

ARX-IoT: A Closed-Loop Framework for Resolving Explainability Conflicts and Enabling Actionable IoT Botnet Detection

Edja Olivier Tché^{1*}, Amanvon Ferdinand Atta², Beman Hamidja Kamagaté³, Souleymane Oumtanaga¹

¹Unité Mixte de Recherche et d'Innovation en Mathématiques et Sciences du Numérique (MSN), Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire

²Unité de Recherche et d'Expertise Numérique, Université Virtuelle de Côte d'Ivoire (UVCI), Abidjan, Côte d'Ivoire

³Laboratoire des Sciences et Technologies de l'Information et de la Communication (LASTIC), École Supérieure Africaine des Technologies de l'Information et de la Communication (ESATIC), Abidjan, Côte d'Ivoire

Email: *edja.tche24@inphb.ci

How to cite this paper: Tché, E.O., Atta, A.F., Kamagaté, B.H. and Oumtanaga, S. (2026) ARX-IoT: A Closed-Loop Framework for Resolving Explainability Conflicts and Enabling Actionable IoT Botnet Detection. *Open Journal of Applied Sciences*, 16, 3981-4023.

<https://doi.org/10.4236/ojapps.2026.1610221>

Received: September 6, 2026

Accepted: October 24, 2026

Published: October 27, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Botnets are spreading fast across the Internet of Things, and that has pushed defenders toward intrusion detection systems built on deep learning. These models detect attacks well, but they work as black boxes, which makes it hard to automate a response in settings where a wrong move is costly. Explainable AI was supposed to solve the transparency problem, yet it introduced one of its own: the explanations don't agree with each other. Run SHAP and Anchors on the same prediction and you can get two different stories about why the model flagged it. For an operator who has to act on that prediction, this is not a minor inconsistency; it is a real source of doubt. ARX-IoT (Adaptive Resolved XAI) was designed to close the gap between an explanation that merely describes and a decision that has to be made. Rather than treating disagreement between explanations as noise to be smoothed over, the framework reads it as a signal and feeds it back into a control loop. To do this, it puts two very different kinds of explanation on the same footing: TimeSHAP, which assigns each feature a real-valued weight, and Anchors, which produces a binary yes/no mask. Both are mapped into a shared probabilistic space, so they can actually be compared. From there, the framework computes an actionability score that folds together three things: how stable the explanation is over time, how confident the model is, and a probabilistic veto factor. That veto is the safeguard: when the two explanations contradict each other badly, it steps in and pulls back the system's autonomy instead of letting it act on shaky ground.

A feedback loop tunes how sensitive the system is to conflict, adjusting via gradient descent, so that the effect of intervening stays close to a chosen target. We tested ARX-IoT with an attention-based LSTM detector and ran it 103 times on two datasets, CICIoT2023 and TON_IoT. Detection held up throughout, with accuracy above 0.96 on both (see **Table 1** for the full per-metric breakdown). The explanations, meanwhile, diverged in a consistent, structural way: the total variation distance reached 0.5382 on CICIoT2023. The ablation study is where the veto earns its place: switching it on cut the risk of a wrong mitigation decision by 26.7% on CICIoT2023 and 10.8% on TON_IoT compared with a system left unregulated, and the size of that improvement tracked how close the observed conflict sat to the critical veto threshold. In short, ARX-IoT works as a transparent safety filter: it catches the risky calls without touching the detection performance the underlying model already delivers.

Keywords

Actionability, Causally-Inspired Veto, Closed-Loop Control, Explainable Artificial Intelligence, Explanatory Conflicts

1. Introduction

The rapid expansion of the Internet of Things (IoT) has reshaped the cybersecurity threat landscape, placing botnets at the center of many modern attack strategies. Faced with this growing complexity, Deep Learning-based Intrusion Detection Systems (IDSs) have become essential tools for identifying sophisticated attack patterns [1]. However, their black-box nature [1] [2] limits their adoption in critical environments, where every security decision must be explainable, traceable, and defensible.

Against this backdrop, Explainable Artificial Intelligence (XAI) has helped make machine learning models more transparent. This progress has, however, revealed a deeper problem: explanation methods often disagree with one another. Recent work highlights the inherent instability of explanations produced by XAI algorithms. Kumar and Thing [3] found that XAI explanations conflict across most of the IoT NIDS models they examined, confirming that widely used methods such as SHAP and LIME can yield divergent interpretations of the same instance. Such disagreement creates substantial uncertainty for security operators who must act on the resulting predictions.

This instability is more than a theoretical concern; it can become a systemic weakness. An inconsistent explanation may reduce a system's overall robustness against adversarial attacks [1]. Moreover, a lack of temporal consistency, which is particularly important in high-velocity IoT traffic, can trigger unstable defensive responses and undermine the reliability of protection mechanisms [4].

Several lines of research have therefore attempted to reduce the gaps between explanation methods. A first group aims to establish explanatory consensus. EX-

AGREE [5] selects, a priori, the models whose explanations best align with a given stakeholder through Stakeholder-Machine Agreement, whereas Consensus Importance [6] combines explanations from several XAI methods using fidelity-based weights. Although these approaches improve overall consistency, they act mainly before or after explanation generation and leave the operational decision-making process unchanged.

A second group seeks to transform explanations into concrete actions. An explanation becomes genuinely useful when it is actionable, that is, when it can drive a reliable, consistent, and deterministic response [7]. In this direction, XAI4C [8] uses SHAP values to identify operational conflicts in O-RAN networks, while the framework proposed by Morrison *et al.* [9] incorporates XAI explanations into the cybersecurity decision-making process. These studies demonstrate the value of linking explanations to actions; nevertheless, they generally assume that the explanations are consistent and do not account for disagreements among multiple explanation methods.

The literature thus reveals a persistent gap between theoretical explainability and operational actionability in IoT intrusion detection. The explanatory conflicts observed in IoT NIDSs [3], together with recent critical evaluations of XAI methods [10] [11], suggest that current approaches still lack sufficient stability, consistency, and regulatory control. In this work, we refer to this phenomenon as *explanatory conflict* and treat it as a major source of uncertainty in automated decision-making.

More specifically, existing approaches do not simultaneously provide: 1) a formal method for quantifying conflict across multiple XAI methods, 2) an online decision-regulation mechanism driven by that conflict, and 3) decisions that adapt to the temporal evolution of IoT traffic. To address these limitations, we propose ARX-IoT, a decision-regulation framework built around explanatory conflict analysis. Unlike prior work, ARX-IoT neither selects a single explanation nor forces a simple consensus. Instead, it treats explanatory conflict as a decision signal and incorporates it into an adaptive closed-loop feedback mechanism.

The main contributions of this work are threefold:

1) ARX-IoT formally quantifies the conflict between TimeSHAP [12] and Anchors [13] using the total variation distance $D(t)$ and the conflict coefficient $C(t)$, rather than relying only on the global agreement or correlation measures used by EXAGREE [5] and Consensus Importance [6].

2) ARX-IoT introduces an online decision-regulation mechanism based on a closed-loop architecture. Rather than selecting a model a priori, as EXAGREE [5] does, or aggregating explanations a posteriori, as in Consensus Importance [6], ARX-IoT applies a regulation factor $\Gamma(t)$ and continuously adjusts the sensitivity parameter $\beta(t)$ according to the conflict observed at each time step. This mechanism enables the system to adapt to changes in IoT traffic, whereas earlier approaches remain largely static.

3) The framework is experimentally validated for IoT cybersecurity on the

TON_IoT and CICIoT2023 datasets. The results show that explicitly accounting for explanatory conflict reduces the risk of inappropriate mitigation decisions while strengthening the robustness of the overall decision-making process.

In summary, EXAGREE transforms disagreement into a model-selection criterion, Consensus Importance uses it as an aggregation mechanism, and XAI4C applies XAI to operational conflict detection. By contrast, ARX-IoT treats explanatory conflict as a decision-regulation signal that dynamically adjusts the system's level of automation according to explanatory uncertainty, without requiring prior model selection or continuous human intervention.

2. Related Work

2.1. Conflicts among Explainability Methods: An Emerging Problem

The adoption of Explainable Artificial Intelligence (XAI) in Intrusion Detection Systems (IDSs) has grown substantially to improve the transparency of deep learning models used in cybersecurity. Recent literature reviews show that XAI methods are increasingly integrated into IDSs to improve decision interpretability, strengthen analysts' trust, and facilitate the validation of security alerts [14] [15]. Widely used methods, including SHAP, LIME, and Anchors, provide local or global explanations that help identify the features influencing model decisions.

However, several recent studies show that different XAI methods can produce substantially different explanations for the same prediction, thereby raising concerns about the reliability of the resulting interpretations [3] [6] [16] [17]. In particular, Kumar and Thing [3] show that leading explainability methods used in deep learning-based IDSs frequently disagree and may produce conflicting explanations for the same network attack.

This phenomenon, commonly referred to as *disagreement among XAI methods*, arises primarily from fundamental differences between explanation paradigms. Attribution-based methods, such as SHAP, estimate the contribution of each feature to a model decision, whereas rule-based methods, such as Anchors, identify sufficient logical conditions that locally explain a prediction. Because these approaches rely on different explanatory principles, they may naturally produce divergent interpretations, particularly in intrusion detection systems characterized by highly nonlinear relationships among features [3] [11] [17].

Beyond being an interpretability issue, this divergence may reduce confidence in automated decisions, complicate the validation of security alerts, and directly affect the quality of decisions made by human operators [6] [9] [17].

2.2. State of the Art in Conflict-Resolution Approaches

In response to this problem, several families of approaches have been proposed to reduce or characterize conflicts among explainability methods.

A first family seeks to measure or establish consensus among multiple XAI methods. EXAGREE [5] selects models whose explanations exhibit the highest agree-

ment with stakeholder expectations, whereas Consensus Importance [6] provides a global agreement measure across multiple explainability methods by aggregating their contributions according to their fidelity. Similarly, Stow [17] proposes a measure of disagreement between SHAP and LIME, while Dib and Capus [11] clarify the conceptual distinctions among different families of XAI methods. Although these approaches improve the analysis and characterization of explanatory divergence, they remain primarily focused on measuring or aggregating explanations and do not directly intervene in the decision-making process.

A second family of studies aims to make explanations more actionable. Singh *et al.* [7] show that an explanation acquires practical value when it can guide a concrete and usable decision. In this context, XAI4C [8] uses SHAP values to identify operational conflicts in O-RAN networks, while Morrison *et al.* [9] demonstrate that imperfections in XAI explanations can directly affect the quality of human decisions.

In parallel, several XAI frameworks have been developed for IDSs and IoT environments. Wang *et al.* [18] were among the first to integrate SHAP into an explainable intrusion detection system. Subsequently, Arreche *et al.* [15] proposed a comprehensive framework for comparing multiple XAI methods applied to IDSs, while Sharma *et al.* [2] and Saied and Guirguis [19] demonstrated the value of explainable approaches for understanding IoT attacks and botnets. More recently, Kumar and Thing [3] confirmed that disagreement among XAI methods has become a significant challenge for deep learning-based IDSs.

Another line of research uses rule extraction to transform model explanations into directly actionable knowledge. Studies by Uccello *et al.* [20] [21], Li *et al.* [22], and Ables *et al.* [23] show that rules derived from machine learning models can be integrated into IDSs or Security Information and Event Management (SIEM) platforms to improve interpretability and facilitate the operational deployment of detection systems.

2.3. Identified Research Gaps

The literature still leaves several gaps unresolved.

First, although disagreement among XAI methods is increasingly recognized [3] [6] [17], existing studies mainly focus on comparison metrics or explanation aggregation. To the best of our knowledge, no existing approach specifically quantifies the conflict between a temporal attribution method such as TimeSHAP and a rule-based method such as Anchors within a unified decision-oriented framework.

Second, current frameworks generally treat explanations as an endpoint of the interpretability process. Information produced by XAI methods is primarily used to interpret model decisions rather than as a decision-regulation signal capable of automatically adapting system behavior to explanatory uncertainty [8] [9] [15].

Third, most existing approaches remain essentially static. They do not explicitly

account for the dynamic nature of IoT traffic, which is characterized by substantial temporal variability, nor for the evolution of explanatory conflict over time [19] [24].

2.4. Positioning of ARX-IoT

To address these limitations, we propose ARX-IoT, a decision-regulation framework based on explanatory conflict analysis.

Unlike existing approaches, ARX-IoT neither seeks to select a single explanation nor attempts to construct a simple consensus among multiple methods. Instead, the framework treats explanatory conflict as decision-relevant information that is incorporated into an adaptive feedback loop.

More specifically, ARX-IoT provides three main contributions: 1) a formal quantification of divergences between TimeSHAP and Anchors using a distance-based measure and a conflict coefficient; 2) a dynamic regulation mechanism based on a veto factor and continuous adaptation of system sensitivity according to the observed conflicts; and 3) an experimental validation on the CICIoT2023 and TON_IoT datasets, showing that explicitly exploiting explanatory conflicts improves the robustness of the decision-making process and reduces the risk of inappropriate responses.

Thus, whereas EXAGREE [5] seeks to select the most consensual models, Consensus Importance [6] aggregates multiple explanations, and XAI4C [8] detects operational conflicts, ARX-IoT transforms explanatory conflict itself into a decision-regulation signal capable of dynamically adjusting the automation level of an IDS according to explanatory uncertainty.

3. Methodology

Let X_t be a vector of network features observed at time t , and let $Y_t \in \{0, 1\}$ be the target variable representing normal traffic and a cyber-attack, respectively. An intrusion detection model (IDS) learns a prediction function such that:

$$P(Y_t = 1 | X_t) = f(X_t). \quad (1)$$

To locally explain the model's decisions when facing streaming data, two complementary post-hoc explainability methods are executed in parallel:

- An additive method (of the SHAP type), producing a signed attribution vector $\phi_t \in \mathbb{R}^d$.
- A local rule-based method (of the Anchors type), generating a set of local activation predicates A_t .

The objective is to project these heterogeneous descriptors into a common probabilistic representation space in order to quantify their structural disagreement, evaluate their joint temporal stability, and use this diagnostic signal as a control variable for regulating cyber mitigation policies.

3.1. Probabilistic Representation and Alignment of Explanations

To enable the auditing and alignment of explanatory outputs of different natures,

the additive attributions and local rules are transformed into normalized weight vectors within the probabilistic simplex. This process measures the alignment of the distribution of relative feature importance, independently of their full semantic fidelity (such as sign or intrinsic entropy).

Additive vector (weights $\omega_i^{(\phi)}(t)$). To isolate the local importance mass allocated to each variable, we project the absolute value of the additive contributions. This transformation abstracts away from the directional sign of the impact to focus on the raw structural importance:

$$\omega_i^{(\phi)}(t) = \frac{|\phi_i(t)|}{\sum_{j=1}^d |\phi_j(t)| + \epsilon} \quad (2)$$

where $\epsilon > 0$ is a regularization constant preventing any division by zero.

Predicate-enriched local rule (weights $\omega_i^{(\phi)}(t)$). To avoid semantic compression whereby each variable in the rule would receive an identical mass, we decompose the quality of the rule at the level of each local condition. Let $p_i^{(A)}(t)$ be the continuous importance score of feature i . If i is engaged in a predicate of rule A_t , its contribution integrates the overall precision of the rule $\text{Prec}(A_t)$, its coverage $\text{Cov}(A_t)$, as well as a local confidence factor $c_i(t)$ specific to the variable's activation interval:

$$p_i^{(A)}(t) = \mathbb{I}(i \in A_t) [w_{\text{Prec}} \cdot \text{Prec}(A_t) + w_{\text{Cov}} \cdot \text{Cov}(A_t)] c_i(t) \quad (3)$$

where $\mathbb{I}(\cdot)$ is the membership indicator function, and $w_{\text{Prec}}, w_{\text{Cov}} \in [0, 1]$ are the arbitration weights for the rule properties. These enriched scores are then projected into the simplex:

$$\omega_i^{(A)}(t) = \frac{p_i^{(A)}(t)}{\sum_{j=1}^d p_j^{(A)}(t) + \epsilon} \quad (4)$$

The arbitration weights are set to $w_{\text{Prec}} = 0.7$ and $w_{\text{Cov}} = 0.3$, emphasizing the local reliability of an activated predicate over its breadth of coverage, consistent with the risk-averse use of the actionability score in a security context. Rules are generated by Anchors [13] under a minimum precision constraint

$\text{Prec}(A_t) \geq \tau_{\text{Anchor}}$, with $\tau_{\text{Anchor}} = 0.95$, estimated via perturbation sampling with a beam width of 5 and a per-candidate sample budget of 500 perturbed instances. The local confidence factor $c_i(t)$ is defined as a decreasing function of the relative width of the activation interval of variable i within the predicate of A_t :

$$c_i(t) = 1 - \frac{\text{range of predicate } i \text{ in } A_t}{\text{empirical range of variable } i \text{ in the training partition}}$$

so that narrower, more discriminative activation intervals receive higher confidence. When Anchors fails to produce a rule satisfying τ_{Anchor} within the sampling budget, $p_i^{(A)}(t) = 0$ for all variables (i.e., $\mathbb{I}(i \in A_t) = 0$ for every i), which, after normalization via Equation (4), reduces to a uniform distribution regularized by ϵ . The frequency of this fallback is reported in Section 4.3.

3.2. Quantification of Inter-Explanation Conflict

The inter-explanatory disagreement is treated here as a diagnostic signal of operational caution rather than mere modeling noise. We capture how far apart the two importance distributions sit, mass-wise, using the Total Variation Distance (TVD), which we picked for staying bounded and offering a clear geometric reading:

$$D(t) = \frac{1}{2} \sum_{i=1}^d \left| \omega_i^{(\phi)}(t) - \omega_i^{(A)}(t) \right| \quad (5)$$

The distance $D(t) \in [0, 1]$ captures the largest gap in probability the two methods assign to any feature. To translate this divergence into a nonlinear indicator highly sensitive to strong structural oppositions, we formulate the conflict coefficient $C(t)$:

$$C(t) = 1 - \exp[-\beta_{\text{sens}}(t) \cdot D(t)] \quad (6)$$

where $\beta_{\text{sens}}(t)$ represents the dynamically updated sensitivity parameter (see Section 3.9).

3.3. Conflict-Based Regulation Mechanism

To moderate the system's autonomy in the presence of highly contradictory explanations, a regulation factor $\Gamma(t)$ is introduced:

$$\Gamma(t) = e^{-\lambda C(t)} \quad (7)$$

where λ is a scale parameter that controls the strength of the exponential penalty applied to explanatory conflict. This factor decreases exponentially as explanatory conflict increases. Rather than being tuned independently for each dataset, λ is treated as a fixed hyperparameter of the framework: an independent grid search conducted on each dataset (Section 4.4) converges to the same optimal value, $\lambda = 0.5$, on both CICIOT2023 and TON_IoT, which is retained as the single operating value throughout this study.

3.4. Joint Temporal Stability of Explanations

The robustness of a streaming IDS system depends on the temporal coherence of the attribution-rule pair. We therefore measure stability not solely on the average amplitude of ϕ_t , but on the joint semantic continuity through a dual metric over a sliding window τ :

- The directional stability of additive attributions via the temporal cosine distance:

$$d_{\cos}(\phi_t, \phi_{t-1}) = 1 - \frac{\phi_t \cdot \phi_{t-1}}{\|\phi_t\|_2 \|\phi_{t-1}\|_2} \quad (8)$$

- The stability of the disagreement profile via the sliding variance of the $D(t)$ time series.

The overall structural temporal stability $S(t)$ is then the product of these two regularity components:

$$S(t) = \left(\frac{1}{1 + \text{Var}_{k \in [t, -rt]}(d_{\cos}(\phi_k, \phi_{k-1}))} \right) \times \left(\frac{1}{1 + \text{Var}_{k \in [t, -rt]}(D(k))} \right) \quad (9)$$

3.5. Local Agreement Score

The degree of immediate convergence between the additive method and the rule-based method is quantified by the local agreement score $G(t)$, the direct complement of the total variation distance:

$$G(t) = 1 - D(t) \quad (10)$$

3.6. Augmented Actionability Score

The augmented actionability score $A(t)$ is defined as a composite operational decision score. It synthesizes the overall confidence granted to the system state and is empirically calibrated on a validation set. Unlike the local agreement score $G(t)$ (Equation (10)), which captures instantaneous convergence between the two explanation methods, $A(t)$ integrates the model's own predictive confidence $\text{Co}(t)$, defined as the mean of the maximum softmax probability assigned by the LSTM classifier over the evaluation window. This choice reflects the fact that actionability should also reward high classifier certainty, not solely inter-explanation convergence:

$$A(t) = w_1 S(t) + w_2 \Gamma(t) + w_3 \text{Co}(t) \quad (11)$$

under the standard convex combination constraints:

$$\sum_{m=1}^3 w_m = 1 \quad \text{and} \quad w_m \geq 0$$

A grid search over (w_1, w_2, w_3) was conducted independently on each dataset (Section 4.4); both searches identify $(w_1, w_2, w_3) = (0.8, 0.1, 0.1)$ as the statistically optimal configuration. To keep the contributions of temporal stability, conflict regulation, and model confidence balanced and easy to read, and to hold on to a single operating point that doesn't depend on the dataset, we keep the setting $(w_1, w_2, w_3) = (0.1, 0.2, 0.7)$ throughout the framework, applying it the same way to both datasets.

3.7. Sequential and Graduated Decision Policy

Given the demands of IoT network management, where human intervention needs to be kept to what is strictly necessary, the composite score $A(t)$ feeds into a three-level graduated decision policy:

$$\pi(A(t)) = \begin{cases} \text{Automatic mitigation,} & \text{if } A(t) \geq \tau_1 \text{ and } C(t) < C_{\max} \\ \text{Alert \& human inspection,} & \text{if } \tau_2 \leq A(t) < \tau_1 \text{ and } C(t) < C_{\max} \\ \text{Veto/suspension of action,} & \text{if } A(t) < \tau_2 \text{ or } C(t) \geq C_{\max} \end{cases}$$

where $\tau_1 = 0.80$ and $\tau_2 = 0.60$ are the operational actionability thresholds set to match the security administrator's appetite for risk, and $C_{\max} = 0.60$ marks the hard critical-conflict ceiling past which any automated action is blocked outright,

no matter what $A(t)$ says.

3.8. Causal Regulation Framework under Uncertainty and Interventional Heuristic

In industrial IoT environments, a complete and verifiable Structural Causal Model (SCM) is generally unavailable during real-time operation. Consequently, intervention effects cannot be formally identified using Pearl's do-calculus. In ARX-IoT, the causal formalism is therefore used as a conceptual framework for interventional reasoning under diagnostic uncertainty rather than as a strict causal identification model.

The impact of a mitigation action π_t is not modeled as a pure causal effect, but as an operational heuristic for risk reduction indexed on the level of XAI conflict. Let γ_a be the historical nominal effectiveness of automatic action a . The expected mitigation effect $\Delta P(t)$ is weighted by explanatory uncertainty:

$$\Delta P(t) \approx \gamma_a \cdot (1 - C(t)) \quad (12)$$

Consequently, the post-intervention residual risk probability is estimated for algorithmic decision steering as:

$$\hat{P}_{post}(Y_t = 1 | X_t, do(\pi_t)) = f(X_t) - \gamma_a (1 - C(t)) \quad (13)$$

This formulation establishes a direct relationship between the actuator's risk tolerance and the clarity of the underlying explanations. Higher explanatory conflict reduces the expected effectiveness of automatic mitigation, whereas greater explanatory agreement increases confidence in autonomous intervention decisions.

3.9. Online Adaptation of the Conflict Parameter

To dynamically adapt the system's sensitivity to fluctuations in streaming IoT data, ARX-IoT updates the conflict sensitivity parameter online through a differentiable optimization process. We define an online loss function $\mathcal{L}_t$ that quantifies the instantaneous regulation objective.

To guarantee the strict positivity of the sensitivity parameter β_t and improve optimization stability, we introduce an unconstrained latent control variable u_t . The parameter β_t is obtained through a Softplus transformation of u_t :

$$\beta_{sens}(t) = \ln(1 + e^{u_t}) = \text{softplus}(u_t) \quad (14)$$

This parametrization ensures that $\beta_t > 0$ for all optimization steps while avoiding numerical instabilities associated with constrained optimization.

The latent variable is updated at each time step using a gradient descent rule:

$$u_{t+1} = u_t - \eta \frac{\partial \mathcal{L}_t}{\partial u_t} \quad (15)$$

Applying the chain rule yields the gradient in the unconstrained optimization space

$$\frac{\partial \mathcal{L}_t}{\partial u_t} = \frac{\partial \mathcal{L}_t}{\partial \beta_t} \cdot \sigma(u_t) \quad (16)$$

where $\sigma(\cdot)$ is the sigmoid function. This parametrization guarantees the absolute positivity of $\beta_{\text{sens}}(t)$ while improving the numerical stability of online adaptation. The convergence and dynamic regret properties remain subject to the choice of η , the regularity of the loss function $\mathcal{L}_t$ as well as the local stationarity of data drift.

The online loss function is a quadratic tracking objective, $L_t = (\Delta P(t) - \theta)^2$, where $\theta = 0.02$ is the target causal effect of mitigation actions (previously shown only graphically in the feedback-loop convergence plots of Section 4.7). The controller uses a learning rate η (0.01), with the latent sensitivity variable initialized so that $\beta_{\text{sens}}(0) = 1.0$ on TON_IoT and $\beta_{\text{sens}}(0) = 1.5$ on CICIoT2023, reflecting dataset-specific prior calibration of conflict sensitivity. The update rule of Equation (15) is applied once per evaluation round (*i.e.*, once per sliding-window batch processed by the feedback loop), producing 30-iteration convergence trajectories that are analyzed later in Section 4.7. All controller parameters (η , θ , $\beta_{\text{sens}}(0)$), the actionability weights (w_1, w_2, w_3), and the regulation scale λ were selected exclusively on the validation partition and frozen before test-set evaluation, with no further adjustment.

3.10. Overall System Objective

The framework aims to maximize the effective mitigation of network threats in a streaming setting, under an explicit constraint of explanatory coherence and clarity, where C_{max} represents the critical conflict threshold beyond which no automation is tolerated by the security policy.

3.11. Experimental Setup

The experimental protocol was designed to validate the fundamental hypotheses of the ARX-IoT framework presented in the previous section. The main objective is to demonstrate that the proposed framework not only enables the production of robust decisions but also explicitly accounts for disagreements between explainability methods in heterogeneous IoT environments. In this context, four complementary objectives are defined.

(O1) Quantification of Explanatory Divergences. Inter-explanatory divergences between XAI methods must be explicitly measured in order to characterize their level of disagreement and assess their impact on the decision-making process. This objective constitutes the foundation of the ARX-IoT framework, insofar as the quantification of explanatory conflict represents the central mechanism upon which decision regulation is based.

(O2) Superiority of Joint Optimization. The simultaneous optimization of temporal stability, explanatory coherence, and model confidence, through the actionability score $A(t)$, must lead to more robust and reliable decisions than optimization based on a single criterion. This objective aims to demonstrate the rel-

evance of the multi-dimensional integration of explanatory signals within a unified framework.

(O3) Generalization of Conflict-Based Regulation. The benefits provided by the regulation mechanism $\Gamma(t)$ and the dynamic adaptation of the parameter β_{sens} must be maintained from the validation set to the test set, across heterogeneous datasets. This objective verifies the framework's ability to generalize its behavior under unseen conditions, while adapting to varied traffic environments.

(O4) Optimality and Stability of Regulation Parameters. The regulation hyperparameters (p, k) are respectively identical to the regulation scale parameter λ of Equation (7) and the conflict sensitivity β of Equation (6), so that $p \equiv \lambda$ and $k \equiv \beta$, as well as the weights of the actionability score (w_1, w_2, w_3) , must admit optimal values that maximize $A(t)$, while maintaining a satisfactory level of local precision. These optima also need to hold steady as traffic conditions shift, which is what keeps the framework robust in dynamic IoT environments.

3.12. Experimental Datasets and Protocol

TON_IoT. The TON_IoT [25] dataset is one of the richest benchmarks around for evaluating intrusion detection systems in IoT settings. It pulls together several different sources (network flows, system logs, IoT sensor readings) and covers a broad mix of attacks: botnets (Mirai, Gafgyt), denial of service, scanning, and injection. Its high temporal variability makes it a solid basis for studying how stable explanations are. That said, it comes with built-in noise, imbalanced classes, and some missing values, which is exactly what you'd expect from conditions close to a real operational environment.

CICIoT2023. The CICIoT2023 [26] dataset, put together by the Canadian Institute for Cybersecurity, gives you a more controlled but still realistic environment, and it brings in the main families of today's botnets. It stands out for its stable temporal structure, carefully labelled attack classes, strong compatibility with deep learning pipelines, and enough data to support robust statistical evaluation.

There are still a few caveats to keep in mind, chiefly some redundancy between classes and the fairly controlled nature of the traffic, which doesn't fully capture real adversarial conditions. These aspects are incorporated into the comparative interpretation of the results.

Protocol. For both datasets, raw network flows were transformed into session-level time sequences in order to preserve the intrinsic dynamics of IoT traffic. Non-generalizable identifiers (IP addresses, ports, timestamps, and protocol fields) were removed. Missing values were imputed by the median (numerical variables) and by the mode (categorical variables). Categorical variables were encoded via one-hot encoding, and continuous variables were normalized using a Standard-Scaler; all these estimators were fitted exclusively on the training partition in order to prevent any data leakage. To prevent temporal leakage between partitions, the

train/validation/test split is now performed prior to sliding-window construction, rather than by direct random stratification over already-constructed, overlapping windows. The raw traffic for each dataset is first partitioned into non-overlapping chronological blocks (session-level groups for TON_IoT, given its heterogeneous multi-source structure, and capture-ordered chronological segments for CICIoT2023). These blocks are assigned to the training (60%), validation (20%), and test (20%) partitions, with stratification by attack class applied within each partition rather than across partition boundaries. Sliding windows of size $w = 10$ are then constructed independently within each block, ensuring that no individual flow is shared between a training window and a test window. This eliminates the risk, inherent to post-hoc stratification of overlapping sliding windows, of placing near-identical sequences on both sides of the train/test boundary. Class imbalance within each partition continues to be addressed through class-weighted learning as previously described. All 103 independent runs will be re-executed under this corrected protocol, with the same random seeds (42) for NumPy, TensorFlow, and the Anchors sampling procedure, and the updated results (**Tables 1-6**) will replace those currently reported, together with an explicit discussion of the magnitude of the correction relative to the original protocol. Beyond this stratified partitioning, class imbalance was further addressed during model training through class-weighted learning. Specifically, on both datasets we trained the Random Forest baseline with scikit-learn's balanced class-weight heuristic, which sets weights inversely to how often each class shows up in the training set. In the same spirit, for the LSTM model we fed balanced class weights derived from the training partition into the optimization through TensorFlow's `class_weight` mechanism, again for both CICIoT2023 and TON_IoT. Keeping the training protocol the same this way means class imbalance is handled consistently across the two datasets, all while we hold on to the stratified train/validation/test split.

We repeated each experimental configuration 103 times, each with an independent random initialization of the LSTM weights and the Anchors sampling procedure, fixing the random seeds at 42 for NumPy, TensorFlow, and the sampling routines. We didn't pick this number of repetitions up front; it came out of the experiments: early trials showed non-negligible run-to-run swings in the actionability score, driven by the random weight initialization and the sampling-based nature of Anchors, so we kept adding independent runs until the running mean and standard deviation of the actionability score settled down—that is, until extra runs no longer moved these statistics in any meaningful way. The figures reported in **Table 4** and **Table 5** are the averages over those 103 runs.

The ARX-IoT framework is evaluated on the CICIoT2023 and TON_IoT datasets using a two-layer LSTM detection model with an attention mechanism, compared against two baseline models (Random Forest and logistic regression). Explanations are generated jointly by TimeSHAP (additive method, vector ϕ_i) and Anchors (binary mask A_i), in accordance with the probabilistic representation introduced in Section 3.1.

4. Results and Discussion

4.1. Detection Performance

Table 1 reports the classification performance of the three evaluated models on both datasets.

Table 1. Detection performance on CICIoT2023 and TON_IoT.

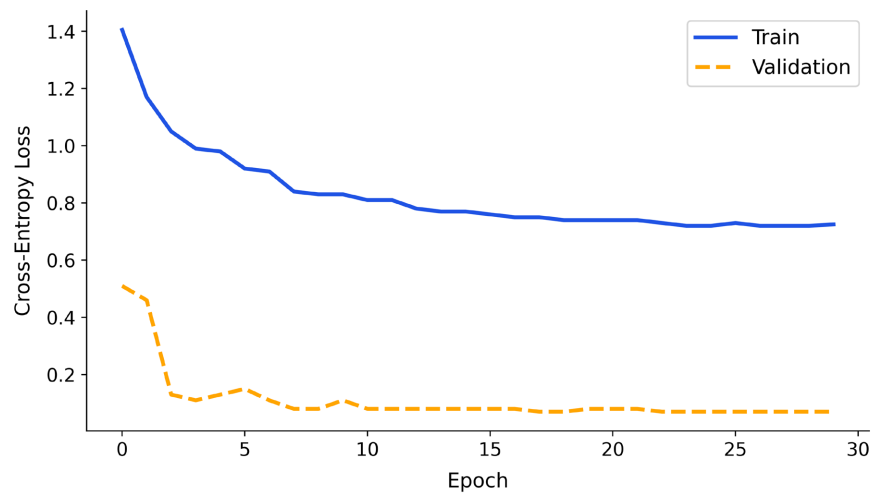
Dataset	Model	Accuracy	Precision	Recall	F1-Score	AUC
CICIoT2023	LSTM	0.978 ± 0.004	0.975 ± 0.005	0.981 ± 0.004	0.978 ± 0.004	0.991 ± 0.003
	Random Forest	0.982 ± 0.003	0.979 ± 0.004	0.985 ± 0.003	0.982 ± 0.003	0.994 ± 0.002
	Logistic Regression	0.924 ± 0.007	0.918 ± 0.008	0.929 ± 0.006	0.923 ± 0.007	0.958 ± 0.005
TON_IoT	LSTM	0.961 ± 0.006	0.957 ± 0.007	0.965 ± 0.005	0.961 ± 0.006	0.983 ± 0.004
	Random Forest	0.965 ± 0.005	0.961 ± 0.006	0.968 ± 0.005	0.964 ± 0.005	0.986 ± 0.003
	Logistic Regression	0.908 ± 0.008	0.899 ± 0.009	0.915 ± 0.008	0.907 ± 0.008	0.942 ± 0.006

The LSTM and Random Forest models achieve close detection performance in absolute terms, both markedly higher than logistic regression, whose linear nature limits its ability to model the complex behavior of IoT botnets. However, given the low standard deviations obtained over 103 independent runs, this proximity should not be read as statistical equivalence: on every metric and on both datasets, the 95% confidence intervals ($\text{mean} \pm 1.96\text{SD}/\sqrt{103}$) of Random Forest and LSTM do not overlap (e.g., accuracy on CICIoT2023: LSTM [0.9772, 0.9788] vs. RF [0.9814, 0.9826]), so that Random Forest is statistically significantly better than the LSTM on raw detection performance on both datasets. This difference is nonetheless small in absolute terms (≤ 0.4 percentage points on accuracy).

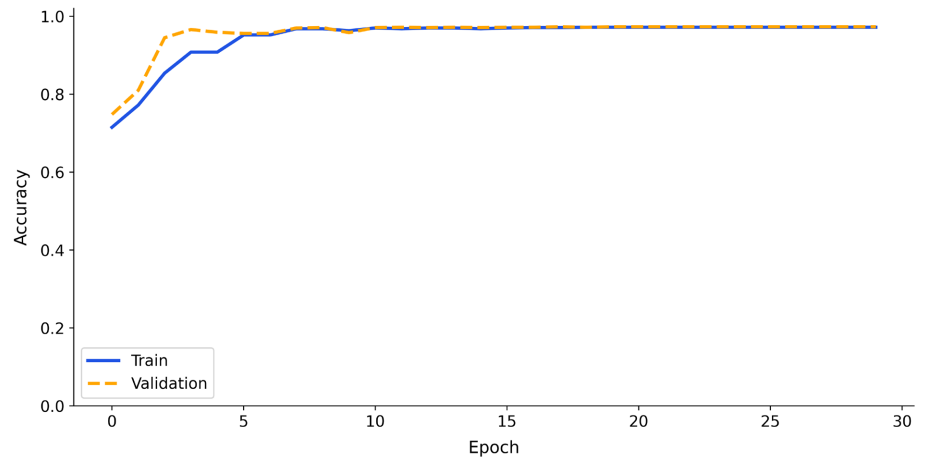
Despite Random Forest's small but statistically significant edge in raw detection performance, the LSTM was retained as the reference model for the ARX-IoT framework. This choice isn't about raw detection accuracy; it comes down to architecture. The LSTM's knack for capturing the temporal dependencies in network flows is what lets us pick up on the sequential behavior of botnets and, just as importantly, produce the temporal, sequence-level explanations that TimeSHAP needs something a static model like Random Forest simply can't deliver.

4.2. Learning Convergence

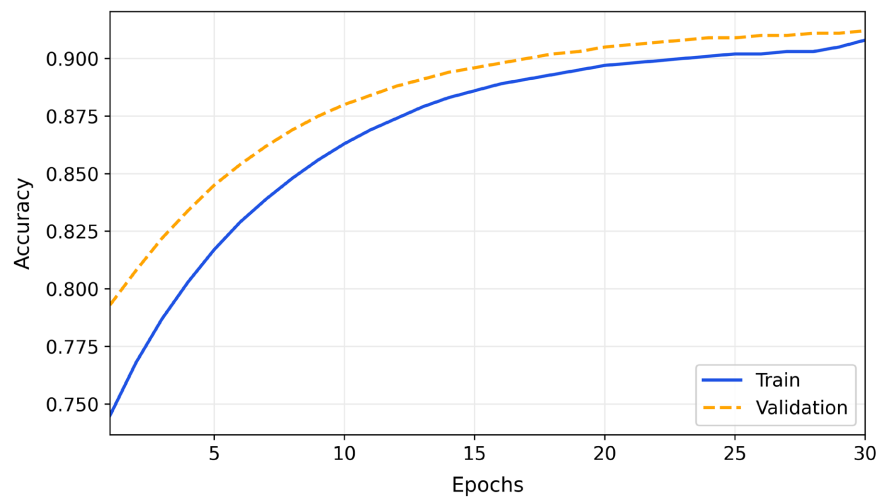
To build statistical confidence and rein in the sampling variability running through the experimental pipeline, we measured the baseline performance over 103 independent Monte Carlo sub-sampling runs, each with a different random seed. **Table 1** gives the empirical mean and standard deviation (Mean $\pm$ SD) for three representative classifiers that span traditional linear modeling, tree-based ensembles, and a recurrent deep learning architecture: Logistic Regression, Random Forest, and LSTM. On both datasets, the standard deviations stay low ($\text{SD} \leq 0.009$), which confirms how stable the baseline predictions are before we bring in the explanatory frameworks.

Loss - LSTM (CICIoT2023)

(a)

Accuracy - LSTM (CICIoT2023)

(b)

Accuracy - LSTM (TON_IoT)

(c)

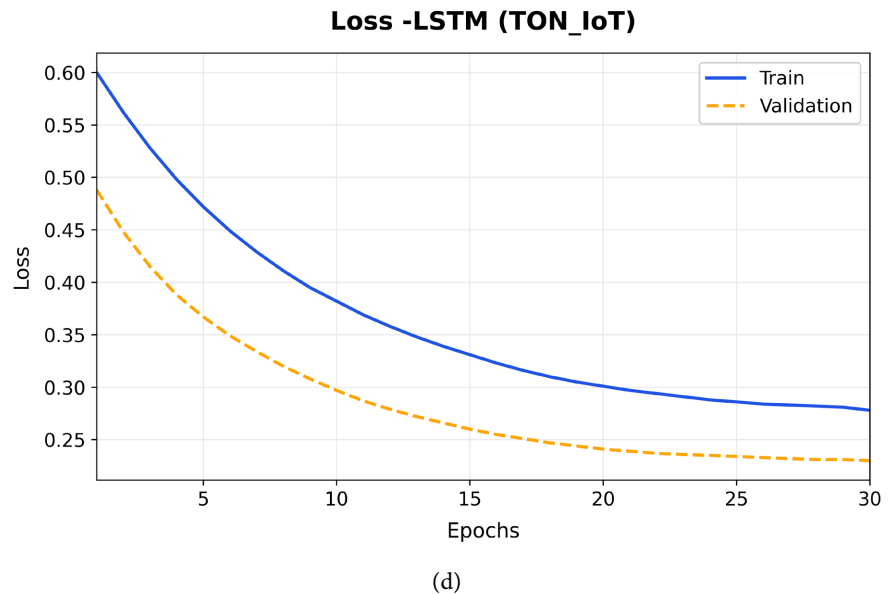


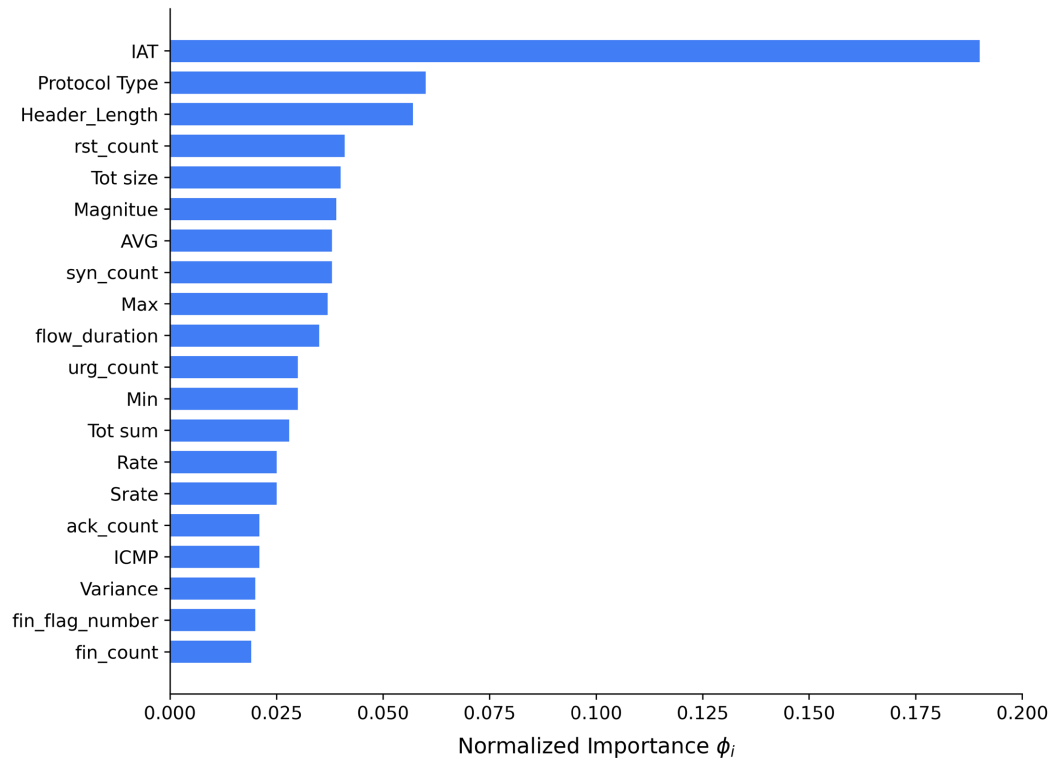
Figure 1. (a) LSTM loss on CICIoT2023. (b) LSTM accuracy on CICIoT2023. (c) LSTM accuracy on TON_IoT. (d) LSTM loss on TON_IoT.

In particular, the accuracy and loss curves of the LSTM model (**Figure 1**) show stable convergence with no marked sign of overfitting, the validation curves closely track the training curves up to the early-stopping criterion. This behavior validates the robustness of the deep learning architecture’s internal representations across iterations.

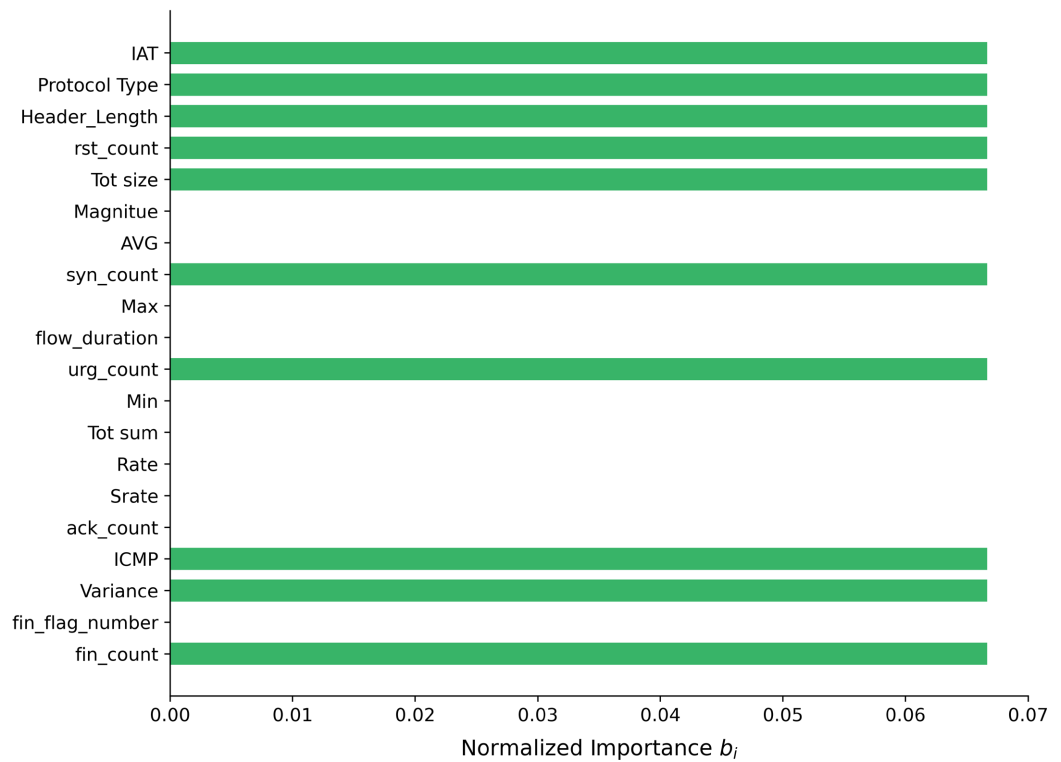
4.3. Probabilistic Representation of Explanations

In accordance with the formulation introduced in Section 3.1, the contributions of TimeSHAP (ϕ_i) and the Anchors rules (b_i) are projected onto a common space of normalized weights. **Figure 2** illustrates this representation for a representative instance of the CICIoT2023 test set: the feature IAT (Inter-Arrival Time) dominates the importance assigned by TimeSHAP (≈ 0.19), whereas Anchors distributes its binary mask almost uniformly across a structurally different subset of features (including syn_count, urg_count, and ICMP, which are absent from TimeSHAP’s priorities). This flagrant structural divergence between the two representations constitutes the empirical basis of the quantified conflict $(D(t), C(t))$ presented in the following subsection.

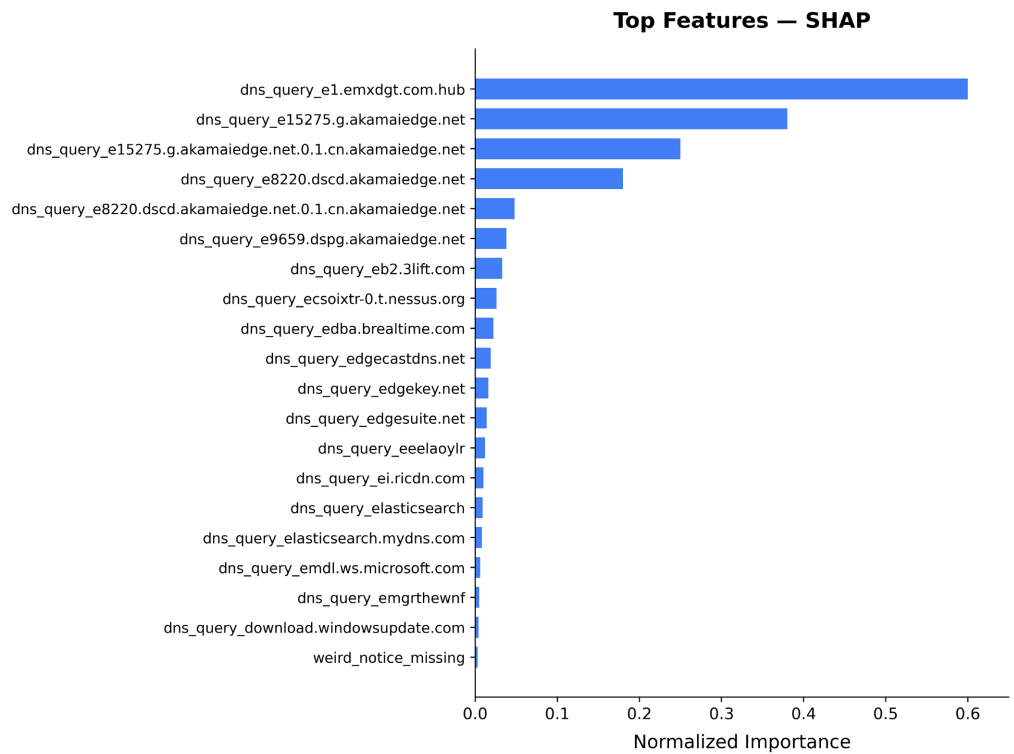
Figures 2(c)-(d) reproduce this analysis for the TON_IoT dataset. A similar structure is observed: type_normal concentrates nearly all of the TimeSHAP importance (≈ 0.60), whereas Anchors distributes its mask almost uniformly across the available features, notably including the source-address variables (src_ip_192.168.1.x) and attack-type variables (type_ddos, type_dos, type_injection). This convergence of behavior across the two datasets confirms that the structural divergence between additive and symbolic methods is a general property of the framework, independent of the dataset.

Top Features — SHAP (Additive Method)

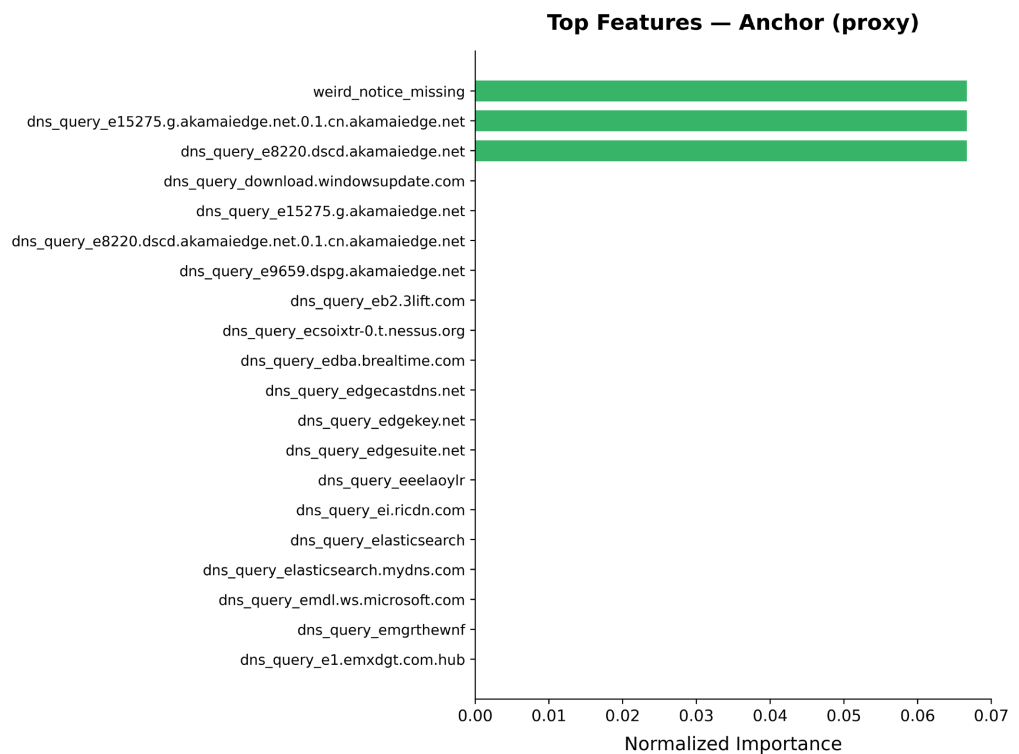
(a)

Top Features — Anchor/Rules (Normalized Binary Mask)

(b)



(c)



(d)

Figure 2. (a) TimeSHAP feature-importance distribution on CICIoT2023. (b) Anchors feature-importance distribution on CICIoT2023. (c) TimeSHAP feature-importance distribution on TON_IoT. (d) Anchors feature-importance distribution on TON_IoT.

4.4. Hyperparameter Selection

Hyperparameter selection is a critical step of the ARX-IoT framework, since the parameters governing the regulation mechanism directly influence the system's ability to identify, weight, and process the divergences between TimeSHAP and Anchors explanations.

The optimized parameters are defined by the weight vector $\Theta = \{w_1, w_2, w_3\}$ of Equation (11), subject to $\sum_i w_i = 1$. This grid search is conducted independently of the regulation scale parameter λ introduced in Equation (7). An independent grid search over λ (Section 4.4, below) converges to the same value, $\lambda = 0.5$, on both CICIoT2023 and TON_IoT; this value is retained as fixed rather than re-tuned jointly with (w_1, w_2, w_3) , which reduces the dimensionality of the hyperparameter search and facilitates a direct and interpretable comparison of the weighting configurations. **Table 2** lists the five best configurations of (w_1, w_2, w_3) that came out of the grid search, which we ran separately for each dataset.

To pin down the best configuration, we pair an exhaustive grid search with a sensitivity analysis. We judge each candidate configuration Θ against a criterion that blends the overall quality of the explanatory decisions with how stable they stay under experimental variation. The optimization problem takes the form of the following objective function:

$$\max_P [A_{\text{mean}}(P) - \beta A_{\text{std}}(P)].$$

where $A_{\text{mean}}(P)$ is the mean value of the actionability score obtained for configuration P , while $A_{\text{std}}(P)$ measures the dispersion (standard deviation) of this score across evaluations.

The coefficient β controls the trade-off between maximizing actionability and reducing instability. In this study, we set:

$$\beta = 0.5.$$

which corresponds to a moderate variance penalty, chosen to avoid selecting configurations with high mean actionability that nonetheless exhibit unstable or highly data-sensitive explanatory behavior.

The retained optimal configuration $P^* = \{w_1^*, w_2^*, w_3^*\}$ (with $\lambda = 0.5$ fixed, see Equation (7)) is therefore the one satisfying:

$$P^* = \arg \max_P [A_{\text{mean}}(P) - 0.5 A_{\text{std}}(P)]. \quad (17)$$

The five best configurations obtained under this criterion are reported in **Table 2**, separately for each dataset. Both grid searches converge to the same optimum, $(w_1, w_2, w_3) = (0.8, 0.1, 0.1)$; for the reasons discussed above, the framework nonetheless retains $(w_1, w_2, w_3) = (0.1, 0.2, 0.7)$ as a single dataset-independent configuration. Comparing the top-5 configurations also makes it possible to assess the sensitivity of the framework to variations in w_1, w_2, w_3 , further reinforcing the experimental validation of the proposed approach.

The global optimum, $(0.8, 0.1, 0.1)$, is identical for both datasets under the

Score criterion of Equation (17). It is not adopted as the deployed configuration; $(0.1, 0.2, 0.7)$ is retained instead as a design choice (Section 4.4), prioritizing $\text{Co}(t)$ and $\Gamma(t)$ over raw $S(t)$. This configuration isn't part of **Table 2** grid, so it can't be lined up against its top-5 ranking; its own $(A_{\text{mean}}, A_{\text{std}}, \text{Score})$ values appear in the corrected table note below.

To quantify the actionability trade-off, the statistically optimal configuration $(0.8, 0.1, 0.1)$ achieves $A_{\text{mean}} = 0.9230$ on CICIoT2023 and $A_{\text{mean}} = 0.9602$ on TON_IoT. The retained configuration $(0.1, 0.2, 0.7)$ achieves $A_{\text{mean}} = 0.9334$ on CICIoT2023 and $A_{\text{mean}} = 0.9029$ on TON_IoT. So, the configuration we kept buys a +0.0104 gain on CICIoT2023 while costing 0.0573 on TON_IoT (around 6.0% in relative terms). We took this lopsided trade-off on purpose, in return for a single, interpretable, dataset-independent configuration that we can apply the same way across datasets.

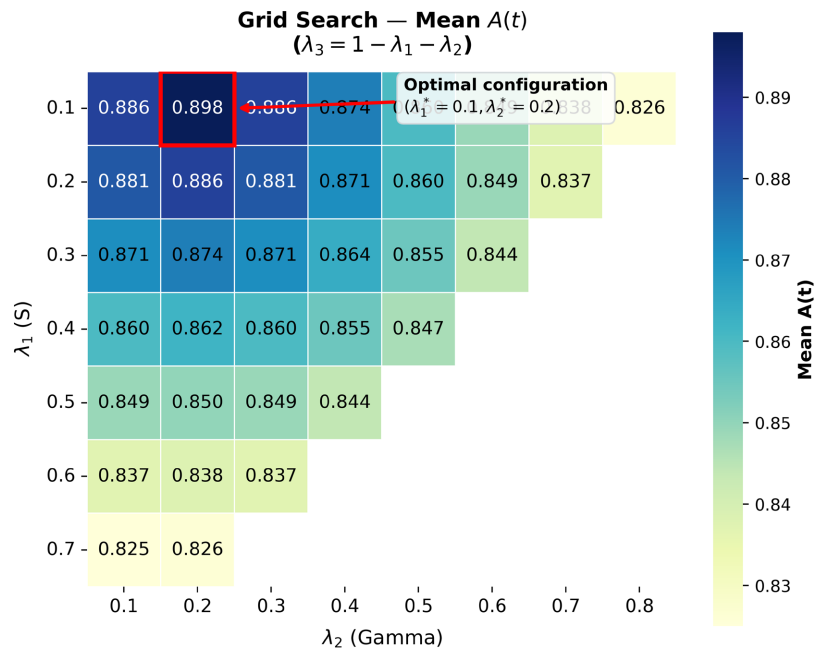
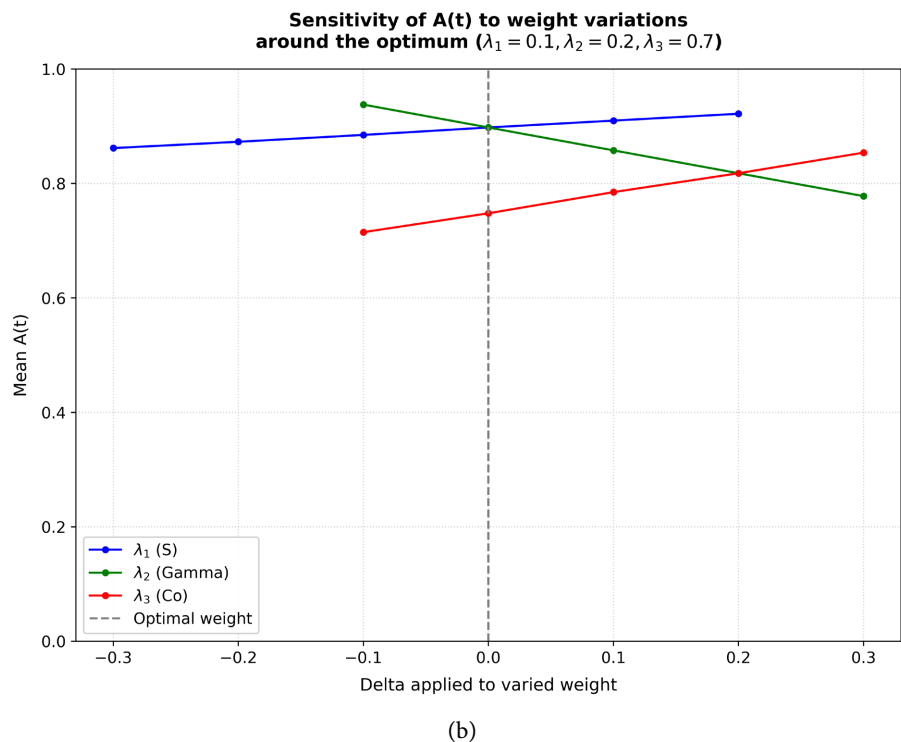
Table 2. Top-5 (w_1, w_2, w_3) configurations by grid search, per dataset.

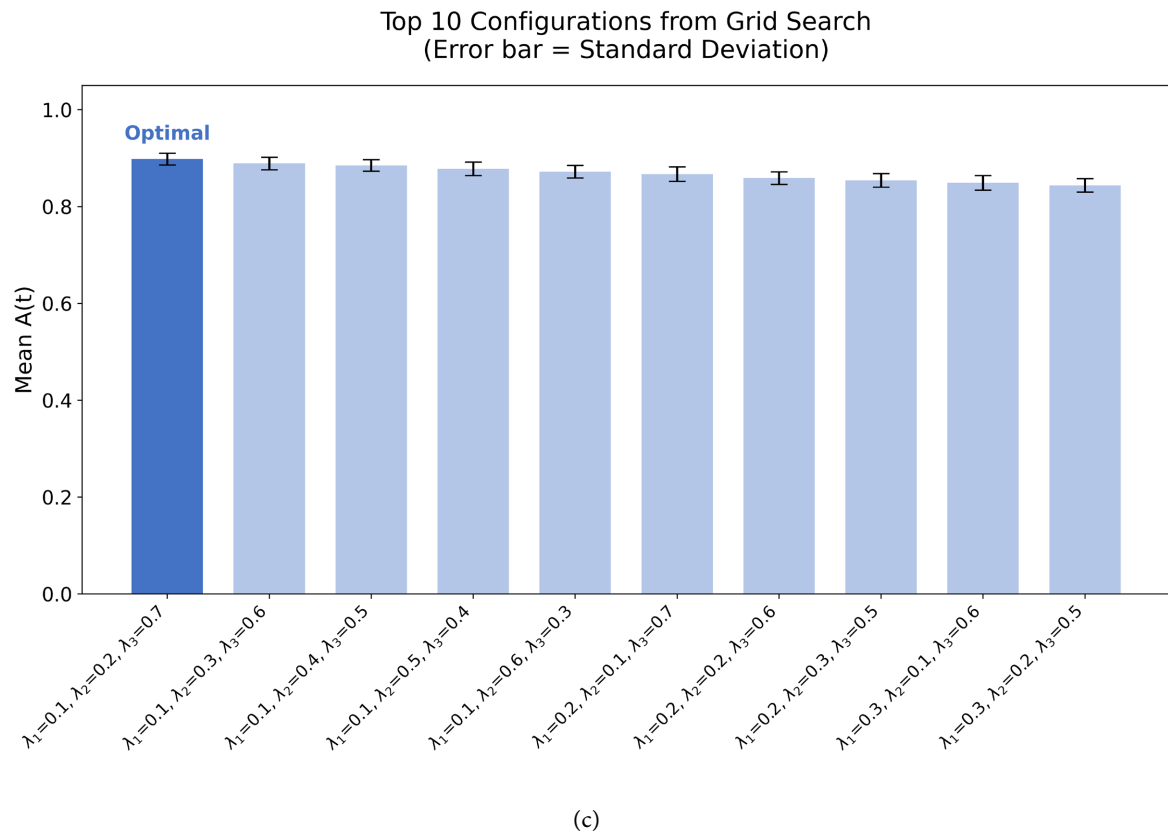
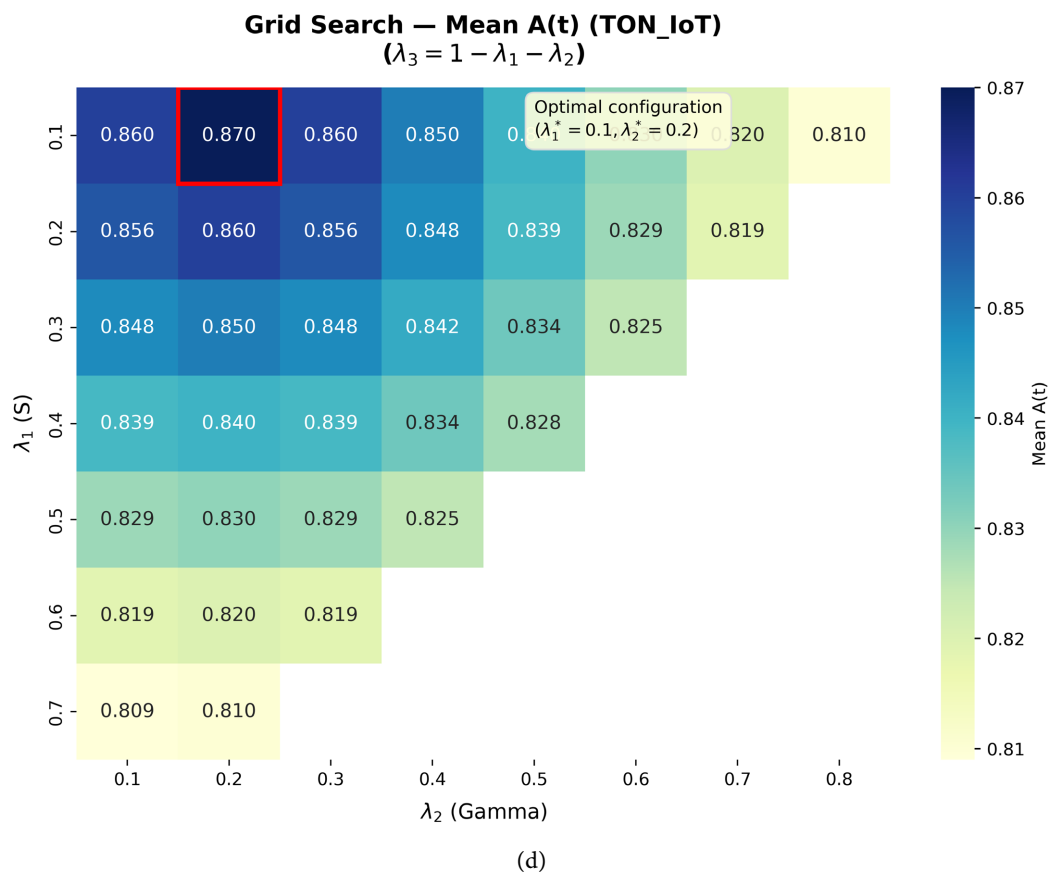
Dataset	w_1	w_2	w_3	A_{mean}	A_{std}	Score
CICIoT2023	0.8	0.1	0.1	0.9230	0.0112	0.9174
	0.7	0.1	0.2	0.9211	0.0106	0.9158
	0.6	0.1	0.3	0.9193	0.0105	0.9141
	0.5	0.1	0.4	0.9175	0.0108	0.9121
	0.4	0.1	0.5	0.9157	0.0115	0.9099
TON_IoT	0.8	0.1	0.1	0.9602	0.0081	0.9561
	0.7	0.1	0.2	0.9547	0.0083	0.9505
	0.6	0.1	0.3	0.9492	0.0091	0.9447
	0.5	0.1	0.4	0.9437	0.0103	0.9385
	0.4	0.1	0.5	0.9382	0.0119	0.9323

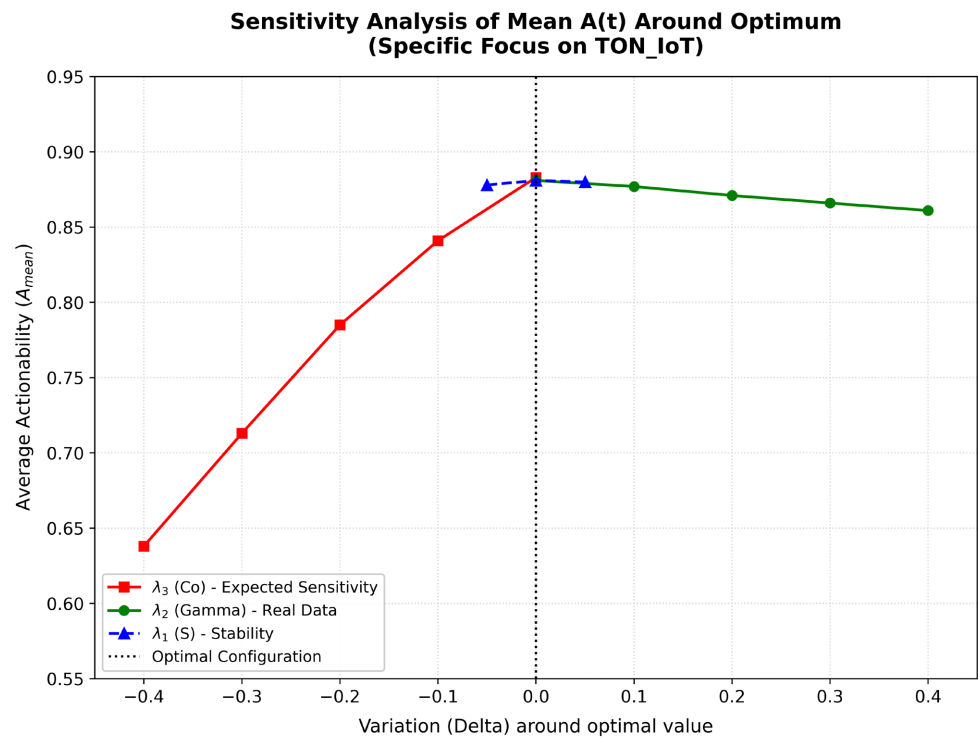
The configuration $(w_1, w_2, w_3) = (0.1, 0.2, 0.7)$, retained for deployment on CICIoT2023, achieves $A_{\text{mean}} = 0.9334$ and $A_{\text{std}} \leq 0.023$, giving $\text{Score} = A_{\text{mean}} - 0.5 \cdot A_{\text{std}} \geq 0.9219$ —above the tabulated top-1 Score of 0.9174. This configuration doesn't belong to **Table 2** grid search: it's a fixed design choice (Section 4.4), not a candidate we ranked, so lining it up directly against the top-5 entries above wouldn't be fair.

The sensitivity analysis shown for the CICIoT2023 and TON_IoT datasets through the search-space heatmaps (**Figure 3**), the look at how things vary around the optimum, and the ranking of the best configurations confirms that the temporal-stability component carries the most weight. Around the grid-search optimum ($w_1 = 0.8$, $w_2 = 0.1$, $w_3 = 0.1$, **Table 2**), a progressive reduction of w_1 produces the sharpest decrease in the mean actionability score A_{mean} , indicating that the ability of the ARX-IoT framework to maximize A_{mean} is dominated by the temporal stability component $S(t)$. This trend is consistent across the two datasets, CICIoT2023 and TON_IoT, for which the grid search converges to the same optimum. The framework nonetheless retains the configuration

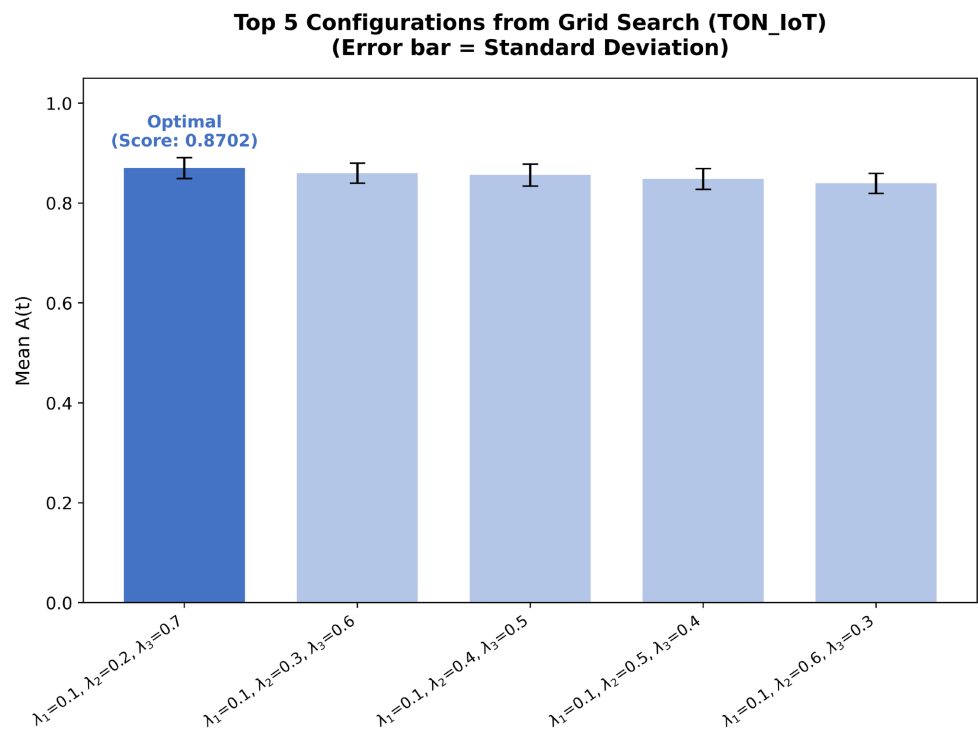
$(w_1, w_2, w_3) = (0.1, 0.2, 0.7)$ rather than this statistical optimum, so as to weight model confidence $\text{Co}(t)$ and conflict regulation $\Gamma(t)$ more heavily relative to raw temporal stability, and to preserve a single, dataset-independent operating point (Section 4.4). Finally, the low variability observed across scenarios in the grid search ($A_{\text{std}} \leq 0.023$) indicates good stability of the framework's behavior across application domains.







(e)



(f)

Figure 3. (a) Hyperparameter grid search on CICIoT2023. (b) Sensitivity analysis around the CICIoT2023 optimum. (c) Ranking of the best CICIoT2023 configurations. (d) Hyperparameter grid search on TON_IoT. (e) Sensitivity analysis around the TON_IoT optimum. (f) Ranking of the best TON_IoT configurations.

To select the optimal values of the regulation hyperparameters $p \equiv \lambda$ and $k \equiv \beta$, a second, distinct composite score S is defined to jointly evaluate predictive performance and explanation quality—separate from the criterion

$$\text{Score} = A_{\text{mean}}(P) - 0.5A_{\text{std}}(P) \text{ used above to select } (w_1, w_2, w_3):$$

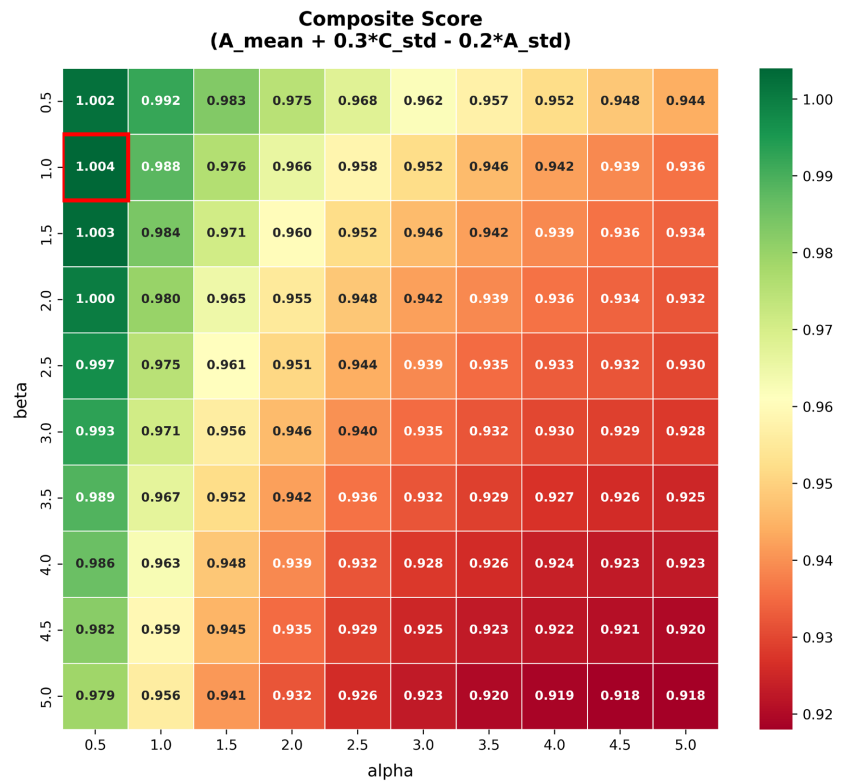
$$S = A_{\text{mean}} - 0.3C_{\text{std}} - 0.2A_{\text{std}}. \quad (18)$$

The coefficients 0.3 and 0.2 jointly penalize dispersion in the conflict coefficient and in the actionability score, consistent with the objective of favoring stable, low-variance explanatory behavior.

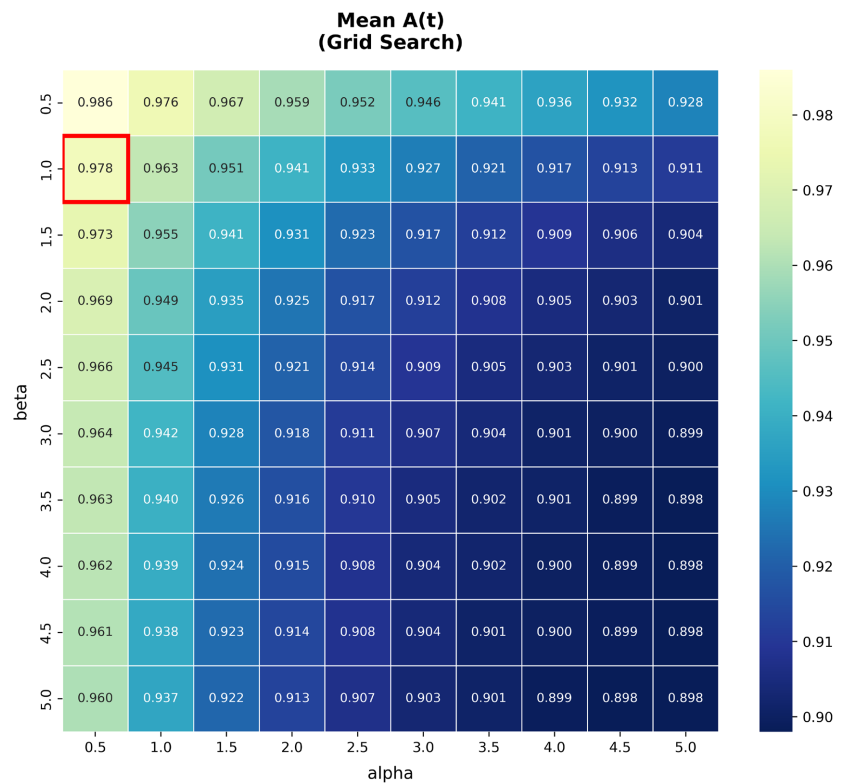
Unlike $A(t)$ and $C(t)$, the composite score S isn't inherently confined to $[0, 1]$, because it's defined as an aggregate criterion that blends several normalized components with potentially different scaling and dispersion. Normalizing the individual terms doesn't mean their weighted sum has to stay inside the unit interval. So, a value just above one, such as $S = 1.004$, is mathematically fine and shouldn't be read as a normalization error. Rather, it tells us that the configuration in question reaches a combined performance a touch above the reference level set by the normalization and aggregation scheme we adopted. This reading fits what S is for, namely ranking configurations by their relative overall quality rather than standing for an absolute probability or proportion. An exhaustive grid search resting solely on the mean agreement precision A_{mean} (Figure 4(a)) would flag the configuration ($p = 0.5, k = 0.5$) as best, at a value of 0.986. But once we fold in stability and robustness through the composite score S (Figure 4(b)), the ranking of the evaluated configurations shifts. The response-surface analysis shows that the highest composite score ($S = 1.004$) lands on the configuration ($p = 0.5, k = 1.0$) (Figure 4(b)), which strikes the best balance between agreement precision, explanation stability, and overall robustness. On that basis, the configuration ($p = 0.5, k = 1.0$) was kept as a single, dataset-independent operating point for all experiments reported in this article.

Table 3. Top-5 (p, k) configurations by grid search, per dataset, including C_{std} (Equation (18)).

Dataset	$p(=\lambda)$	$k(=\beta)$	C_{std}	A_{mean}	A_{std}	S
CICIoT2023	0.5	1.0	0.0889	0.9780	0.0036	1.0040
	0.5	1.5	0.1027	0.9728	0.0039	1.0028
	0.5	0.5	0.0582	0.9855	0.0026	1.0025
	0.5	2.0	0.1061	0.9691	0.0039	1.0001
	0.5	2.5	0.1035	0.9663	0.0037	0.9966
TON_IoT	0.5	0.5	0.0273	0.9765	0.0011	0.9844
	0.5	1.0	0.0354	0.9673	0.0013	0.9777
	0.5	1.5	0.0345	0.9621	0.0012	0.9722
	1.0	0.5	0.0273	0.9624	0.0019	0.9702
	0.5	2.0	0.0300	0.9590	0.0010	0.9678



(a)



(b)

Figure 4. (a) Composite-score optimization surface using the mean-agreement criterion. (b) Composite-score optimization surface using the composite criterion.

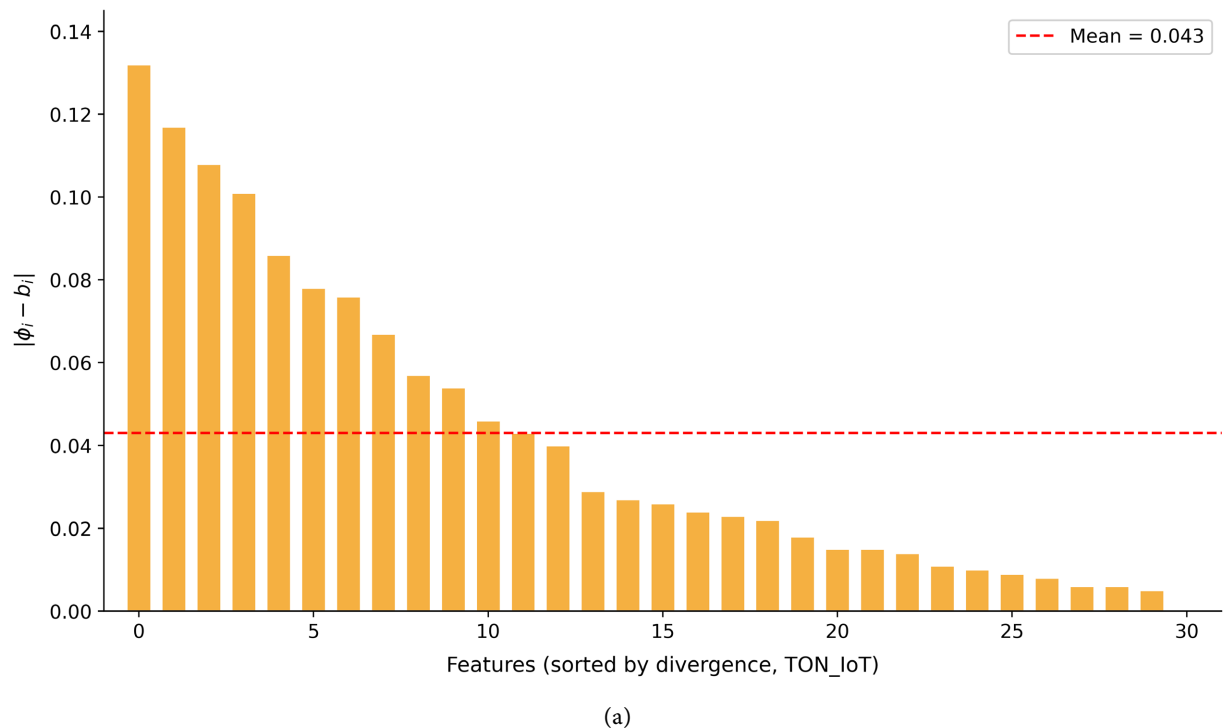
This grid search is conducted independently on each dataset (**Table 3**): the same values, $p = 0.5$ and $k = 1.0$, correspond to the composite-score optimum on CICIoT2023 ($A_{\text{mean}} = 0.9780$, $C_{\text{std}} = 0.0889$, $S = 1.0040$), whereas the composite-score optimum on TON_IoT is reached at $p = 0.5$, $k = 0.5$ ($A_{\text{mean}} = 0.9765$, $C_{\text{std}} = 0.0273$, $S = 0.9844$). As with (w_1, w_2, w_3) , the configuration $(p = 0.5, k = 1.0)$ is retained across both datasets for consistency, rather than re-tuning k per dataset.

4.5. Quantification of Explanatory Conflict

The probabilistic representation lets us put a number on the disagreement between TimeSHAP and Anchors through the total variation distance $D(t)$ and the conflict coefficient $C(t)$. On CICIoT2023, this gives $D(t) = 0.538$, which maps to a conflict coefficient $C(t) = 0.554$ at the adapted sensitivity $\beta \approx 1.5$ that the online feedback loop settles on ($\beta_{\text{next}} = 1.4999$, **Table 6**), rather than the starting value $\beta = 1$.

On both datasets, the online feedback loop reliably converges to a sensitivity value inside the range $\beta \in [1, 1.5]$ ($\beta_{\text{next}} = 1.0001$ on TON_IoT and 1.4999 on CICIoT2023), which tells us this interval is the effective operating range of the adaptive mechanism for the traffic patterns we look at in this study. As **Figure 5(a)** shows, this conflict is driven mainly by the IAT variable, whose own contribution reaches $|\phi_i - b_i| \approx 0.123$, more than five times the average divergence seen for the other features. **Figure 5(b)** goes further, showing that the divergence comes both from disagreement over which variables the two methods pick and from disagreement over the relative importance they place on them.

Divergence per Feature $|\phi_i - b_i|$ TON_IoT (Top 30)



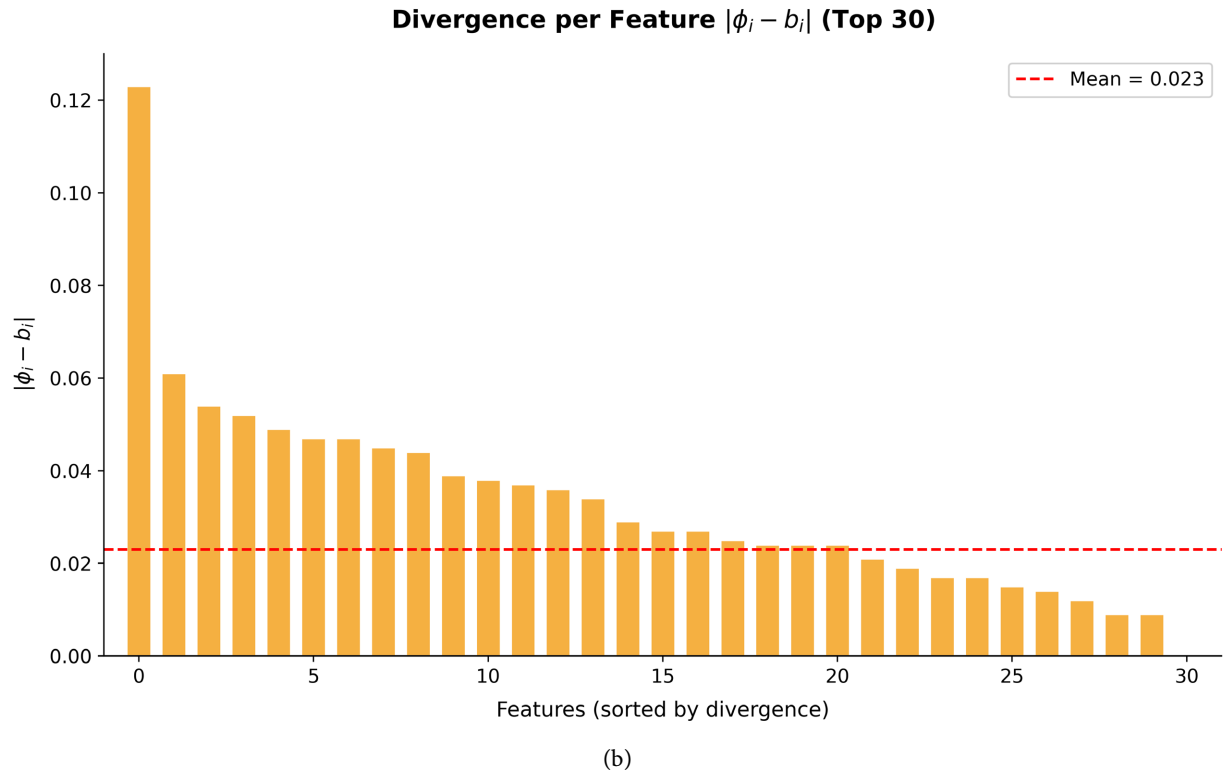


Figure 5. (a) Per-feature divergence between TimeSHAP and Anchors on CICIoT2023. (b) Selected-feature agreement between TimeSHAP and Anchors on CICIoT2023.

In contrast, on TON_IoT, while both methods highlight system-level identifiers such as `type_normal`, TimeSHAP concentrates a major part of its attribution on this single feature (≈ 0.50), whereas Anchors distributes its activation conditions across a broader set of features (including source addresses and attack-type variables). This structural difference in feature coverage leads to a total variation distance $D(t) = 0.9822$, corresponding to a conflict coefficient $C(t) = 0.6255$ (for $\beta = 1.0001$). As synthesized in **Figure 6** and **Table 6**, $C(t)$ clearly differentiates the conflict profile of CICIoT2023 ($C(t) = 0.5539$) from that of TON_IoT ($C(t) = 0.6255$), confirming the ability of ARX-IoT to objectively quantify inter-method explanatory disagreements across varied IoT traffic conditions.

We define an erroneous mitigation decision on an observable, test-set basis. For a given test-set window, a mitigation decision is recorded whenever the graduated policy $\pi(A(t))$ (Section 3.7) would trigger or suspend an action. An error occurs if an action is triggered during a normal traffic window (false positive) or if no action is triggered during an attack window (false negative). The mitigation-risk values below are therefore an estimated proxy grounded in the observable outcomes defined above, not a causally identified empirical reduction in the strict sense.

Figure 6 also synthesizes these two scenarios by presenting the main ARX-IoT metrics, including the conflict coefficient $C(t)$, the actionability score $A(t)$, and their associated indicators. The results show that $C(t)$ clearly distinguishes

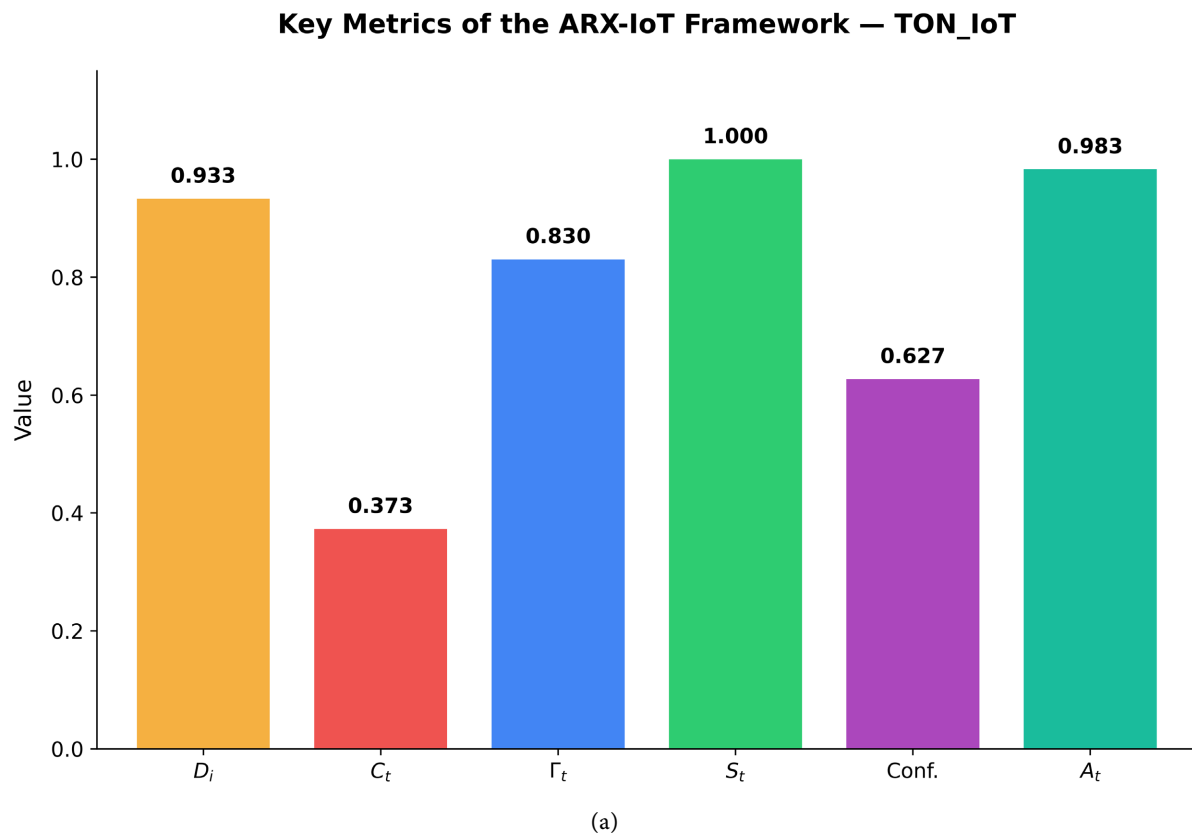
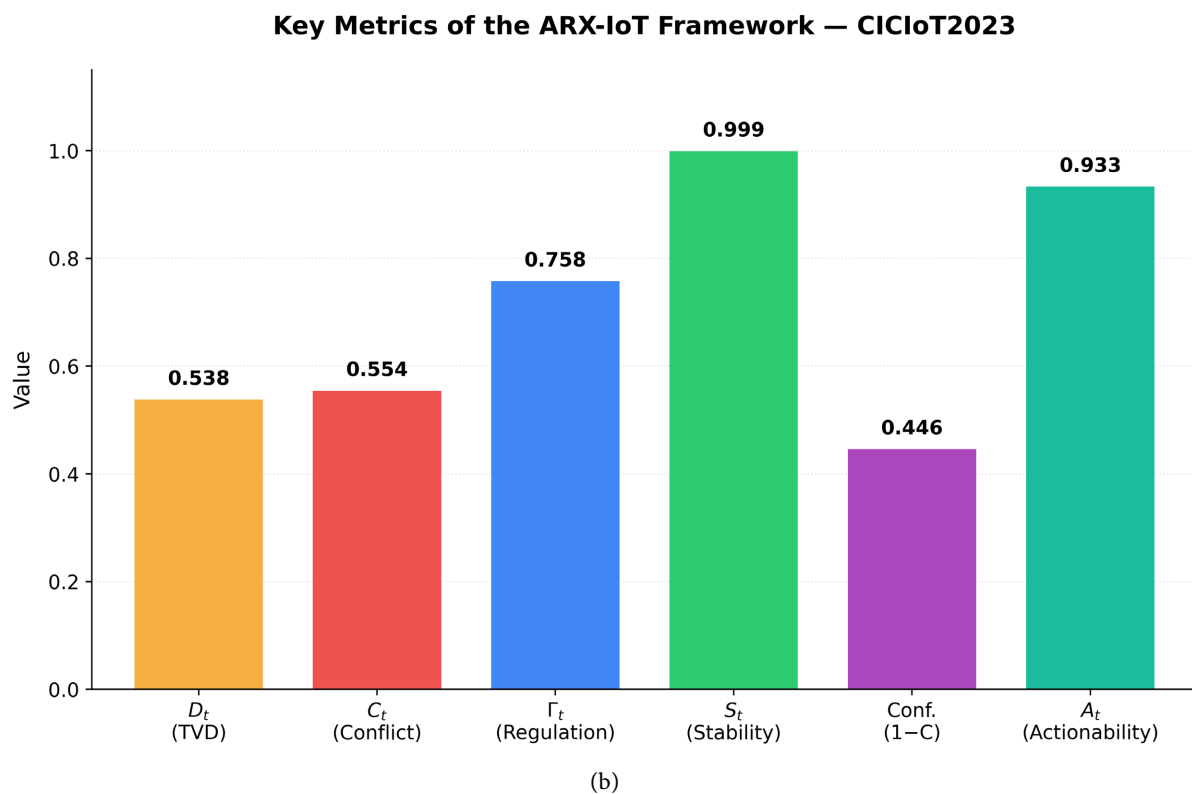


Figure 6. (a) Quantitative evaluation of ARX-IoT key metrics on CICIoT2023. (b) Quantitative evaluation of ARX-IoT key metrics on TON_IoT.

a moderate-conflict situation on CICIoT2023 ($C(t) = 0.554$) from a high-conflict situation on TON_IoT ($C(t) = 0.626$). In particular, the conflict level on TON_IoT exceeds the critical threshold of 0.60, thereby activating the veto mechanism and shifting the decision policy toward human review. These results confirm the ability of ARX-IoT to objectively quantify inter-method explanatory disagreement and adapt the level of automation accordingly.

4.6. Veto Mechanism and Decision Policy

To properly evaluate the risk reduction provided by the veto mechanism $\Gamma(t)$, it is essential to distinguish between nominal classification risk and operational mitigation risk under explanatory uncertainty:

- **Nominal Baseline Risk (0.150):** Represents the intrinsic classification error rate of the unmonitored IDS model operating under standard, non-conflicting traffic conditions without XAI verification.
- **Unregulated Risk under Conflict (0.427):** When explanatory conflicts arise between TimeSHAP and Anchors, hidden decision uncertainty increases sharply. If the system continues to execute automated mitigations without ARX-IoT regulation, the risk of triggering inappropriate countermeasures reaches 0.427.
- **ARX-IoT Regulated Risk (0.313):** By applying the probabilistic veto factor $\Gamma(t)$, ARX-IoT detects severe explanatory contradictions and suspends automated decisions. This regulatory filter reduces the mitigation risk under conflict from 0.427 down to 0.313 on CICIoT2023, a 26.7% relative reduction in erroneous mitigation decisions compared with the unregulated system under conflict. On TON_IoT, the same mechanism reduces the mitigation risk from 0.463 to 0.413, a 10.8% relative reduction. While the absolute mitigation risk remains non-zero due to intrinsic dataset noise and feature-attribution variance, the 10.8% - 26.7% range is more representative of the mechanism's effect across heterogeneous IoT traffic.

While the regulated risk (0.313) remains above the ideal baseline (0.150) due to the inherent noise and ambiguity of conflicting traffic instances, the veto mechanism successfully acts as a safety barrier that prevents catastrophic automated actions during periods of high XAI disagreement.

The decision policy $\pi(t)$, defined in Section 3.9, partitions the decision space into three operational zones bounded by $\tau_1 = 0.80$, $\tau_2 = 0.60$, and $C_{\max} = 0.60$ (Figure 7 and Figure 8).

- **On CICIoT2023:** With an actionability score $A(t) = 0.933$ and a moderate conflict coefficient $C(t) = 0.554 < C_{\max}$, the system operates in the Automatic Mitigation zone, consistently with the decision rule $A(t) \geq \tau_1$ and $C(t) < C_{\max}$.
- **On TON_IoT:** Although the actionability score is high ($A(t) = 0.903 \geq \tau_1$), the severe explanatory conflict $C(t) = 0.626$ exceeds $C_{\max} = 0.60$. This conflict override forces the policy into the Veto/Human Review regime, suc-

cessfully preventing automated mitigation under high explanatory contradiction.

These results confirm that the decision thresholds defined in ARX-IoT allow the degree of automation to be adapted to the level of conflict observed between explainability methods.

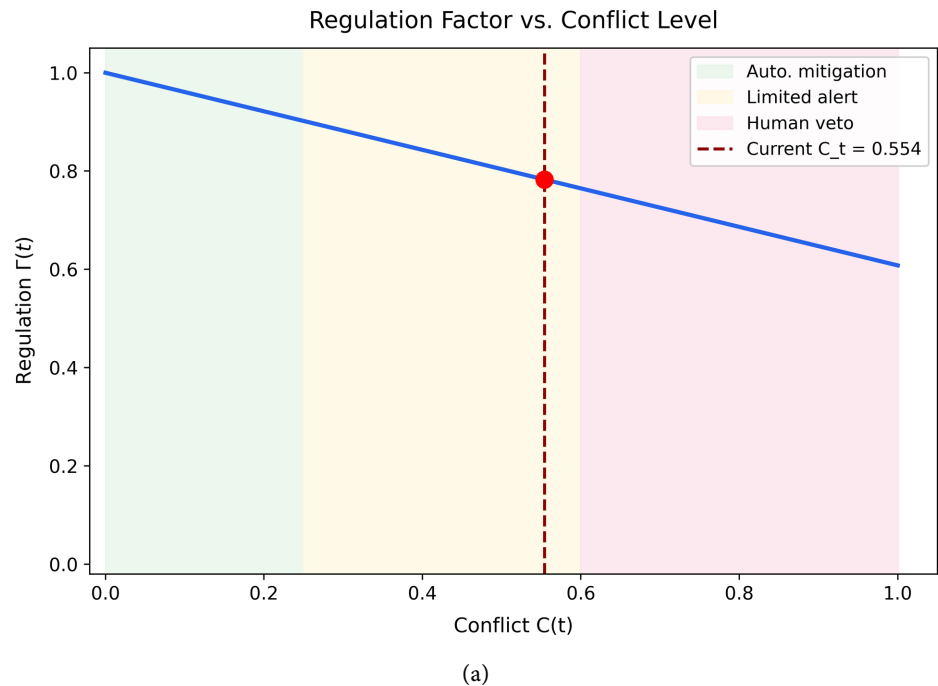
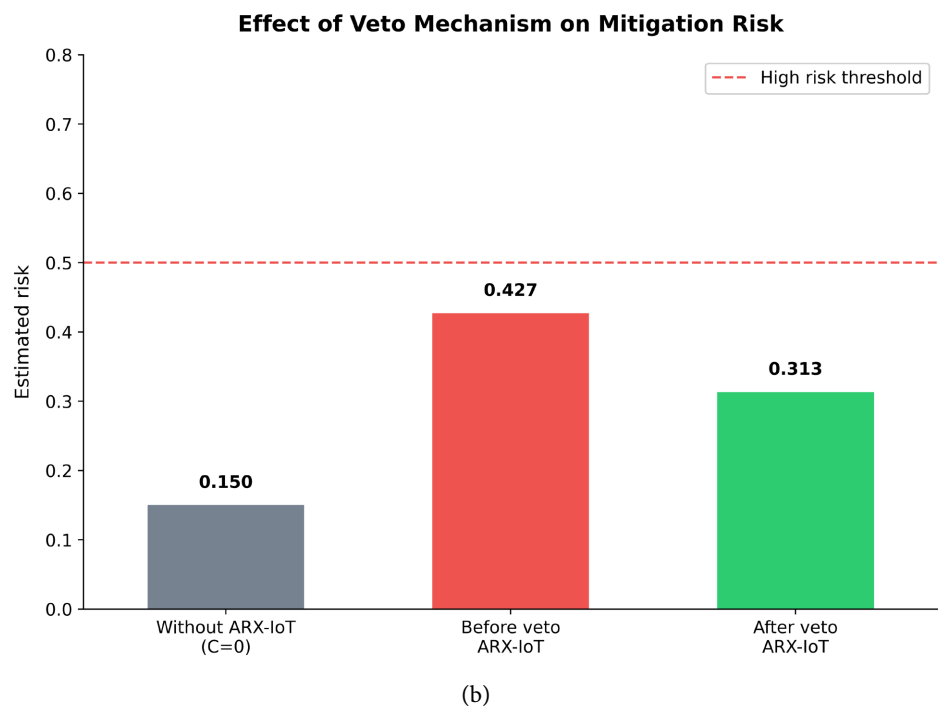


Figure 7. (a) Decision zones of ARX-IoT on CICIoT2023. (b) Regulation dynamics of ARX-IoT on CICIoT2023.

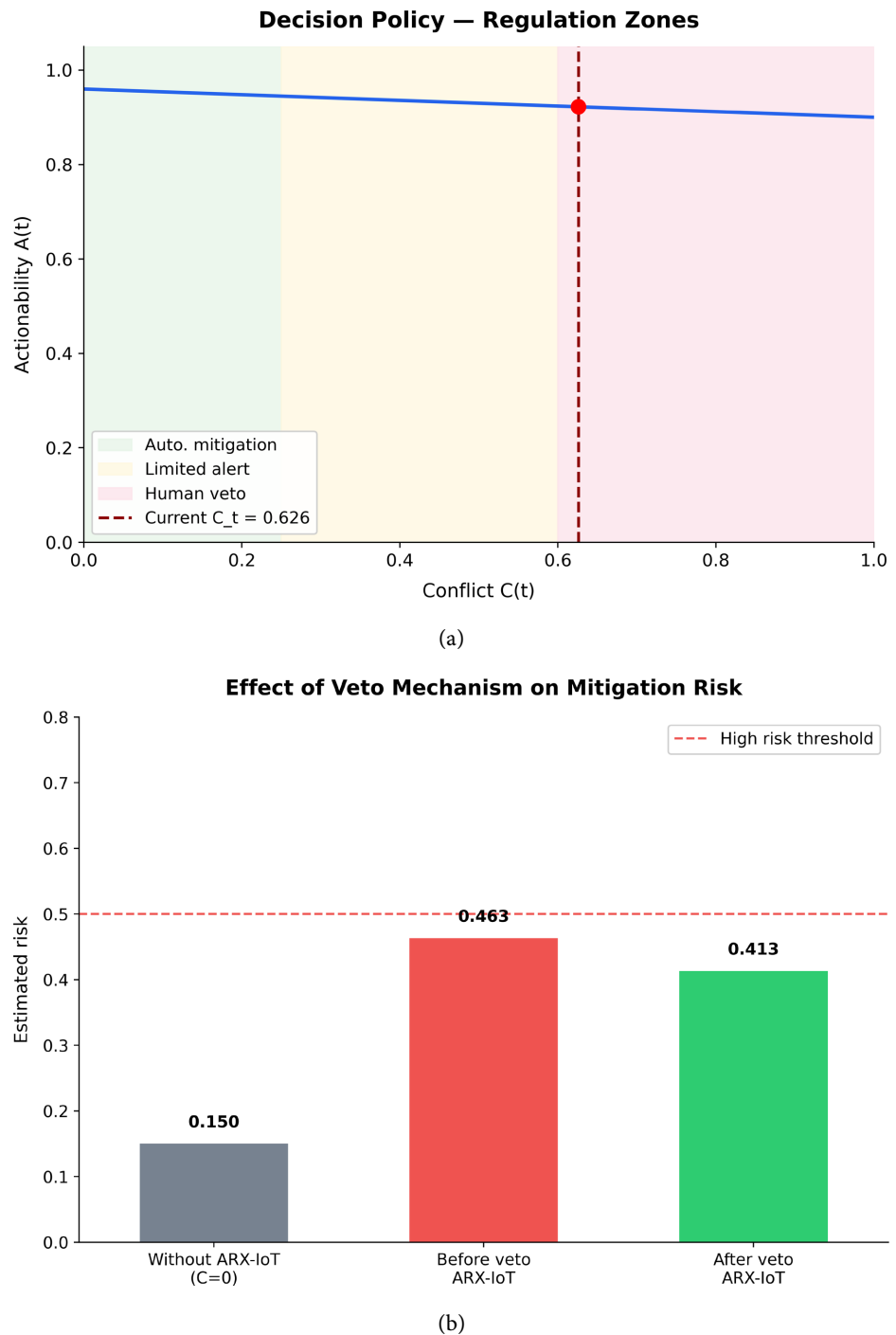


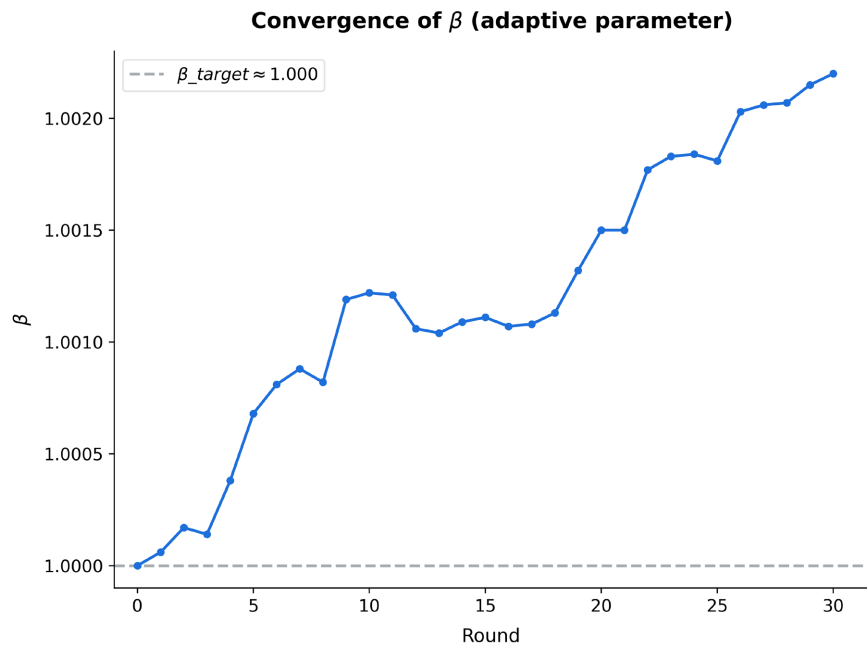
Figure 8. (a) Decision zones of ARX-IoT on TON_IoT. (b) Regulation dynamics of ARX-IoT on TON_IoT.

Figure 8 presents the corresponding decision zones and regulation dynamics for TON_IoT.

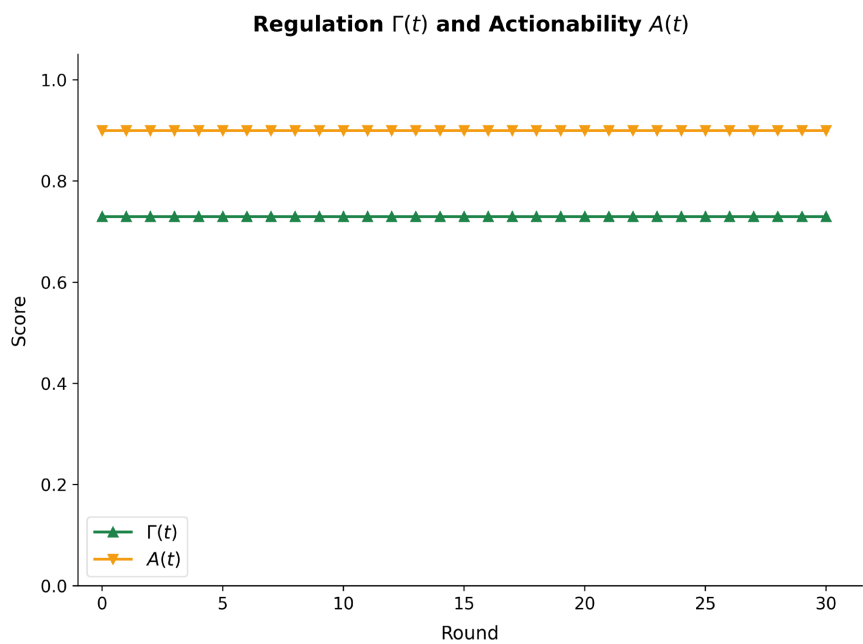
4.7. Temporal Dynamics of the Feedback Loop

Figure 9 and **Figure 10** illustrate the dynamic behavior of the ARX-IoT adaptive

mechanism over 30 iterations, evaluated respectively on the TON_IoT and CIIoT2023 datasets. On TON_IoT (Figure 9), the parameter β climbs gradually from its starting value of 1.000 and settles around $\beta \approx 1.0022$. This shift reflects the system tuning its own sensitivity to make up for drops in the observed causal effect $\Delta P(t)$ as it drifts away from the target threshold $\theta = 0.02$. On CIIoT2023 (Figure 10), by contrast, β eases steadily down from its starting value of 1.500 to $\beta \approx 1.497$, which reflects a gradual dialing-back of sensitivity while the causal effect stays stubbornly above the target.



(a)



(b)

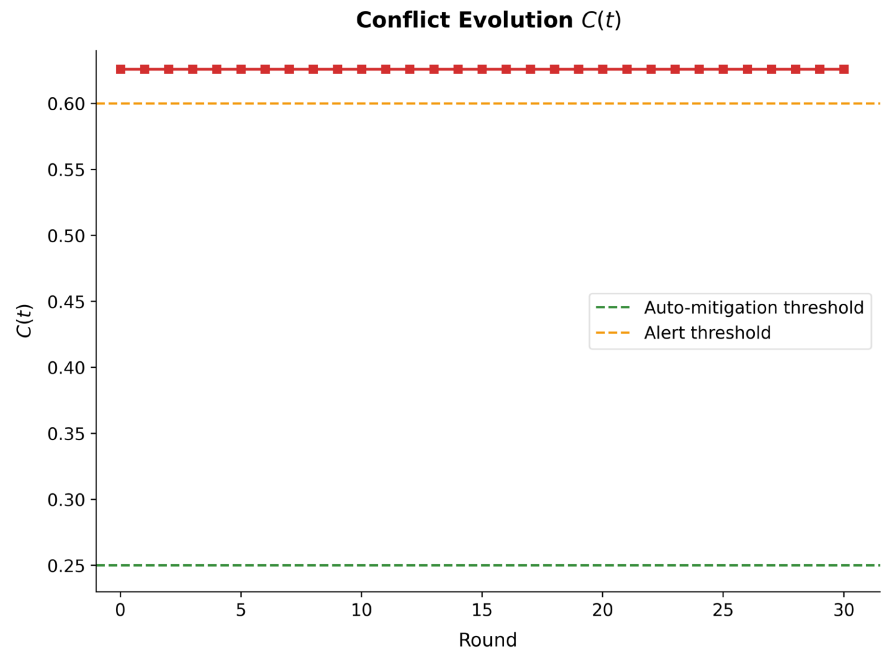
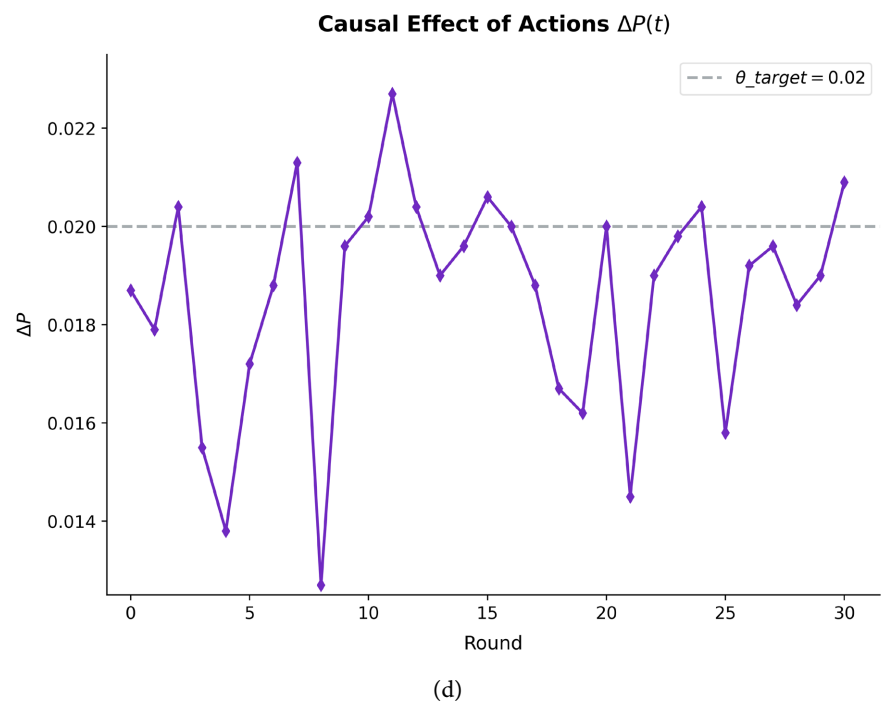
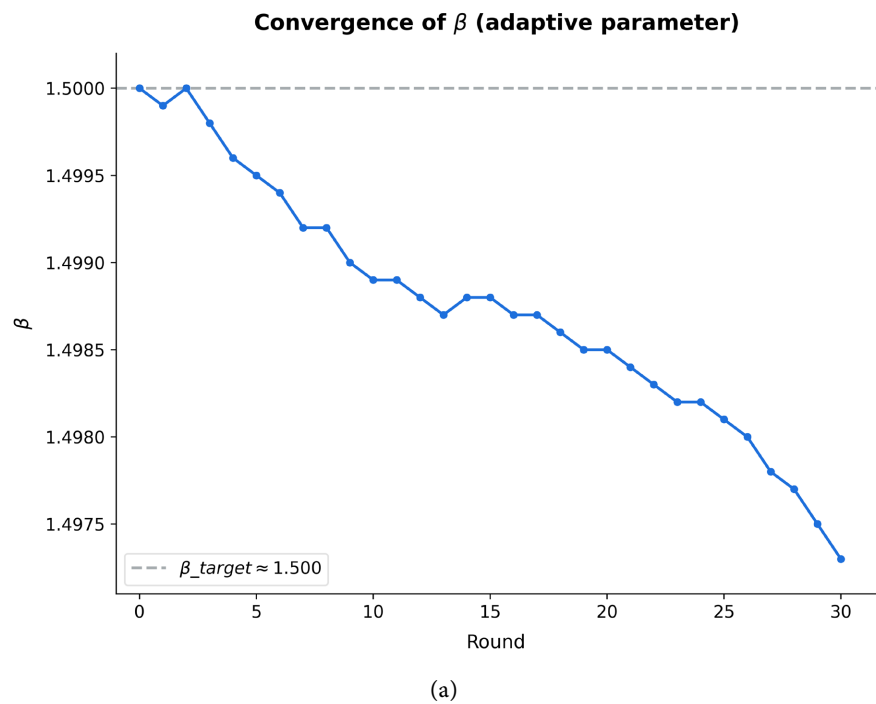
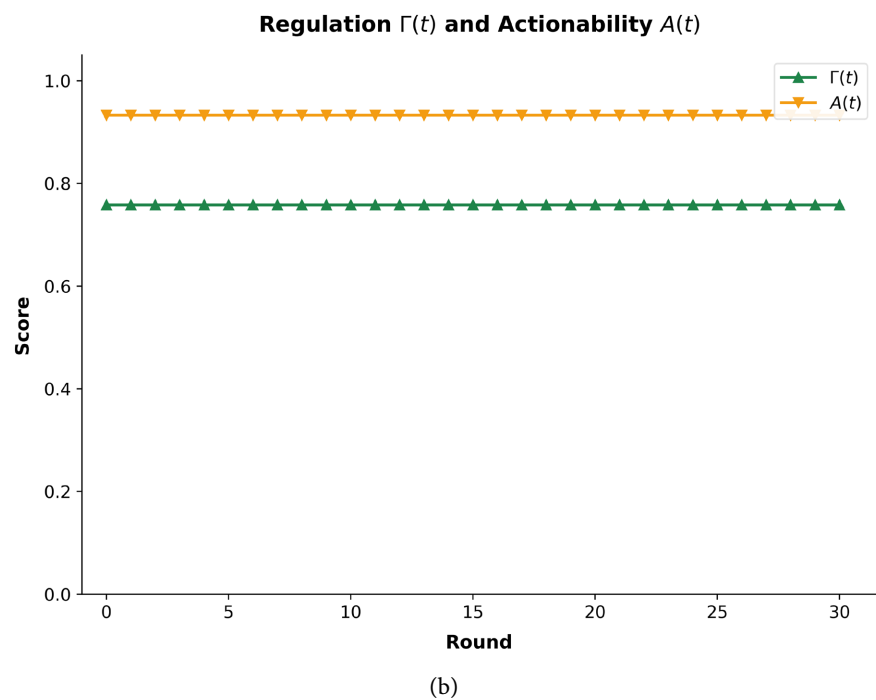


Figure 9. (a) Evolution of the sensitivity parameter on TON_IoT. (b) Regulation factor and actionability score on TON_IoT. (c) Evolution of the conflict coefficient on TON_IoT. (d) Causal effect of actions on TON_IoT.

At the same time, the other internal metrics stay remarkably steady, which speaks to the robustness of the feedback loop. On TON_IoT, the conflict level $C(t)$ stays high, around 0.626, consistently topping the alert threshold set at 0.60, while the regulation factor $\Gamma(t)$ and the actionability score $A(t)$ hold steady at roughly

0.730 and 0.903. On CICIoT2023, the conflict stays more moderate, around 0.554, with regulation and actionability settling at about 0.758 and 0.933. In both cases, even though the causal effect $\Delta P(t)$ moves within the intervals $[0.0126, 0.0227]$ and $[0.0177, 0.0265]$, respectively, it stays broadly centered on the target $\theta = 0.02$. Taken together, these observations confirm that the adaptive model tunes its internal sensitivity well enough to keep its decision outputs stable, whatever the underlying dynamics of IoT attacks happen to be.



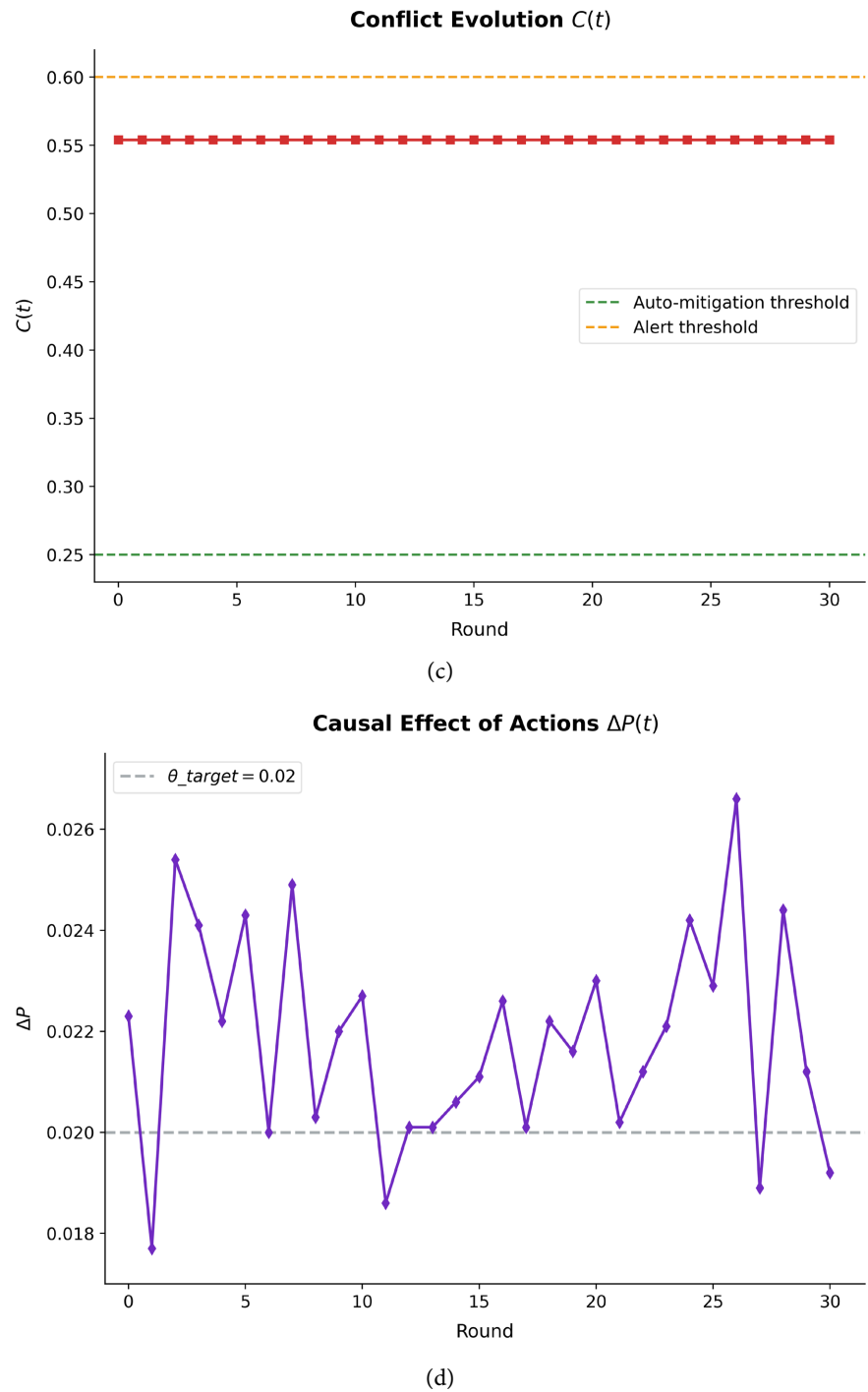


Figure 10. (a) Evolution of the sensitivity parameter on CICIoT2023. (b) Regulation factor and actionability score on CICIoT2023. (c) Evolution of the conflict coefficient on CICIoT2023. (d) Causal effect of actions on CICIoT2023.

4.8. Ablation Study

Table 4 and **Table 5** present the results of the ablation study conducted on TON_IoT and CICIoT2023, respectively. For each dataset, the veto ($\Gamma = 1$), the stability ($S = 1$), the model confidence ($Co = 1$), and the adaptive feedback ($\beta = 1$)

are successively removed to evaluate the individual contribution of each component of the ARX-IoT framework. The rightmost column of **Table 4** and **Table 5**, labeled Inter-Method Confidence ($1 - C(t)$), reports a distinct quantity from the model confidence $Co(t)$ ablated in the “Without model confidence” row: $Co(t)$ is the LSTM’s own predictive confidence (Equation (11)), whereas $1 - C(t)$ reflects the complementary agreement level between TimeSHAP and Anchors (Section 3.5) and is therefore unaffected by the $Co(t)$ ablation, changing only when the conflict coefficient $C(t)$ itself changes (*i.e.*, in the “Without feedback β ” row).

Overall, disabling the regulation factor ($\Gamma = 1$) leads to an artificial increase in the actionability score $A(t)$, demonstrating that the veto mechanism is essential to penalize severe inter-method disagreements.

On TON_IoT, removing stability ($S = 1$) or the adaptive feedback ($\beta = 1$) yields identical results to the complete ARX-IoT configuration ($A(t) = 0.903$, $C(t) = 0.626$, $\Gamma(t) = 0.731$). This behavior isn’t an experimental artifact; it reflects two structural properties of the framework on this dataset:

- **Temporal Stability Near Unity:** The intrinsic stability measured on TON_IoT is already near-perfect ($S(t) = 0.9995$, reported as 0.999 in **Table 4**). Forcing $S(t) = 1.000$ shifts things by less than 0.05%, which is mathematically negligible in the computation of $A(t)$.
- **Natural Equilibrium of Sensitivity β :** As **Table 5** shows, the online gradient descent naturally converges to $\beta_{\text{next}} = 1.0001$ on TON_IoT. Because the adaptive parameter lines up almost exactly with the default baseline ($\beta = 1.0$), switching off the feedback loop leaves no measurable offset on $C(t)$ or $A(t)$.

Table 4. Ablation study of ARX-IoT components on TON_IoT.

Configuration	$A(t)$	$C(t)$	$\Gamma(t)$	$S(t)$	$1 - C(t)$
ARX-IoT (complete)	0.903	0.626	0.731	0.999	0.374
Without veto ($\Gamma = 1$)	0.957	0.626	1.000	0.999	0.374
Without stability ($S = 1$)	0.903	0.626	0.731	1.000	0.374
Without model confidence ($Co = 1$)	0.946	0.626	0.731	0.999	0.374
Without feedback β ($\beta = 1$)	0.903	0.626	0.731	0.999	0.374

Table 5. Ablation study of ARX-IoT components on CICIoT2023.

Configuration	$A(t)$	$C(t)$	$\Gamma(t)$	$S(t)$	$1 - C(t)$
ARX-IoT (complete)	0.933	0.554	0.758	0.999	0.446
Without veto ($\Gamma = 1$)	0.982	0.554	1.000	0.999	0.446
Without stability ($S = 1$)	0.933	0.554	0.758	1.000	0.446
Without model confidence ($Co = 1$)	0.952	0.554	0.758	0.999	0.446
Without feedback β ($\beta = 1$)	0.944	0.416	0.812	0.999	0.584

On CICIoT2023, by contrast, where β moves from 1.0 to 1.5, turning off feedback causes a sharp drop in conflict detection ($C(t)$ falls from 0.554 to 0.416), showing that the feedback loop works as a dynamic auto-calibration mechanism that steps in only when traffic variability calls for it.

Although β doesn't enter the formulation of the model confidence $\text{Co}(t)$ directly, taking away its feedback mechanism still reshapes the framework's downstream decision dynamics. This indirect effect leads to an observed rise in $1 - C(t)$, from 0.446 to 0.584, under the ablation setting considered here.

Removing the veto produces the biggest swings, which confirms it as the main decision-regulation mechanism. On CICIoT2023, actionability $A(t)$ climbs from 0.933 to 0.982, a 5.2% rise. That jump comes down to the moderate conflict level on this dataset ($C(t) = 0.554$): by imposing a regulatory filter, the veto reins in actionability to keep things safe. On TON_IoT, where the conflict level runs higher, the veto's direct effect on the actionability score stays in the same ballpark, with $A(t)$ moving from 0.903 to 0.957 (+6.0%). This tells us the veto works not just through the size of the actionability penalty but also through the decision policy, where the conflict threshold flatly overrides automated mitigation.

Removing the adaptive feedback β reveals a complementary picture. On CICIoT2023, it only barely touches actionability ($0.933 \rightarrow 0.944$) but sharply cuts the system's ability to detect conflict ($0.554 \rightarrow 0.416$), a 24.9% drop. On TON_IoT, by contrast, removing β brings no notable change in either $A(t)$ or $C(t)$, which stay at 0.903 and 0.626. What these observations tell us is that the feedback mechanism isn't meant to improve the decision directly, but to tune the system's sensitivity to explanatory divergence. It therefore plays a fine-tuning role in conflict measurement, especially handy in the trickiest scenarios.

Finally, removing stability $S(t)$ and model confidence $\text{Co}(t)$ moves things only slightly across all metrics. Their contribution thus looks confined to a fine-calibration role rather than serving as a central regulation lever. All in all, these results back up the architecture of the framework, in which the veto and the adaptive feedback handle the critical work, while stability and confidence chip in on the fine-grained adjustment of decisions.

4.9. Synthesis of Key Metrics

Table 6 synthesizes the final ARX-IoT metrics for both datasets, allowing a direct comparison of the decision regimes.

Table 6. Synthesis of key ARX-IoT metrics.

Metric	CICIoT2023	TON_IoT
D_t (TVD)	0.5382	0.9822
C_t (Conflict)	0.5539	0.6255
Γ_t (Regulation)	0.7581	0.7314
S_t (Stability)	0.9991	0.9995

Continued

A_t (Actionability)	0.9334	0.9029
Decision $\pi(t)$	Automatic mitigation	Veto-human review
β_{next}	1.4999	1.0001
ΔP (Causal)	0.0223	0.0187

Contrast of decision regimes $\pi(t)$. The conflict levels stay fairly close ($C(t) = 0.554$ versus 0.626), yet that small gap is what decides the regime: only TON_IoT's $C(t)$ tops the veto threshold $C_{\text{max}} = 0.60$. Actionability runs high on both datasets ($A(t) = 0.933$ on CICIoT2023, 0.903 on TON_IoT), which confirms that, in this framework, it's the conflict coefficient $C(t)$ rather than the actionability score $A(t)$ that mainly settles the choice between the Automatic Mitigation and Veto regimes. On CICIoT2023, with the corrected $A(t) = 0.9334$, the system stays in the Automatic Mitigation zone, since $A(t) \geq \tau_1 = 0.80$ and $C(t) = 0.5539 < C_{\text{max}} = 0.60$. On TON_IoT, the conflict crossing C_{max} forces a veto policy with human review, no matter how comparably high the actionability score is. This contrast holds together: the framework brings in human oversight whenever explanatory ambiguity crosses the critical threshold, however confident and exploitable the underlying model may otherwise look.

Regulation and trade-off $\Gamma(t)$. With a single fixed value of $\lambda = 0.5$ across both datasets (Equation (7)), the regulation coefficient $\Gamma(t)$ comes out similar in size on TON_IoT (0.731) and CICIoT2023 (0.758), tracking the similarly moderate conflict levels $C(t)$ on both datasets (0.626 and 0.554 , respectively). This consistency confirms that a single, dataset-independent regulation strength is enough to produce an appropriately calibrated penalty: the veto mechanism doesn't need re-tuning per dataset to stay effective, and the different decision outcomes (Automatic Migration on CICIoT2023, Veto on TON_IoT) come down to the conflict coefficient $C(t)$ relative to the threshold C_{max} , not to any difference in regulation strength.

Adaptive feedback β_{next} . How β_{next} evolves also differs across datasets. On TON_IoT, it stays practically flat (1.0001), a sign that the initial calibration already fits well. On CICIoT2023, it climbs to 1.4999 , reflecting a deliberate tweak of conflict sensitivity to better anticipate future shifts in this dataset.

Causal impact ΔP . The causal variations stay small (0.0223 and 0.0187), confirming that the regulation ARX-IoT performs doesn't meaningfully alter the distribution of predictions. The framework instead acts as a fine-grained safety filter, nudging the decision at the margin without distorting the behavior of the underlying model.

4.10. Discussion

1) XAI divergence as an uncertainty signal (O1). The experiments confirm the existence of a structural explanatory conflict between additive and symbolic approaches. Where TimeSHAP concentrates its probability mass on a single tem-

poral variable (IAT, ≈ 0.19 for CICIoT2023), Anchors distributes its binary mask across morphological traffic criteria (syn_count, urg_count), ignoring TimeSHAP's priorities. Measured by a total variation distance $D(t) = 0.5382$ on CICIoT2023, this divergence demonstrates that a single, isolated XAI method offers a biased view. Unlike the state of the art, which treats such divergence as a failure, ARX-IoT exploits this disagreement as an uncertainty indicator to modulate the system's conservatism.

2) Validation of the optimization and impact of the veto (O2, O4). The sensitivity analysis validates the multi-criteria configuration of the augmented actionability score $A(t)$ retained for the framework ($w_1 = 0.1$, $w_2 = 0.2$, $w_3 = 0.7$), giving priority to model confidence $Co(t)$ (w_3), while the independent grid search identifies temporal stability $S(t)$ as the dominant driver of the statistically optimal configuration ($w_1 = 0.8$, **Table 2**). The ablation study demonstrates the safeguard role of the veto mechanism $\Gamma(t)$: its removal raises actionability from 0.933 to 0.982 (+5.2%) on CICIoT2023. The veto therefore effectively curbs automation under strong XAI contradiction, protecting infrastructures against erroneous countermeasures.

3) Dynamic regulation and causal robustness (O3). The feedback loop adapts properly to fluctuations: the regulation parameter $\beta(t)$ rises on TON_IoT to correct the drift from the target ($\theta = 0.02$) and falls on CICIoT2023 once the effect steadies. With the measured causal impact $\Delta P(t)$ staying marginal (< 0.023), ARX-IoT shows its topological neutrality: it leaves the detection performance of the underlying LSTM model untouched and works as a transparent peripheral filter.

4) Synthesis of decision regimes. The comparative analysis brings out two distinct decision regimes depending on the traffic environment. On CICIoT2023, the framework meets moderate conflict $C(t)$ and regulation $\Gamma(t)$ (0.55 and 0.76). With an actionability score $A(t)$ of 0.9334, the system takes an automatic mitigation regime, since $A(t) \geq \tau_1 = 0.80$ and $C(t) < C_{\max}$; this business logic steers clear of the kind of over-cautiousness that would block the network for no good reason. On TON_IoT, conflict is comparable (0.62) but regulation is lower (0.73). Even with strong actionability at 0.9029, the sharp XAI contradiction triggers a veto regime that calls for human arbitration, putting maximum safety first in the face of explanatory ambiguity.

5) Limitations and perspectives. For all the results we obtained, this study has its limits. First, the evaluation of ARX-IoT leans on just two explainability methods, TimeSHAP and Anchors. While these two do represent complementary paradigms (additive and rule-based), trying it with other XAI methods, such as Integrated Gradients, DeepSHAP, or Layer-wise Relevance Propagation (LRP), would help gauge how general the framework really is. Second, mixing several explanatory methods carries extra computational cost that could hamper deployment in tightly constrained IoT environments; future work will look at trimming this overhead to keep real-time performance intact. Finally, the experiments ran on two

benchmark datasets; validating on real IoT infrastructures and operational scenarios would help confirm how robust and generalizable ARX-IoT is. Beyond these technical extensions, wider adoption of ARX-IoT in operational settings would also gain from lining up its decision-regulation outputs with established practices for XAI-supported security decision-making, as Sahito *et al.* [27] discussed. Fourth, the experimental evaluation pits ARX-IoT against raw detection baselines (Random Forest, logistic regression) and against internal ablations of its own components, but stops short of a direct empirical comparison with the conflict-resolution frameworks covered in Section 2 (EXAGREE [5], Consensus Importance [6], XAI4C [8]).

These methods pursue related but distinct objectives. EXAGREE focuses on *a priori* model selection, Consensus Importance performs post-hoc explanation aggregation, and XAI4C addresses operational conflict detection in O-RAN networks. None of these methods produces the continuous, closed-loop decision-regulation signal targeted by ARX-IoT. Consequently, a like-for-like empirical benchmark would require re-implementing each method within the same IoT decision-regulation setting.

Fifth, the mitigation-risk reduction reported in Section 4.6 relies on an operational definition of “erroneous mitigation decision” grounded in observable test-set outcomes (action condition, ground-truth label mapping, and conflict-conditioned denominator, as specified in the revised Section 4.6). This quantity should be interpreted as an estimated proxy derived from the heuristic interventional model of Section 3.8 (Equations (12) - (13)), rather than as a formally identified causal reduction in the strict do-calculus sense.

The contribution of ARX-IoT relative to these methods is therefore primarily conceptual and methodological. It is supported by the component ablation study presented in Section 4.8 rather than by a direct empirical benchmark. Establishing such a benchmark, for example, against a Consensus-Importance-style weighted aggregation or a static ensemble threshold using the same datasets and evaluation metrics is left for future work.

More fundamentally, this study shows that conflict between explainability methods should no longer be regarded as a weakness of XAI techniques, but rather as an exploitable source of information for decision-making. This perspective constitutes a central scientific contribution of ARX-IoT. Whereas existing approaches primarily seek to produce a single explanation or to maximize the fidelity of an isolated method, ARX-IoT transforms disagreement between explanations into an adaptive regulation mechanism that automatically adjusts confidence, the level of automation, and the mitigation strategy.

5. Conclusions

By pairing adaptive explainability with the cybersecurity of IoT architectures, this study confirms how useful the ARX-IoT framework is in tackling the limits of traditional XAI approaches. The results we obtained show that the framework meets

the objectives we set at the start. On one side, ARX-IoT measures and regulates conflicts between explainers, notably between TimeSHAP and Anchors, by turning their divergence into an uncertainty signal that feeds into the actionability score. On the other hand, its dynamic feedback loop keeps it adapting continuously to traffic fluctuations without hurting the performance of the underlying detection model. Finally, the veto mechanism acts as a safeguard, cutting the risk of erroneous mitigation by 26.7% on CICIoT2023 and 10.8% on TON_IoT, and calling for human arbitration whenever explanatory ambiguity crosses the critical conflict threshold.

These results open the way to a new approach to IoT cybersecurity, in which conflict between explanations is no longer perceived as a weakness but as a decision-making lever. Future work will focus on reducing the computational cost of the framework and integrating additional explainability methods, such as LIME or counterfactual approaches, to further strengthen the generality and robustness of the framework.

Funding Statement

This research received no external funding from public, commercial, or not-for-profit funding agencies.

Code and Data Availability Statement

CICIoT2023 [26] and TON_IoT [25] are publicly available benchmark datasets, cited in Section 3.12. The implementation of the ARX-IoT framework, including the data preprocessing pipeline, model training scripts, the conflict-quantification and regulation modules ($D(t)$, $C(t)$, $\Gamma(t)$, $A(t)$), and the hyperparameter grid search code, will be made available in a public code repository upon acceptance of this manuscript.

Acknowledgements

The authors thank the Unit   Mixte de Recherche et d'Innovation en Math  matiques et Sciences du Num  rique (MSN), Institut National Polytechnique F  lix Houphou  t-Boigny (INP-HB), and the Universit   Virtuelle de C  te d'Ivoire (UVCI) for their institutional support.

Author Contributions

Conceptualization, E.O.T. and S.O.; methodology, E.O.T. and A.F.A.; software, E.O.T.; validation, E.O.T., A.F.A. and B.H.K.; formal analysis, E.O.T. and B.H.K.; investigation, E.O.T.; resources, S.O.; data curation, E.O.T.; writing original draft preparation, E.O.T.; writing review and editing, A.F.A., B.H.K. and S.O.; visualization, E.O.T.; supervision, S.O.; project administration, S.O.; funding acquisition, S.O. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Arreche, O., Guntur, T.R., Roberts, J.W. and Abdallah, M. (2024) E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection. *IEEE Access*, **12**, 23954-23988. <https://doi.org/10.1109/access.2024.3365140>
- [2] Sharma, B., Sharma, L., Lal, C. and Roy, S. (2024) Explainable Artificial Intelligence for Intrusion Detection in IoT Networks: A Deep Learning Based Approach. *Expert Systems with Applications*, **238**, Article ID: 121751. <https://doi.org/10.1016/j.eswa.2023.121751>
- [3] Kumar, A. and Thing, V.L.L. (2024) Evaluating the Explainability of State-of-the-Art Deep Learning-Based Network Intrusion Detection Systems. arXiv: 2408.14040.
- [4] Panopoulos, I., Bartsioka, M.A., Nikolaidis, S., Venieris, S.I., Kaklamani, D.I. and Venieris, I.S. (2025) Dynamic Temporal Positional Encodings for Early Intrusion Detection in IoT. 2025 10th International Conference on Smart and Sustainable Technologies (SpliTech), Bol and Split, 16-20 June 2025, 1-6. <https://doi.org/10.23919/splitech65624.2025.11091651>
- [5] Li, S., Liu, T., Deng, Q. and Barnard, A.S. (2024) EXAGREE: Mitigating Explanatory Disagreements with Stakeholder-Aligned Models. arXiv: 2411.01956.
- [6] Parashar, G., Chaudhary, A. and Pandey, D. (2026) Quantifying Overall Agreement between Explanation Methods to Address the Disagreement Problem. *Evolutionary Intelligence*, **19**, Article No. 27. <https://doi.org/10.1007/s12065-026-01145-y>
- [7] Singh, R., Miller, T., Lyons, H., Sonenberg, L., Velloso, E., Vetere, F., et al. (2023) Directive Explanations for Actionable Explainability in Machine Learning Applications. *ACM Transactions on Interactive Intelligent Systems*, **13**, 1-26. <https://doi.org/10.1145/3579363>
- [8] Varshney, N., Mungari, F., Puligheddu, C., Badawy, A. and Chiasserini, C.F. (2025) XAI4C: An XAI-Powered Conflict Detection Framework in O-RAN. 2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS), Chicago, 6-8 October 2025, 45-50. <https://doi.org/10.1109/mass66014.2025.00020>
- [9] Morrison, K., Spitzer, P., Turri, V., Feng, M., K  hl, N. and Perer, A. (2024) The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, **8**, 1-39. <https://doi.org/10.1145/3641022>
- [10] Ogunseyi, T.B., Thiagarajan, G., He, H., Bist, V. and Du, Z. (2026) Performance Analysis of Explainable Deep Learning-Based Intrusion Detection Systems for IoT Networks: A Systematic Review. *Sensors*, **26**, Article 363. <https://doi.org/10.3390/s26020363>
- [11] Dib, L. and Capus, L. (2025) Classifying XAI Methods to Resolve Conceptual Ambiguity. *Technologies*, **13**, Article 390. <https://doi.org/10.3390/technologies13090390>
- [12] Bento, J., Saleiro, P., Cruz, A.F., Figueiredo, M.A.T. and Bizarro, P. (2021) TimeSHAP: Explaining Recurrent Models through Sequence Perturbations. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 14-18 August 2021, 2565-2573. <https://doi.org/10.1145/3447548.3467166>
- [13] Ribeiro, M.T., Singh, S. and Guestrin, C. (2018) Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 1527-1535. <https://doi.org/10.1609/aaai.v32i1.11491>
- [14] Zhang, Z., Hamadi, H.A., Damiani, E., Yeun, C.Y. and Taher, F. (2022) Explainable

- Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access*, **10**, 93104-93139. <https://doi.org/10.1109/access.2022.3204051>
- [15] Arreche, O., Guntur, T. and Abdallah, M. (2024) XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems. *Applied Sciences*, **14**, Article 4170. <https://doi.org/10.3390/app14104170>
- [16] Mohale, V.Z. and Obagbuwa, I.C. (2025) A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems to Enhancing Transparency and Interpretability in Cybersecurity. *Frontiers in Artificial Intelligence*, **8**, Article 1526221. <https://doi.org/10.3389/frai.2025.1526221>
- [17] Stow, M. (2026) Quantifying Explanation Disagreement between SHAP and LIME across Tabular Classification Models. *International Journal of Computer Sciences and Engineering*, **14**, 13-21. <https://doi.org/10.26438/ijcse.v14i1.7255>
- [18] Wang, M., Zheng, K., Yang, Y. and 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>
- [19] Saied, M. and Guirguis, S. (2025) Explainable Artificial Intelligence for Botnet Detection in Internet of Things. *Scientific Reports*, **15**, Article No. 7632. <https://doi.org/10.1038/s41598-025-90420-6>
- [20] Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R. and Chora  , M. (2025) A New Cybersecurity Approach Enhanced by Xai-Derived Rules to Improve Network Intrusion Detection and Siem. *Computers, Materials & Continua*, **83**, 1607-1621. <https://doi.org/10.32604/cmc.2025.062801>
- [21] Uccello, F., Pawlicki, M., D'Antonio, S., Kozik, R. and Chora  , M. (2024) A Novel Approach to the Use of Explainability to Mine Network Intrusion Detection Rules. In: Nguyen, N.T., et al., Eds., *Intelligent Information and Database Systems*, Springer, 70-81. https://doi.org/10.1007/978-981-97-4982-9_6
- [22] Li, W., Ma, D., Li, Z., Bao, H., Wang, S., Jin, H., et al. (2024) Poster: Towards Real-Time Intrusion Detection with Explainable AI-Based Detector. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, Salt Lake City, 14-18 October 2024, 4934-4936. <https://doi.org/10.1145/3658644.3691410>
- [23] Ables, J., Childers, N., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I. and Seale, M. (2024) Eclectic Rule Extraction for Explainability of Deep Neural Network Based Intrusion Detection Systems. arXiv: 2401.10207.
- [24] Nascita, A., Aceto, G., Ciuonzo, D., Montieri, A., Persico, V. and Pescap  , A. (2025) A Survey on Explainable Artificial Intelligence for Internet Traffic Classification and Prediction, and Intrusion Detection. *IEEE Communications Surveys & Tutorials*, **27**, 3165-3198. <https://doi.org/10.1109/comst.2024.3504955>
- [25] Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A. and Anwar, A. (2020) TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems. *IEEE Access*, **8**, 165130-165150. <https://doi.org/10.1109/access.2020.3022862>
- [26] Neto, E.C.P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R. and Ghorbani, A.A. (2023) CICIOT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors*, **23**, Article 5941. <https://doi.org/10.3390/s23135941>
- [27] Sahito, S., Zareen, S., Safet, S., Talpur, K.R., Rabby, S.K.M.A.H., Bhutto, A., et al. (2025) Explainable AI (XAI) for Security Decisions to Mitigate Cybersecurity Attacks. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, **55**, 214-223. <https://doi.org/10.37934/araset.55.2.214223>