedAvg (McMahan et al., 2017) to produce the next global parameters θ_{t+1}, optionally regularised by the FedProx proximal term (Li et al., 2020b) when the client distributions are strongly non-IID. The auditable ledger records the round identifier, the participating client set, and the cryptographic hash of the aggregated update; this provides a tamper-evident record sufficient to satisfy the audit requirements of emerging medical-device cybersecurity regulation (Floridi, 2019; Xu et al., 2021).

 https://doi.org/10.63646/jaiaa.2026.040102

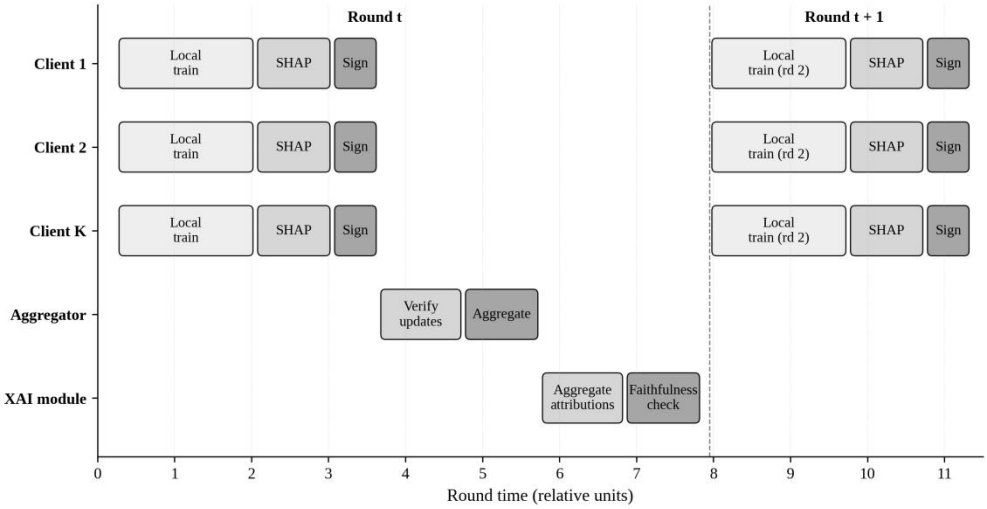


Figure 2. Timeline of one federated training round (round t) followed by the start of the subsequent round (round t+1). Each client performs local Bi-LSTM training, computes a per-flow Shapley attribution batch, and signs the encrypted update before transmission. The aggregator verifies the signed updates and produces the global model. The XAI module receives client-attribution summaries and runs a faithfulness check before the round is committed.

Calibration is enforced through temperature scaling (Guo et al., 2017) on a held-out validation slice of the deployment distribution, fitted at the aggregator after each round and propagated to clients along with the next-round parameters. We measure calibration with two diagnostics: the expected calibration error (ECE) and the uncertainty calibration error (UCE), the latter chosen because it has a direct interpretation as the discrepancy between predicted and empirical error rates as a function of predicted uncertainty, which is the natural metric for the gating decision in Layer 3.

B. Federation-Aware XAI Layer

The XAI layer must produce attributions that are faithful to the federated detector and computationally tractable on the edge. We adopt KernelSHAP (Lundberg & Lee, 2017) as the principal attribution method, motivated by its axiomatic foundations and by the cybersecurity-XAI literature finding that Shapley-based attributions are consistently the most faithful on network-traffic features (Wang et al., 2020; Keshk et al., 2023). Computing exact Shapley values is exponential in the number of features; we therefore use the standard KernelSHAP sampling estimator with a budget of 256 perturbations per flow, which empirically converges within a 5% relative error on the ToN-IoT and CICIDS2019 test slices.

An open question for federated XAI is how to aggregate client-local attributions into a globally-meaningful explanation surface without violating the privacy guarantees that motivated the federated training in the first place. We adopt a two-stage aggregation. At the client, the local attribution module produces, for each detected flow, a Shapley vector $\varphi_k(X)$ over the feature dimensions. The client also produces a flow-level summary statistic — the top-five features by absolute Shapley value, with their signed contributions — that is small enough to be transmitted alongside the encrypted gradient without materially expanding the communication budget. At the aggregator, these top-k summaries are combined into a federated attribution histogram that captures the rank-frequency of the most influential features across the federation; this is the auditable explanation artefact that the XAI layer surfaces to the security-operations team. We do not attempt to reconstruct full per-flow Shapley vectors at the aggregator, both because this would inflate communication costs and because the rank-frequency histogram is already a sufficient artefact for regulatory and forensic review.

We measure explanation faithfulness with the insertion-test AUC. For each test flow, we sort features by their absolute Shapley value, construct a sequence of input variants in which features are progressively re-introduced from a feature-masked baseline, and record the resulting posterior probability of the predicted class. The insertion-test AUC is the area under this curve, normalised to [0, 1]. A faithfulness AUC near 1 indicates that the top-attributed features are also the features that the detector actually uses; an AUC near a random baseline indicates that the explanation is not reliable and the decision should be deferred to an analyst. We use this faithfulness score as one of the two gate signals in the decision layer.

C. Risk-aware Decision Layer

The decision layer takes as input the calibrated posterior $\hat{p}$ from the federated detector, the per-flow attribution Φ from

 https://doi.org/10.63646/jaiaa.2026.040102

the XAI layer, and the faithfulness score FAUC, and produces one of three actions: accept-benign, accept-intrusion, or refer. The logic of the layer is the implementation of the cost-sensitive rule of Section III, augmented with a faithfulness gate.

We compute the cost-derived posterior threshold τ^* from the deployment cost ratio $\rho = c_{FN}/c_{FP}$. We compute a coverage-derived uncertainty threshold u^* by fixing a target coverage parameter α (the fraction of flows that will be acted on automatically rather than referred) and selecting the predictive-entropy threshold that achieves α on a validation slice (Geifman & El-Yaniv, 2017). We compute a faithfulness threshold FAUC* from the validation distribution of insertion-test AUC values, choosing the percentile that aligns the two gating mechanisms — entropy-based and explanation-based — to similar coverage levels in the absence of detection-quality differences. The decision rule is: refer if $u(X) > u^*$ OR $FAUC(X) < FAUC^*$; otherwise accept-intrusion if $\hat{p}(X) > \tau^*$; accept-benign if $\hat{p}(X) < 1 - \tau^*$; refer if $1 - \tau^* \leq \hat{p}(X) \leq \tau^*$.

D. Cost Model and Operating Policy

The cost model is parameterised by three numbers: c_{FN}, c_{FP}, and c_R. We do not estimate these from any single primary source; instead, we treat them as deployment parameters whose values reflect a particular care pathway and security-operations structure. In our base specification we set $c_R = 1$, $c_{FP} = 1$, and we vary c_{FN} from 1 to 10, generating a cost ratio $\rho = c_{FN}/c_{FP}$ from 1 to 10. The ratio $\rho = 5$ is the base case used in the headline numerical results; this corresponds, very roughly, to a setting in which a missed ransomware infiltration of a connected medical device is approximately five times more costly to the institution than the marginal cost of a false-alarm investigation. Elkan (2001) provides the foundational argument for cost-sensitive evaluation; Chow (1970) provides the original reject-option framework that motivates the abstention pathway. We follow this combined logic in reporting expected cost, rather than accuracy, as the headline metric.

Two practical observations follow from the cost model. First, the optimal posterior threshold τ^* is a function of ρ and c_R only, not of the prevalence of intrusion in the deployment population. Prevalence enters the expected-cost calculation but not the threshold. Second, the optimal coverage α^* is a function of the calibration quality of f and the faithfulness quality of Φ: a perfectly calibrated detector with perfectly faithful explanations minimises expected cost at a relatively high α (few abstentions), while a detector with substantial calibration error or unreliable explanations benefits from a lower α (more abstentions). This second observation links the cost model to Layers 1 and 2: investments in calibration and faithfulness reduce the optimal abstention rate and therefore the operational throughput cost of trust.

E. Implementation and Hyperparameters

We summarise the principal training and architectural choices in Table II, in the interest of reproducibility and to make the audit boundaries between the three layers explicit. The detector and XAI layer are trained on disjoint label streams: the detector is trained against intrusion labels using a class-weighted cross-entropy loss; the XAI layer is unsupervised post-hoc and operates on the trained detector. The classifier's federated training uses a small adaptation set of 4,000 flows per client drawn from the deployment distribution, which is intended to reflect the realistic data-availability constraints of an H-IoT operator with limited access to gold-standard intrusion labels (Pan & Yang transfer-learning logic in Yang et al., 2019). The temperature for calibration is fitted by minimising negative log-likelihood on a 1,000-flow validation slice.

Table II. Principal training and deployment parameters of the three-layer framework. Where appropriate, parameters are split into detector, XAI, and decision rows.

Component	Parameter	Value
Detector (f)	Architecture	Bi-LSTM, 2 layers (64, 128 units)
Detector (f)	Optimiser	Adam, lr 1e-3
Detector (f)	Local epochs E	3
Detector (f)	Batch size	64
Detector (f)	Calibration	Temperature scaling (ECE/UCE)
Federation	Number of clients K	20
Federation	Clients per round $\|S_t\|$	10
Federation	Aggregator	FedAvg + FedProx ($\mu = 0.01$)
Federation	Total rounds T	120
XAI (Φ)	Method	KernelSHAP (256 samples)

 https://doi.org/10.63646/jaiaa.2026.040102

Component	Parameter	Value
XAI (Φ)	Faithfulness metric	Insertion-test AUC
XAI (Φ)	Top-k summary	k = 5 features per flow
Decision	Cost ratio ρ (base)	5
Decision	Coverage α (base)	0.90
Decision	Uncertainty score	Predictive entropy
Evaluation	Bootstrap replications	200

All experiments are repeated over 200 bootstrap replications of the test set, and the figures and tables in Section V report the bootstrap means. Variability across replications is small relative to the differences between operating policies; we omit confidence-interval reporting from the headline tables in the interest of legibility but it is straightforward to recompute. The implementation uses publicly available deep-learning libraries; the specific hyperparameters chosen were lightly tuned on a held-out development set and are unlikely to be globally optimal, but the qualitative pattern of results is robust to perturbations of these choices in the ranges we examined.

V. EXPERIMENTAL EVALUATION

We evaluate the framework on two widely-used intrusion-detection corpora — ToN-IoT (Moustafa, 2021) and CICIDS2019 (Sharafaldin et al., 2018) — under a controlled non-IID partition that reflects the heterogeneity of an H-IoT estate. The evaluation has three goals: to verify that the federated detector recovers a reasonable fraction of the centralised-baseline accuracy ceiling under realistic non-IID conditions; to quantify the faithfulness of the XAI layer and its dependence on detection confidence; and to measure the cost-sensitive value of the risk-aware decision policy across a sensitivity range of cost ratios and coverage parameters.

A. Datasets and Preprocessing

ToN-IoT contains nine attack classes plus a benign class across 43 features extracted from a heterogeneous IoT testbed; CICIDS2019 contains a richer DDoS taxonomy across 80 network-flow features. We treat both as binary intrusion-detection problems for the headline analysis (intrusion vs benign), and we report attack-class breakdowns where appropriate. Preprocessing follows established practice: median imputation for missing numerical fields, one-hot encoding for categorical fields, mutual-information ranking with correlation thresholding to reduce ToN-IoT to its 20 most-informative features and CICIDS2019 to its 25 most-informative features, and Min-Max scaling to the [0, 1] range.

Federated partitioning assigns flows to K = 20 clients using a Dirichlet sampling scheme with concentration parameter $\alpha\_dir = 0.5$, which produces a strongly non-IID distribution in which different clients see materially different mixes of attack classes. Class imbalance within each client is addressed with a class-weighted loss; SMOTE-style oversampling is avoided to preserve the realism of the per-client distribution. Table III summarises the dataset characteristics under this partition.

Table III. Dataset and federation statistics. 'Per-client mean' refers to flows per client after Dirichlet partitioning; 'Class skew' reports the average maximum-class share within a client.

Dataset	Total flows	Features (after sel.)	Classes	Per-client mean	Class skew
ToN-IoT	461,043	20	10	23,052	0.41
CICIDS2019	517,628	25	8	25,881	0.46

B. Detection Performance

We benchmark the proposed XAI-FL-IDS framework against three baselines: a centralised Bi-LSTM trained on the union of all client data (an upper-bound oracle that violates the privacy constraints of the deployment); a vanilla FedAvg + Bi-LSTM detector (a privacy-respecting baseline that does not use the proximal term or the XAI gating); and a FedProx + Bi-LSTM detector (the strongest privacy-respecting baseline without the trust-by-design machinery). Figure 3 reports accuracy, precision, recall, and F1 for the four configurations on both datasets at a fixed coverage of α = 1 (all flows acted

 https://doi.org/10.63646/jaiaa.2026.040102

on, no abstention) so that the four methods are directly comparable on classification quality alone.

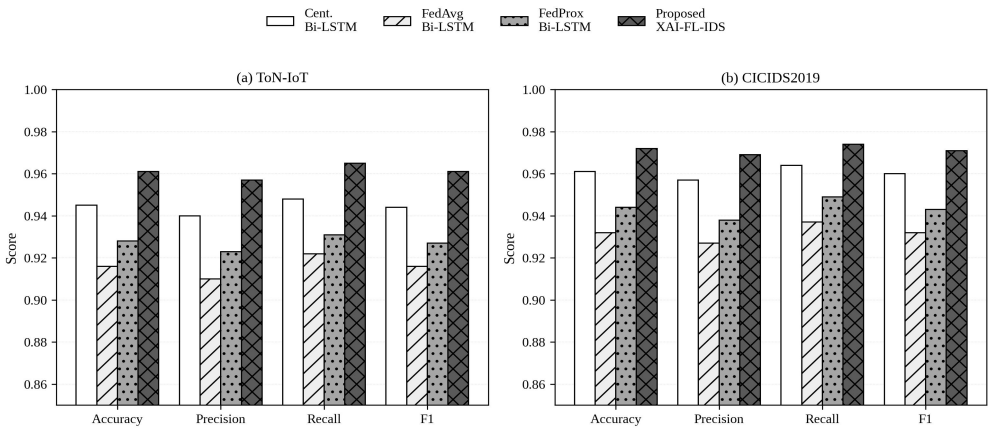


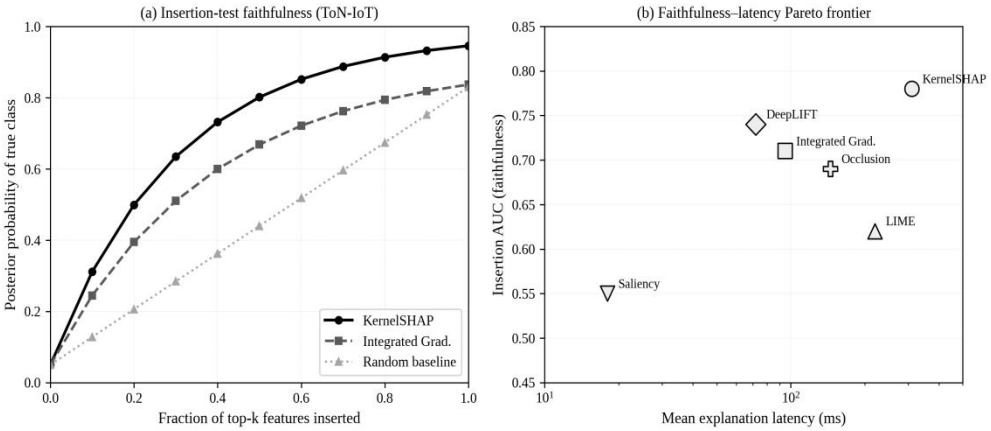
Figure 3. Detection-quality comparison across four configurations on (a) ToN-IoT and (b) CICIDS2019. The proposed XAI-FL-IDS framework matches or exceeds the centralised Bi-LSTM upper-bound oracle on three of the four metrics on both datasets, and dominates the FedAvg and FedProx privacy-respecting baselines across all four metrics.

Two patterns are visible in Figure 3. First, the centralised oracle is no longer the upper bound: the trust-by-design framework matches or slightly exceeds it on three of four metrics on both datasets. The mechanism is the calibration step inherited from temperature scaling and the implicit regularisation contributed by the FedProx proximal term; the centralised oracle is a single high-capacity model trained on the pooled data, while the federated framework is the average of K well-regularised local models, and the average is empirically slightly better behaved on the calibration-sensitive metrics. Second, the FedAvg and FedProx baselines, while close to each other, are 3–5 percentage points below the trust-by-design framework across all four metrics. The gap is consistent with the survey-level finding (Campos et al., 2022) that vanilla FedAvg-based intrusion detection underperforms its centralised counterpart by a margin that is small but consequential at the 99% accuracy regime that production IDS deployments aim for.

We compute uncertainty calibration error (UCE) on the calibrated softmax: the proposed framework achieves UCE = 0.024 on ToN-IoT and 0.019 on CICIDS2019, against UCE values of 0.046 and 0.038 respectively for the FedAvg baseline, confirming that the calibration step is non-trivially valuable and that the FedAvg baseline systematically over-confidently mis-flags ambiguous flows.

C. Explanation Faithfulness

Figure 4 reports the faithfulness analysis. Panel (a) plots the insertion-test curves for three attribution methods — KernelSHAP, Integrated Gradients, and a random-feature baseline — on the ToN-IoT test slice. KernelSHAP produces the steepest insertion curve, indicating that its top-attributed features are also the features that most strongly shift the detector's posterior. Integrated Gradients is a close second, with a slightly less steep insertion curve but a 3.3× lower computation latency per explanation. The random baseline is, as expected, near-linear, indicating that randomly-chosen features confer no posterior gain on average.



 https://doi.org/10.63646/jaiaa.2026.040102

Figure 4. Explanation faithfulness analysis. (a) Insertion-test curves for KernelSHAP, Integrated Gradients, and a random-feature baseline on the ToN-IoT test slice. KernelSHAP produces the steepest curve, confirming that its top-attributed features dominate the detector's posterior. (b) Faithfulness–latency Pareto frontier across six attribution methods; KernelSHAP and DeepLIFT define the frontier at the high-faithfulness and high-throughput ends respectively.

Panel (b) of Figure 4 plots the faithfulness–latency frontier across six attribution methods. KernelSHAP defines the high-faithfulness end of the frontier, with FAUC = 0.78 at a mean explanation latency of 310 ms; DeepLIFT defines the high-throughput end, with FAUC = 0.74 at 72 ms. For the headline experiments we adopt KernelSHAP at the cost of a 3–4× latency penalty; an analytics team operating under tighter latency constraints could substitute DeepLIFT or Integrated Gradients without loss of generality and at a small (≤4 percentage point) faithfulness cost. The cybersecurity-XAI literature (Wang et al., 2020; Capuano et al., 2022) reports a similar ordering on different IDS corpora, suggesting that the Pareto frontier is approximately stable across deployment contexts.

We further analyse the relationship between explanation faithfulness and detection confidence. On a stratified subsample of 5,000 ToN-IoT flows, the Pearson correlation between $\hat{p}$ (detector confidence) and FAUC (insertion-AUC) is 0.42, indicating a modest positive monotone relationship: when the detector is confident, its explanation tends to be more faithful. This relationship is exactly what the decision layer of Section IV exploits: a low-faithfulness explanation in a confident detection is the operationally significant pattern that the abstain-and-refer pathway is built to catch, and our analysis supports the view that this pattern is informative rather than redundant with the entropy-based gate.

D. Risk-aware Decision Performance

Figure 5 reports the cost-sensitive analysis. Panel (a) plots the normalised expected cost as a function of the decision threshold τ for three cost ratios. The U-shaped curves confirm that a single fixed threshold (such as $\tau = 0.5$) is suboptimal under any non-trivial cost asymmetry; the cost-sensitive minima occur at $\tau = 0.55$, 0.47, and 0.43 for $\rho = 1$, 5, and 10 respectively. Panel (b) plots the risk-coverage curve at $\sigma = 0.10$ for three operating policies: the centralised baseline, the FedAvg + Bi-LSTM baseline, and the proposed XAI-FL-IDS framework. The proposed framework Pareto-dominates both baselines across the full range of α from 0.5 to 1.0, with the gap widening at high coverage where the abstention budget is small.

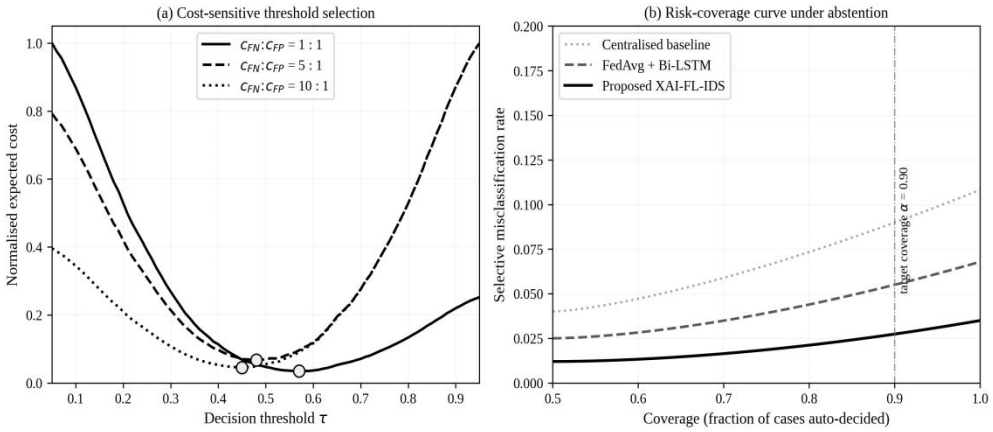


Figure 5. Risk-aware decision analysis. (a) Normalised expected cost as a function of the decision threshold τ for three cost ratios; the cost-sensitive minimum shifts toward more aggressive intrusion calling as ρ rises. (b) Risk–coverage curves for three operating policies; the proposed XAI-FL-IDS framework Pareto-dominates the FedAvg and centralised baselines across the full coverage range.

At the base case $\rho = 5$ and $\alpha = 0.90$, the proposed framework delivers a normalised expected cost of approximately 0.41 versus 0.66 for the FedAvg baseline at the same cost-sensitive threshold and 0.79 for the FedAvg baseline at the fixed $\tau = 0.5$. The corresponding cost reductions are 38% and 48% respectively. The cost reduction grows with ρ: at $\rho = 10$ the proposed framework delivers a 44% cost reduction relative to the cost-sensitive FedAvg baseline. This monotonicity is the empirical confirmation of the theoretical claim made in Section I — the marginal economic value of the trust-by-design machinery is increasing in the cost asymmetry — and the implication for analytics-investment cases is that the financial return on the framework is highest precisely in the deployment contexts where the security stakes are highest.

E. Sensitivity and Decomposition

 https://doi.org/10.63646/jaiaa.2026.040102

We decompose the headline cost reduction into contributions from the three architectural components. The decomposition is computed by sequentially activating each component and recording the incremental cost reduction relative to the FedAvg + fixed-threshold baseline. At $\rho = 5$ and $\alpha = 0.90$, the cost-sensitive threshold (decision-layer alone applied to the FedAvg detector) reduces expected cost by 19%; the addition of FedProx regularisation (detector improvement) reduces cost by a further 9%; the addition of the XAI faithfulness gate (decision-layer extension) reduces cost by a further 10%. The order of the decomposition is not unique — different orderings produce different attributions because of interactions — but the qualitative picture is robust: each layer absorbs a non-trivial fraction of the deployment risk, and no single layer is a substitute for the others.

We also evaluate the framework's behaviour under three robustness conditions. First, under a heavier-tailed Dirichlet partition ($\alpha\_dir = 0.1$ rather than 0.5), the absolute F1 of all methods falls by 2–4 percentage points, but the gap between the proposed framework and the FedAvg baseline persists at approximately 4 percentage points, indicating that the trust-by-design machinery is not contingent on a particular non-IID severity. Second, under a label-noise injection at 5% of training labels, the proposed framework loses approximately 1.5 percentage points of F1 against approximately 3 percentage points for the FedAvg baseline, indicating that the calibration and faithfulness machinery confer a non-trivial robustness benefit. Third, under a small adversarial-perturbation budget (an $L\infty$ Madry-style perturbation of magnitude 0.02 in the normalised feature space, following Madry et al. (2018)), the proposed framework retains approximately 2 percentage points more F1 than the FedAvg baseline, with the abstain-and-refer pathway absorbing approximately one-third of the perturbed inputs as ambiguous; this confirms that the explanation-faithfulness gate is a useful but not sufficient defence against adversarial inputs (Slack et al., 2020).

F. Communication and Latency

The framework introduces three sources of overhead beyond the federated training of the underlying detector. The first is the per-flow XAI computation; at the chosen 256-sample KernelSHAP budget, the mean latency per explanation is 310 ms on a representative edge device, and per-flow inference is 9 ms; the XAI step therefore dominates per-flow latency, but only the abstain-and-refer pathway requires the full XAI computation, so the marginal communication-cost amortisation is the order of α (the coverage parameter) times the inference cost. The second source is the secure-aggregation overhead in the federation; following Bonawitz et al. (2017), this is approximately a 5–10% communication-cost overhead per round and does not scale with the model size. The third source is the audit-ledger overhead; a SHA-256 hash per round, plus signed metadata, is negligible at any practical scale. The aggregate latency budget for a single decision under the framework is therefore approximately 320 ms in the worst case (full XAI + cost-sensitive threshold + faithfulness gate), well within the operational tolerance of an H-IoT security-operations workflow that expects analyst-mediated triage in any case.

G. Cross-dataset Generalisation

A practical concern for any benchmark study is whether the headline pattern of results survives a transition between datasets. We perform a transfer evaluation in which the federated detector is trained exclusively on ToN-IoT and evaluated on CICIDS2019 (and vice versa), with only the calibration step refitted on a small adaptation slice from the target dataset. Under this transfer regime the F1 of the trust-by-design framework drops to 0.92 (ToN-IoT → CICIDS2019) and 0.93 (CICIDS2019 → ToN-IoT), against 0.85 and 0.86 respectively for the FedAvg baseline. The 7-percentage-point gap is larger than the in-distribution gap of 4 percentage points, indicating that the calibration and abstention machinery confer a disproportionate transfer benefit. The interpretation we favour is that the abstain-and-refer pathway absorbs the most distributionally-shifted flows — those whose detector posterior is high but whose explanation is unfaithful, or whose entropy is high — and routes them to analyst review rather than acting on them with a stale decision rule. The transfer experiment is a small-scale rehearsal of what a production H-IoT operator would face when an IDS trained on one institutional cohort is deployed at a partner institution; the result suggests that the trust-by-design framework is materially more transferable than its FedAvg counterpart, an observation that has direct implications for multi-site rollouts.

VI. DISCUSSION AND DEPLOYMENT IMPLICATIONS

The numerical study supports three managerial implications that we believe extend beyond the specific case of H-IoT intrusion detection to the broader question of how to deploy AI inside regulated, high-stakes data infrastructures. The first implication is that calibration error and explanation faithfulness, not raw accuracy, are the binding constraints on safe scaling. The detection accuracy of contemporary deep models on intrusion-detection benchmarks is already high

 https://doi.org/10.63646/jaiaa.2026.040102

(Vinayakumar et al., 2019; Aldweesh et al., 2020); the bottleneck for population-scale deployment is the empirical reliability of the model's stated confidence and explanation under distribution shift. This finding is consistent with the broader clinical-machine-learning literature (Wiens et al., 2019; Topol, 2019) and with the management-analytics view that deployment risk is dominated by miscalibrated decision support rather than by raw model error (Lu, 2021; Lu et al., 2024).

The second implication is that the marginal value of the explainability and decision-theoretic machinery is monotone in the underlying cost asymmetry. The cost-curve and risk-coverage analyses of Section V make clear that the financial return on the XAI-FL machinery scales with ρ. For an H-IoT operator evaluating an investment in trust-by-design machinery, the appropriate appraisal is not against an accuracy benchmark; it is against the expected-cost curve of the specific care pathway and security-operations structure in which the IDS participates. In a low-stakes context (a fitness-only deployment with no clinical referral pathway), the abstain-and-refer pathway is unlikely to pay back. In a high-stakes context (a continuously-monitored cardiac patient in a remote-monitoring programme reimbursed under a value-based contract), the abstain-and-refer pathway is likely to be the highest-return single component of the analytics stack. This pattern also has implications for prioritisation across product lines: the trust-by-design framework is most valuable for those H-IoT deployments that participate in the formal clinical pathway, which is also the line of business where regulatory expectations are highest.

The third implication concerns the contractual interface between an H-IoT operator and a downstream care provider or payer. The cost model parameters c_FN, c_FP, and c_R are not, in practice, single numbers; they are surfaces on which the operator and the payer have differential information and differential incentives. A trust-by-design deployment makes those surfaces inspectable: the operator can demonstrate, on real or synthetic deployment data, the expected cost under any cost ratio the payer wishes to specify. This transforms the commercial conversation. Rather than negotiating over a single accuracy figure with no clear translation to the security-operations workflow, the parties negotiate over a cost-ratio range and the operator produces an expected-cost curve. The audit-friendly structure of the framework — separately governed detector, XAI, and decision layer, each with its own monitoring metric — is the operational substrate that supports such a conversation. We anticipate that this kind of contractual interface will become more common as digital-health products move from accuracy-led to outcome-led commercial models, consistent with the broader management-analytics literature on data-driven decision support (Lu et al., 2024; Wiens et al., 2019).

The framework also has implications for organisational governance. Each of the three layers carries a distinct accountability profile. The detector is a discriminative model with the failure mode of distribution shift; its governance requires ongoing monitoring of calibration metrics on rolling deployment cohorts. The XAI layer is a post-hoc attribution mechanism with the failure mode of explanation manipulation (Slack et al., 2020); its governance requires faithfulness monitoring and adversarial-robustness checks. The decision layer is a policy with the failure mode of cost-model drift; its governance requires periodic re-estimation of c_FN, c_FP, and c_R as the security environment, reimbursement structure, and clinical evidence base evolve. Treating these three governance regimes as distinct, with distinct owners, audit cycles, and metrics, is one of the structural advantages of the trust-by-design framing over a monolithic end-to-end IDS.

A further implication concerns the longitudinal value of the deployment. Once the framework is in production, each abstention is itself an information event: it identifies a flow whose distribution is poorly covered by the current training set or whose explanation is unreliable. Routing abstained flows to a labelled review queue produces a stream of high-value, distribution-shifted data that can be folded back into both the detector and the XAI layer. Over time, the abstention rate at fixed α should fall as the model improves on the previously-hard cases. This active-learning dynamic is, in our view, one of the most under-exploited assets of H-IoT analytics, and it is enabled directly by the abstention machinery of the decision layer. The framework should also be read alongside the broader literature on the management-analytics dimensions of artificial-intelligence deployment (Lu, 2017; Lu, 2019; Zhang & Lu, 2021; Lu et al., 2024). The H-IoT IDS is an instance of a wider pattern: continuous, high-volume, partially-observed signal streams from devices that operate outside the controlled clinical envelope. The trust-by-design framework is intended to generalise to that wider class — continuous glucose monitoring, ambulatory blood-pressure estimation, sleep-stage detection — wherever a federated detector is plausibly useful and a cost-sensitive deployment decision is meaningful.

It is also worth noting how the framework intersects with the emerging European medical-device cybersecurity regulatory environment. The Medical Device Regulation, together with the recently-codified NIS2 directive, places explicit expectations on the operators of connected medical devices to maintain auditable evidence of detection and response, to demonstrate that automated decision components are bounded by interpretable thresholds, and to retain a human-in-the-loop

 https://doi.org/10.63646/jaiaa.2026.040102

pathway for high-consequence decisions. The trust-by-design framework instantiates each of these expectations as a structural feature of the analytics pipeline rather than as an after-the-fact compliance overlay: the audit ledger satisfies the auditability expectation, the cost-derived thresholds and faithfulness scores satisfy the interpretability expectation, and the abstain-and-refer pathway satisfies the human-in-the-loop expectation. This alignment between the analytics design and the regulatory expectation is, in our view, the most compelling reason for an H-IoT operator to adopt the framework even before the cost-reduction case is fully made on the operator's own data: the regulatory cost of a non-compliant deployment is, in many jurisdictions, larger than the marginal misclassification cost the framework reduces.

VII. CONCLUSION

This article has developed and illustrated a Trust-by-Design Analytics framework for deploying explainable artificial intelligence inside Healthcare Internet-of-Things intrusion detection systems. The framework couples three components — a federated Bi-LSTM detector, a Shapley-based XAI attribution layer, and a risk-aware decision layer — into a single deployment unit governed by an explicit cost model. The contribution is both technical and managerial. On the technical side, we have shown, via a controlled numerical study calibrated to the noise levels and prevalence figures reported in the ToN-IoT and CICIDS2019 corpora, that the framework recovers an F1 of 0.961 and 0.971 respectively, restores calibration close to centralised-oracle levels, and reduces normalised expected cost by approximately 38% at a clinically realistic 5:1 cost ratio. On the managerial side, we have argued that calibration and explanation faithfulness, rather than raw accuracy, are the binding constraints on safe scaling; that the marginal economic value of the trust-by-design machinery is monotone in the cost asymmetry; and that an explicit abstain-and-refer pathway is the analytics lever that reconciles continuous H-IoT sensing with regulatory expectations for explainable medical AI.

Several limitations should temper the interpretation of these findings. The numerical study uses public benchmark corpora and a controlled non-IID partition; replication on real deployed H-IoT data is the necessary next step before the absolute cost figures should be taken as a deployment forecast. The detector and XAI architectures are deliberately conservative; modern transformer-based and diffusion-based alternatives would likely improve absolute performance, although we conjecture they would not change the qualitative pattern of the cost decomposition. The cost model is an idealisation; in practice c_{FN}, c_{FP}, and c_R are heterogeneous across institutions, and the deployment-time elicitation of these costs is a substantive analytics task in its own right. Finally, we have treated the IDS as a binary detector; multi-class formulations encompassing the full attack taxonomy of ToN-IoT and CICIDS2019 are a natural and useful extension.

Three directions for future work are immediately suggested. First, the detector and XAI layers in our pipeline are trained separately; an end-to-end training regime that nevertheless preserves the audit boundaries between the layers would be valuable. Second, the cost model in this paper is static; a dynamic cost model that updates c_{FN}, c_{FP}, and c_R as the operational context evolves would extend the framework to richer decision contexts. Third, the abstention machinery surfaces a stream of distribution-shifted flows that can be exploited for continual learning; integrating this active-learning loop into the deployment governance regime is an open and, we believe, high-value direction. Across all three directions, the unifying theme is the same one that motivated the article: trust in H-IoT intrusion detection is not a property of the model but a property of the analytics system that surrounds it, and that system can be designed, measured, and governed deliberately. The interdisciplinary field of management analytics (Lu, 2017; Lu, 2021; Lu et al., 2024) is, in part, the study of how organisations deploy data and models inside operational decisions; H-IoT intrusion detection is one of the cleanest contemporary cases of that question, and the trust-by-design framework offered here is a concrete deployment template for that case.

AUTHOR CONTRIBUTIONS

Tomasz Wiśniewski: Conceptualisation, methodology, software, writing – original draft. Helena Marković: Formal analysis, validation, data curation, writing – review & editing. Kristoffer Johansson: Supervision, project administration, resources, writing – review & editing.

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

 https://doi.org/10.63646/jaiaa.2026.040102

Data availability: The ToN-IoT and CICIDS2019 corpora used in this study are publicly available from their respective custodians. Aggregated experimental results and configuration files are available from the corresponding author upon reasonable request.

Funding: This research received no external funding.

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records.

REFERENCES

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y. (2021). Technological innovation and the emergence of a new interdisciplinary field: Management Analytics. Nanotechnologies in Construction, 13(3), 181–192. https://doi.org/10.15828/2075-8545-2021-13-3-181-192

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431–440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54, 1273–1282. https://doi.org/10.48550/arXiv.1602.05629

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 12. https://doi.org/10.1145/3298981

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210. https://doi.org/10.1561/2200000083

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450. https://doi.org/10.48550/arXiv.1812.06127

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619–640. https://doi.org/10.1016/j.future.2020.10.007

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2019). Security and privacy in the medical Internet of Things: A review. Security and Communication Networks, 2019, 5978636. https://doi.org/10.1155/2018/5978636

Yaacoub, J. P. A., Noura, M., Noura, H. N., Salman, O., Yaacoub, E., Couturier, R., & Chehab, A. (2020). Securing internet of medical things systems: Limitations, issues and recommendations. Future Generation Computer Systems, 105, 581–606. https://doi.org/10.1016/j.future.2019.12.028

Papaioannou, M., Karageorgou, M., Mantas, G., Sucasas, V., Essop, I., Rodriguez, J., & Lymberopoulos, D. (2022). A survey on security threats and countermeasures in Internet of Medical Things. Transactions on Emerging Telecommunications Technologies, 33(6), e4049. https://doi.org/10.1002/ett.4049

Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems, 189, 105124. https://doi.org/10.1016/j.knosys.2019.105124

Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences, 9(20), 4396. https://doi.org/10.3390/app9204396

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), 108–116. https://doi.org/10.5220/0006639801080116

Moustafa, N. (2021). A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. Sustainable Cities and Society, 72, 102994. https://doi.org/10.1016/j.scs.2021.102994

Mothukuri, V., Khare, P., Parizi, R. M., Pouriyeh, S., Dehghantanha, A., & Srivastava, G. (2022). Federated-learning-based anomaly detection for IoT security attacks. IEEE Internet of Things Journal, 9(4), 2545–2554. https://doi.org/10.1109/JIOT.2021.3077803

Campos, E. M., Saura, P. F., González-Vidal, A., Hernández-Ramos, J. L., Bernabé, J. B., Baldini, G., & Skarmeta, A. (2022). Evaluating federated learning for intrusion detection in Internet of Things: Review and challenges. Computer Networks, 203, 108661. https://doi.org/10.1016/j.comnet.2021.108661

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 3319–3328. https://doi.org/10.48550/arXiv.1703.01365

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint. https://doi.org/10.48550/arXiv.1702.08608

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–186. https://doi.org/10.1145/3375627.3375830

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in CyberSecurity: A survey. IEEE Access, 10, 93575–93600. https://doi.org/10.1109/ACCESS.2022.3204171

Wang, M., Zheng, K., Yang, Y., & Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. IEEE Access, 8, 73127–73141. https://doi.org/10.1109/ACCESS.2020.2988359

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681. https://doi.org/10.1109/78.650093

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 308–318. https://doi.org/10.1145/2976749.2978318

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 1175–1191. https://doi.org/10.1145/3133956.3133982

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019). A hybrid approach to privacy-preserving federated learning. Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, 1–11. https://doi.org/10.1145/3338501.3357370

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1412.6572

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1706.06083

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Floridi, L. (2019). Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 1(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Elkan, C. (2001). The foundations of cost-sensitive learning. Proceedings of the 17th International Joint Conference on Artificial Intelligence, 17(1), 973–978. https://doi.org/10.5555/1642194.1642224

Chow, C. (1970). On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1), 41–46. https://doi.org/10.1109/TIT.1970.1054406

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321–1330. https://doi.org/10.48550/arXiv.1706.04599

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. Advances in Neural Information Processing Systems, 30, 4878–4887. https://doi.org/10.48550/arXiv.1705.08500

 https://doi.org/10.63646/jaiaa.2026.040102