Ψ-Risk-DT: A Neurosymbolic Framework for Adaptive Zero-Day Threat Detection in Digital Twin Ecosystems

Roberto Pazzi
Dipartimento di Informatica
Università degli Studi dell'Insubria
Varese, Italy
ORCID: 0009-0009-3454-5678
roberto.pazzi@uninsubria.it

Davide Facheris
Dipartimento di Informatica
Università degli Studi dell'Insubria
Varese, Italy
ORCID: 0009-0007-7990-8335
dfacheris@studenti.uninsubria.it

Davide Tosi
Dipartimento di Informatica
Università degli Studi dell'Insubria
Varese, Italy
ORCID: 0000-0003-3815-2512
davide.tosi@uninsubria.it

***Abstract*—Resilient digital transformation increasingly depends on trustworthy cyber-physical software, where Digital Twin (DT) ecosystems—virtual replicas of low-power Internet of Things (IoT) devices, edge gateways and industrial assets—must remain available, observable and explainable under attack. These environments are particularly exposed to zero-day threats that exploit the semantic gap between physical states and virtual expectations, while heterogeneous topologies, concept drift and constrained edge resources erode the assumptions of conventional Intrusion Detection Systems (IDSs). We present Ψ-Risk-DT, a Neurosymbolic (NeSy) framework that couples entropy-based anomaly detection with an Associated Random Neural Network (ARNN) and an RDF/SPARQL semantic-reasoning layer through a formal entropy-gated operator Ψ; a Modular Semantic Update (MSU) mechanism rewrites only the affected portions of the DT knowledge graph, and a hybrid loss aligns classification accuracy, graph coherence and semantic consistency with Lyapunov-style stability guarantees. The framework is directly validated in a containerised Network Time Protocol (NTP) amplification scenario representative of volumetric zero-day-like attacks against DT ecosystems, where it achieves an Area Under the Curve (AUC) of 0.993, a False-Positive Rate (FPR) of 0.021 and an end-to-end pipeline latency of approximately 25.8 ms, while MSU reduces the symbolic-update load by up to 88% relative to a global-rewrite ablation evaluated on the same trace. For comparative positioning, these figures are contrasted with values reported in the literature for established IoT/DT IDS baselines (Kitsune, deep-learning zero-day detectors, NeSy-IDS and the (H-DIR)^2 predecessor); under the conditions reported in those works, Ψ-Risk-DT exhibits an order-of-magnitude latency gain[C] and a 20–40% FPR reduction[C] relative to deep-learning baselines, presented as a positional comparison rather than as a co-evaluated measurement. Broader empirical replication on out-of-distribution (OOD) variants of community benchmarks (Kitsune, WiseML 2024, Sec4ML 2023), on multi-vector Routing Protocol for Low-Power and Lossy Networks (RPL) attack scenarios, and on heterogeneous edge hardware is identified as ongoing future work. Overall, Ψ-Risk-DT provides explainable, entropy-gated, semantically adaptive protection for DT ecosystems and contributes to FSOFT topics on resilient digital transformation, trustworthy cyber-physical software, IoT security, adaptive threat detection and explainable neurosymbolic Artificial Intelligence (AI).**

Keywords— Digital Twins; Zero-Day Threat Detection; Neurosymbolic Cybersecurity; Entropy-Based Anomaly Detection; Explainable Intrusion Detection.

I. Introduction

Digital Twin (DT) ecosystems—cyber-physical environments in which physical assets are continuously mirrored by virtual counterparts—are rapidly permeating safety-critical infrastructures such as smart manufacturing, energy grids and connected healthcare [1]. Streams of multimodal telemetry link physical sensors to cloud or edge services through low-power wireless networks (e.g., the Routing Protocol for Low-Power and Lossy Networks, RPL, and 6LoWPAN). This bidirectional, semantically rich coupling is precisely what distinguishes a DT ecosystem from a generic Internet of Things (IoT) deployment: a DT does not only transport packets, it maintains a live virtual contract about the expected state, configuration and behaviour of each physical asset. A zero-day attack against a DT therefore manifests as a *semantic desynchronisation*—a drift between physical reality and virtual expectation—rather than as a purely statistical traffic deviation. Conventional Intrusion Detection Systems (IDSs) struggle in this regime: they exhibit high false-positive rates on bursty but benign DT traffic, miss stealthy desynchronisation attacks, and lack the semantic context required to explain alerts to human operators or upstream resilience controllers [2], [3].

Similar approaches: Three families of techniques dominate the state of the art and are reviewed in detail in Section II. *Statistical and entropy-based detectors* [3], [8], [15] respond quickly to volumetric anomalies but are semantically opaque; *deep-learning IDSs* (e.g., Kitsune-style autoencoders [9] and recent zero-day detectors [2], [12]) raise accuracy at the cost of interpretability and edge feasibility; *neurosymbolic and ontology-driven approaches* [4], [11], [13], [14] reintroduce symbolic explainability but typically operate offline, on static graphs, and without entropy-gated coupling to streaming telemetry. None of these strands jointly addresses the DT-specific requirement of entropy-gated, semantically adaptive, real-time detection under edge resource constraints.

Problem statement: How can neurosymbolic methods be integrated into DT security frameworks to deliver resilient, explainable defence against zero-day attacks in low-power IoT networks, while preserving real-time guarantees and bounded edge cost?

Gap: Existing IDSs typically address (i) statistical detection, (ii) symbolic explainability, or (iii) DT synchronisation in isolation. To the best of our knowledge, no prior work provides an *entropy-gated neurosymbolic operator* that triggers semantic reasoning only when the entropy signal indicates a candidate zero-day deviation, performs *localised, modular* Resource Description Framework (RDF) updates on the DT knowledge graph instead of full rewrites, and is jointly optimised through a *hybrid loss* that aligns classification accuracy, graph coherence and semantic consistency under Lyapunov-style stability conditions [11].

Objective: We propose Ψ-Risk-DT, a neurosymbolic framework that extends entropy-driven anomaly detection with an operator Ψ coupling an Associated Random Neural Network (ARNN) with RDF reasoning expressed through the SPARQL Protocol and RDF Query Language (SPARQL) [5], [7], [11]. The operator Ψ injects entropy-aware neural activations into the DT knowledge graph, aligning statistical evidence with contextual knowledge to maintain stability and explainability.

Contributions:

1) A formal *entropy-gated neurosymbolic operator* Ψ for conditional RDF updates, coupling ARNN activations with entropy deviation signals.
2) A *Modular Semantic Update* (MSU) mechanism enabling localised, coherent graph updates without full rewrites.
3) A *hybrid loss function* jointly optimising classification accuracy, graph coherence and semantic alignment, with Lyapunov-style stability guarantees [11].
4) A *directly validated* prototype on a containerised NTP amplification scenario, complemented by a *comparative positioning* against representative IoT/DT IDS baselines from the literature.

Validation boundary: The empirical evaluation reported in this paper is intentionally scoped to a single, fully reproducible scenario—a containerised Network Time Protocol (NTP) amplification attack representative of volumetric zero-day-like behaviour in DT ecosystems. The reported figures (AUC, FPR, latency, MSU update load) refer to this scenario; cross-baseline comparisons against Kitsune [9], WiseML 2024, Sec4ML 2023 and the (H-DIR)^2 predecessor [5] are based on values reported in their respective publications. Broader empirical replication on out-of-distribution (OOD) variants of these benchmarks, on multi-vector RPL attacks (e.g., Destination Advertisement Object–DODAG Information Object, DAO–DIO, manipulation [10]), and on heterogeneous edge hardware is identified as future work in Section V-D and Section VI.

Relevance to FSOFT: Ψ-Risk-DT addresses several core FSOFT 2026 themes: *resilient digital transformation* through fault-tolerant DT defence; *trustworthy cyber-physical software* via explainable, traceable reasoning paths; *DT ecosystem and IoT security*; *adaptive threat detection* under concept drift; and *explainable neurosymbolic AI* as an enabler of operator-facing forensic analysis.

Paper organisation: The remainder of the paper is structured as follows. Section II reviews related work on IoT/DT anomaly detection, neurosymbolic AI and semantic reasoning. Section III introduces the proposed Ψ-Risk-DT framework, including the threat model, the operator Ψ, the MSU mechanism and the hybrid loss with stability analysis. Section IV details the experimental setting and the evaluation protocol, including the explicit validation scope. Section V reports and discusses the empirical results, hyperparameter sensitivity, resource feasibility and limitations. Section VI concludes and outlines future work. Appendix A consolidates the mathematical refinements.

II. Related Work

Anomaly detection in IoT/DT: Early IDSs relied on static rules and statistical profiling, with limited generalisation under distributional drift [2], [6]. Deep learning improved accuracy but introduced opacity and high computational cost. Lightweight approaches [3] trade accuracy for efficiency while remaining semantically opaque. In RPL-based networks, ARNN models have shown promise for probabilistic attack propagation [7], [8], but lack semantic traceability. Benchmarks such as Kitsune [9], WiseML 2024 and Sec4ML 2023 are widely used in the literature to quantify false-positive rates and zero-day resilience under varied attack mixtures [10]; in this paper, they are referenced as *comparative literature baselines* and as *future benchmark targets* (Section IV-A), not as datasets directly used to train or test Ψ-Risk-DT.

Neurosymbolic AI for cybersecurity: Neurosymbolic AI seeks to merge neural adaptability with symbolic interpretability [4], [11]. Recent frameworks combine deep representations with rule-based reasoning [12], but remain offline or simulation-bound and lack entropy-gated, closed-loop integration with streaming telemetry. Graph-based approaches [9] omit dynamic entropy modelling, and formal protocol reasoning [11] operates without neural feedback.

Semantic reasoning and knowledge graphs: Resource Description Framework (RDF) and Web Ontology Language (OWL), together with the Structured Threat Information eXpression (STIX) and the Malware Information Sharing Platform (MISP) ontologies, support interoperability and explainability in cybersecurity [11], [13]. DT-enhanced frameworks [1] emphasise secure synchronisation but lack real-time neurosymbolic coupling. Static RDF pipelines [6], [14] cannot synchronise neural inference with graph semantics under streaming conditions.

Table I positions Ψ-Risk-DT against representative baselines, highlighting entropy-gated coupling and modular RDF updates as key differentiators. The values reported in the "Explainability", "Real-Time DT" (RT DT), "RPL validation" (RPL val.) and "RDF integration" (RDF int.) columns are

Table I: Comparison of related approaches. Symbols: ✓ = fully supported; × = not supported; ∼ = partial. Values are based on each work's primary publication. For Ψ-Risk-DT, RPL support is architectural; empirical validation on multi-vector RPL scenarios is left to future work.

Approach	Explain.	RT DT	RPL val.	RDF int.
Kitsune (2019) [9]	×	×	×	×
WiseML (2024)	∼	∼	×	×
NeSy-GNN (2023) [4]	✓	×	×	∼
Sec4ML (2022–23)	✓	∼	×	×
Ψ-Risk-DT	✓	✓	∼	✓

based on the respective primary publications and on our own reading of their reported scope, not on a re-execution of their codebases.

III. Proposed Framework: Ψ-Risk-DT

Ψ-Risk-DT is a neurosymbolic advancement of the (H-DIR)^2 framework [1], [5], extending entropy-based anomaly detection to the specific challenges of DT ecosystems. The architecture separates the semantic control plane (RDF store, messaging) from the batch-oriented data plane (DT execution, simulation), and retains the six-stage (H-DIR)^2 pipeline—RDF serialisation, entropy-based selection, vectorisation, ARNN inference, semantic injection and dynamic feedback—while introducing three key innovations: the neurosymbolic operator Ψ, the Modular Semantic Update (MSU) mechanism, and a hybrid loss function with Lyapunov-style stability guarantees [11]. Deployment on Apache Spark supports horizontal and vertical scaling within the distributed framework of (H-DIR)^2 [5].

A. Threat Model and Digital Twin Specificity

Unlike standard IoT networks, a DT ecosystem relies on continuous bidirectional synchronisation between physical assets and virtual replicas. A zero-day threat in this context is an event that exploits the *semantic gap* between physical reality and virtual expectation—not just an unknown malware signature. A high-entropy traffic burst may represent a legitimate firmware transfer during maintenance, while a slow sensor-drift attack may be statistically invisible yet fatal to synchronisation logic.

Our framework targets this desynchronisation risk by employing Shannon entropy as a proxy for epistemic uncertainty. As established in recent neurosymbolic literature [11], entropy robustly detects out-of-distribution samples that deviate from the learned synchronisation manifold. Zero-day detection is thus defined as a two-step process: (i) a *statistical trigger*, where an entropy spike flags an anomaly candidate, and (ii) a *semantic validation*, where the operator Ψ queries the RDF knowledge graph. If the graph explains the deviation (e.g., an asset registered in a maintenance state), the alert is suppressed as a benign burst; otherwise, the event is classified as a semantic desynchronisation attack and mitigation procedures are triggered. The directly validated instance of this threat model in the present paper is the containerised NTP amplification scenario detailed in Section IV-A.

B. Neurosymbolic Operator Ψ

The operator Ψ formalises the coupling between entropy-driven neural inference and semantic graph updates:

$$\Psi : (\Delta H_t, G_t) \longmapsto (a_t, G_{t+1}). \tag{1}$$

a) Entropy signal: At time t, the entropy deviation is $\Delta H_t = H(X_t) - H_{\text{baseline}}$, with $H(X_t) = -\sum_{x_i} P(x_i)\log_2 P(x_i)$, computed over a multimodal DT feature set X_t (IP addresses, protocol flags, sensor values). Unlike (H-DIR)^2, ΔH_t incorporates physical–virtual signals, capturing concept drift due to DT mismatches. Deviations above a dynamic threshold τ_s (e.g., $\tau_s = 0.15$) are anomaly candidates.

b) Neural activation: The ARNN activation vector is

$$a_{t+1} = f(W\, a_t + b + x_t), \qquad f(z) = \frac{1}{1+e^{-z}}, \tag{2}$$

where W denotes trainable weights, b a bias, and $x_t = \phi(S_{\Delta t})$ the entropy-weighted input embedding. The selection function $S_{\Delta t} = \{i : |\Delta H_t(i)| > \tau_s(i)\}$ retains only entropy-significant features [7]. Thresholding ($\theta = 0.7$) yields probabilistic attack flags: $P_{\text{attack}} = \sigma(a_t) > \theta$.

c) Update rule: The RDF graph update is gated by both entropy and activation:

$$G_{t+1} = \begin{cases} G_t \cup \Psi_{\text{inject}}(a_t) & \text{if } \Delta H_t > \tau_s \wedge a_t > \theta, \\ G_t & \text{otherwise,} \end{cases} \tag{3}$$

where Ψ_{inject} maps activations into RDF triples aligned with STIX/MISP ontologies [13]. This conditional update reduces unnecessary graph rewrites by up to $88\%^{[\text{M}]}$ on the NTP amplification trace described in Section IV-A.

d) Stability: Discrete-time analysis yields the Lyapunov decrease condition [11]

$$L_{t+1} - L_t \;\le\; -\epsilon\, \|a_t - \theta\|^2, \qquad \epsilon > 0, \tag{4}$$

ensuring bounded convergence under gated updates. ARNN inference scales as $\mathcal{O}(|E|)$ on sparse graphs; RDF updates scale as $\mathcal{O}(r)$ for r symbolic rules.

C. Modular Semantic Update

Most DT cybersecurity approaches treat the twin as a passive replica [1]. MSU instead embeds security reasoning within the synchronisation loop by partitioning G_t into component-level subgraphs:

$$\text{MSU} : G_t \;\mapsto\; \{\, G_t(i) \mid i \in C \,\}, \tag{5}$$

$$C \;=\; \{\,\text{nodes},\ \text{sensors},\ \text{flows},\ \text{links},\ \text{controllers}\,\}. \tag{6}$$

For each subgraph $G_t(i) = (V_i, E_i)$, the operator is applied locally: $\Psi_{\text{local}}(i) : (\Delta H_t(i), G_t(i)) \mapsto (a_t(i), G_{t+1}(i))$, where $\Delta H_t(i) = H(X_t(i)) - H_{\text{baseline}}(i)$ is the component-specific entropy deviation. Graph $G_{t+1}(i)$ is updated only if $\Delta H_t(i) > \tau_s(i) \wedge a_t(i) > \theta(i)$.

a) Coherence and complexity: Semantic consistency is measured as $\mathrm{Coh}(G_t(i)) = 1 - \frac{\#\text{inconsistent triples}}{\#\text{total triples}}$, with minimum threshold $\kappa = 0.95$. Updates follow $P(\text{update}_i) = \sigma\big(\Delta H_t(i) + a_t(i) - \tau_s(i) - \theta(i)\big)$, with convergence analysed via modular Lyapunov functions:

$$L_i = \alpha \, \|\Delta H_t(i)\|^2 + \beta \left(1 - \mathrm{Coh}(G_t(i))\right)^2. \tag{7}$$

MSU reduces the update cost from $\mathcal{O}(r)$ to $\mathcal{O}(r_i)$ per subgraph ($r_i \ll r$). *Empirical scope*: on the NTP amplification trace this yields a maximum symbolic-update load reduction of 88%[M] relative to a global-rewrite ablation on the same trace (Section V-A). Quantitative validation of the expected packet-delivery improvement (target range 90–96%[F]) on multi-vector RPL scenarios such as DAO–DIO [10] is reserved for future work (Section V-D).

D. Hybrid Loss Function

Ψ-Risk-DT optimises a three-term loss:

$$\mathcal{L}_{\text{total}} = \alpha \, \mathcal{L}_{\text{cls}} + \beta \, \mathcal{L}_{\text{graph}} + \gamma \, \mathcal{L}_{\text{sem}}, \quad \alpha + \beta + \gamma = 1, \tag{8}$$

with values $(\alpha, \beta, \gamma) = (0.4, 0.3, 0.3)$ obtained via grid search over the simplex $\{\alpha + \beta + \gamma = 1\}$ [4]. The components are:

- $\mathcal{L}_{\text{cls}}$: cross-entropy for zero-day classification, $\mathcal{L}_{\text{cls}} = -\sum_{(x,y)} y \log \hat{y}$, $\hat{y} = \sigma(a_t)$.
- $\mathcal{L}_{\text{graph}} = \|W_t - W_{\text{gt}}\|_F^2$: structural consistency between the ARNN propagation matrix and a teacher graph $W_{\text{gt}} = \mathcal{M}(G_t)$ [7].
- $\mathcal{L}_{\text{sem}} = \sum_{(s,p,o) \in T_t} \|\mathrm{emb}(a_t) - \mathrm{emb}((s,p,o))\|_2^2$: semantic alignment between neural activations and RDF triples (embeddings via RDF2Vec).

a) Entropy-aware modulation: To emphasise non-stationary regimes, the loss is modulated:

$$\hat{\mathcal{L}}_{\text{total}} = \mathcal{L}_{\text{total}} \cdot (1 + \lambda \, \Delta H_t), \qquad \lambda \leq 0.2, \tag{9}$$

activated when $\Delta H_t > \tau_s$, with $\lambda = 0.1$ to prevent gradient instability [8].

b) Stability: A Lyapunov functional [11]

$$V(\vartheta) = \alpha \, \|y - \hat{y}\|^2 + \beta \, \|W_t - W_{\text{gt}}\|_F^2 + \gamma \, (1 - \mathrm{Coh}(G_t))^2 \tag{10}$$

decreases monotonically under gradient descent if the learning rate satisfies $\eta < 2/L$, where L is the Lipschitz constant of the gradient. Training minimises $\hat{\mathcal{L}}_{\text{total}}$ via Adam, with backpropagation through the Ψ interface jointly optimising all three terms.

IV. Experimental Setting

The operator Ψ is implemented as a six-stage deterministic pipeline that processes streaming data in micro-batches over Apache Spark 3.5.0: (1) RDF serialisation, (2) entropy-based selection, (3) vectorisation, (4) ARNN inference, (5) semantic injection, and (6) dynamic update with feedback. The pipeline extends the distributed framework of (H-DIR)^2 [5], [7].

Symbolic reasoning leverages STIX 2.1 and MISP ontologies [13]. Bidirectional ARNN–graph communication is handled via RESTful interfaces (PyTorch 2.x for neural inference; Apache Jena 4.x for RDF storage and SPARQL execution), configured with Open Authorization (OAuth) 2.0 and Transport Layer Security (TLS) 1.3. The full implementation, SPARQL rules, REST API specifications and reproducibility scripts are released in the companion repository [18].

A. Evaluation Protocol and Validation Scope

This subsection makes the validation protocol explicit, in line with the reviewers' request for tighter alignment between empirical claims and the actual experiment.

(1) Validated scenario: The directly validated scenario is a containerised NTP amplification attack, treated as a representative volumetric zero-day-like threat against a DT ecosystem. NTP amplification is selected because it (i) produces strong, well-characterised entropy spikes at the network layer; (ii) is reproducible deterministically inside Docker; and (iii) stresses the entropy-gated activation path of Ψ end-to-end.

(2) Trace source and traffic construction: Network traces are generated on a Docker Compose topology comprising one victim NTP server, one observer (the DT virtual replica), and a configurable set of reflector and attacker containers. *Benign* traffic is generated by replaying timing-query patterns drawn from the (H-DIR)^2 baseline corpus [5], extended with low-rate periodic firmware-update bursts to emulate maintenance traffic. *Malicious* traffic is produced by orchestrated `monlist`-based amplification campaigns, with attacker rates scanned over five orders of magnitude. Captures are collected via `tcpdump`/`pcap` and pre-processed into per-window flow records.

(3) Feature extraction: From each one-second observation window we extract: source/destination IP entropy, port entropy, packet-size entropy, protocol-flag distributions, inter-arrival time statistics and DT-side state-mismatch indicators (asset registration vs. observed sender). The resulting feature vector feeds both the entropy module and the ARNN.

(4) Entropy baseline computation: The baseline H_{base} is estimated on benign-only windows using a sliding-window Exponentially Weighted Moving Average (EWMA) with decay 0.95, bootstrapped over the first 10% of the benign trace. ΔH_t is then computed as in Eq. (11) and normalised to $[0, 1]$.

(5) Threshold calibration: The entropy threshold $\tau_s = 0.15$ is set at approximately 1σ above the mean baseline entropy, consistent with reported IoT entropy thresholds of 0.10–0.20 [15]. The ARNN activation gate $\theta = 0.7$ is selected by Receiver Operating Characteristic (ROC) analysis on a held-out portion of the benign+malicious training partition, reflecting the asymmetric DT cost structure where missed zero-days dominate the loss [7].

(6) Splits and runs: The labelled corpus is partitioned chronologically into train/validation/test = $60/20/20\%$. Reported figures are the mean over $N = 10$ independent runs with different random seeds for ARNN initialisation and Spark micro-batch ordering; 95% confidence intervals are within ± 0.4 percentage points on AUC and within ± 1.2 ms on latency.

(7) Metrics: We report Area Under the ROC Curve (AUC), False-Positive Rate (FPR) at the operating point θ, end-to-end pipeline latency in milliseconds, semantic traceability (binary: existence of a SPARQL-recoverable explanation chain for each alert), and MSU update load measured as the ratio of triples rewritten per detection event over a global-rewrite baseline.

(8) Software/hardware stack: Experiments run on Apache Spark 3.5.0, PyTorch 2.x, Apache Jena 4.x and Docker 25.x, on a workstation-class host (Intel Xeon-class CPU, 32 GB RAM, no GPU). Edge feasibility on Raspberry Pi 4 / Jetson Nano is discussed architecturally in Section V-C and is left to future device-level benchmarking.

(9) Role of Kitsune, WiseML 2024 and Sec4ML 2023: These benchmarks are *not* used as direct training or test data in the present experiment. They are referenced in two distinct roles: (a) as *comparative literature baselines*, whose AUC, FPR and latency ranges—as reported in their original publications—are placed alongside our measurements in Table II for positional context only; and (b) as *future benchmark targets* for OOD replication of the Ψ-Risk-DT pipeline (Sections V-D and VI).

V. Results and Discussion

This section reports figures obtained within the validation scope of Section IV-A and discusses their positioning against the literature. To preserve claim–evidence alignment, every quantitative statement is annotated with one of the following tags: [M] directly measured on the NTP scenario; [L] reproduced from prior work; [C] comparative claim derived from [M]+[L]; [A] architectural estimate; and [F] future target.

A. Detection Performance

Within the validated scenario of Section IV-A, Ψ-Risk-DT attains an AUC of $0.993^{[M]}$ and an FPR of $0.021^{[M]}$ at the calibrated operating point ($\tau_s = 0.15$, $\theta = 0.7$). The mean ARNN inference time is $15.3\,\text{ms}^{[M]}$, and the end-to-end six-stage pipeline latency is $25.8\,\text{ms}^{[M]}$. The MSU mechanism reduces the number of rewritten RDF triples per detection event by up to $88\%^{[M]}$ with respect to a global-rewrite ablation evaluated on the same trace. All four figures are directly measured on the containerised NTP amplification scenario; they have not yet been established on Kitsune, WiseML 2024 or Sec4ML 2023.

Table II positions these measurements against representative IoT/DT IDS baselines whose AUC, FPR and latency ranges are reproduced from the cited publications.

Discussion: The NTP amplification experiment indicates that entropy-gated neurosymbolic feedback is able to reach AUC values comparable to or exceeding those reported for deep-learning baselines on similar volumetric tasks (typically 0.92–0.96 [9], [2]), while keeping FPR low under a non-stationary regime[C]. The end-to-end latency of 25.8 ms—an order of magnitude below the sub-second threshold commonly assumed for real-time DT synchronisation—is consistent with the design intent of the entropy gate, which acts as a computational filter that avoids unnecessary neural and symbolic processing on benign traffic[M][C]. Compared with the (H-DIR)^2 predecessor [5], Ψ-Risk-DT improves latency by roughly an order of magnitude in the same pipeline geometry (25.8 ms vs. approximately 247 ms[L], i.e. a factor of about $9.6\times$[C]), primarily through MSU-driven update locality. These results suggest, but do not yet prove, generalisability across broader DT attack classes[F]; replication on OOD variants of Kitsune [9], WiseML 2024 and Sec4ML 2023, and on multi-vector RPL scenarios, is reserved for future work (Sections V-D and VI). The frequently cited target range of 85–95% detection rate on broader OOD attack classes is therefore reported here as a future target[F] rather than as a measured outcome.

Table II: Quantitative positioning of Ψ-Risk-DT against representative IoT/DT IDS baselines. Values for Ψ-Risk-DT are directly measured on the containerised NTP amplification scenario of Section IV-A (mean over 10 runs). Values for the other methods are reproduced from the cited publications and serve as comparative literature baselines. "–" = not reported.

Method	AUC	FPR	Lat. (ms)	Sem. trace.
Kitsune [9]	0.92–0.96	0.05–0.08	∼350	×
DL zero-day [2]	∼0.95	0.04–0.06	>500	×
NeSy-IDS [4]	∼0.93	∼0.04	offline	partial
(H-DIR)^2 [5]	0.978	–	∼247	partial
Ψ-Risk-DT	**0.993**[M]	**0.021**[M]	**25.8**[M]	✓

Table III: Key hyperparameters of Ψ-Risk-DT.

Param.	Value	Role
τ_s	0.15	Entropy deviation threshold
θ	0.7	ARNN activation gate
α, β, γ	$0.4, 0.3, 0.3$	Hybrid loss weights ($\sum = 1$)
λ	0.1	Entropy-aware loss modulation

B. Hyperparameter Sensitivity

Table III lists the key hyperparameters. The entropy threshold $\tau_s = 0.15$ was selected on (H-DIR)^2 NTP traces [2], [5], representing approximately 1σ above the mean baseline entropy—consistent with IoT entropy thresholds of 0.10–0.20[L] reported in [15]. The ARNN activation threshold $\theta = 0.7$ was derived from ROC analysis on the NTP training partition, reflecting the asymmetric DT cost structure where false negatives (missed zero-days) are costlier than false positives (unnecessary RDF updates). As shown by Nakıp and Gelenbe [7], ARNN activation distributions exhibit well-separated benign/malicious modes, supporting $\theta \in [0.6, 0.8]$[L].

The loss weights $(\alpha, \beta, \gamma) = (0.4, 0.3, 0.3)$ were determined via grid search over the simplex $\{\alpha + \beta + \gamma = 1\}$. Sensitivity analysis on the NTP partition shows AUC variations smaller than 2%[M] for ± 0.1 perturbations in any single weight; FPR is most sensitive to γ (semantic alignment), confirming that semantic consistency acts as the primary regulariser against false positives. Bayesian optimisation of these weights is a planned refinement for deployment-specific tuning [16]. The

modulation gain $\lambda = 0.1$ ensures that entropy spikes amplify the loss by at most 10–15%[M], preventing gradient instability while following the entropy-as-soft-attention paradigm [8]. Fully autonomous threshold calibration via online learning remains an open challenge.

C. Resource Efficiency and Edge Feasibility

The architecture explicitly decouples offline training (cloud/server) from lightweight inference to enable edge deployment.

- **Entropy-based OOD filter.** Shannon entropy acts as an $\mathcal{O}(1)$[A] filter on IoT nodes, discarding in-distribution traffic before the neural layer [11].
- **ARNN inference.** $\mathcal{O}(|E|)$[A] on sparse graphs, CPU-bound (no GPU), and expected to be executable on Raspberry Pi 4 or Jetson Nano. The intrinsic ability of the ARNN to estimate attack graphs maps naturally onto DT topologies [7].
- **MSU.** Localises updates to $\mathcal{O}(r_i)$[A] ($r_i \ll r$), minimising radio transmission and preventing "broadcast storms" in RPL networks.
- **Hybrid loss optimisation.** Confined to offline training, with zero impact on edge latency.

These arguments establish architectural feasibility; device-specific benchmarks are left to future empirical validation.

D. Limitations

Three limitations should be acknowledged.

L1: Validation scope (single attack class, single hardware tier): The current empirical validation is restricted to a containerised NTP amplification scenario. This setting deliberately stresses the entropy-gated activation path of Ψ on a volumetric, well-characterised attack, but it does *not* yet exercise four classes of threats and conditions that are highly relevant to DT ecosystems: (i) *multi-vector RPL routing attacks*, including DAO–DIO manipulation, version-number attacks and rank inconsistencies [10], where the diagnostic signal is topological rather than volumetric; (ii) *stealthy, low-amplitude concept drift*, where the entropy deviation per window is below τ_s but accumulates over time, requiring sequential or change-point detection on top of the current single-window gate; (iii) *physical–virtual desynchronisation attacks*, in which the network-layer signal is benign but the digital twin's state estimate diverges from the physical asset (e.g., sensor-spoofing, replay-with-drift, false data injection); and (iv) *heterogeneous edge hardware conditions*, including Raspberry Pi 4 and Jetson Nano deployments, intermittent connectivity and energy-constrained duty cycles. Extending validation to (i)–(iv) and replicating the protocol on OOD variants of Kitsune [9], WiseML 2024 and Sec4ML 2023 constitutes the primary future direction; recent work on continuous learning in DT environments [17] further motivates a longitudinal evaluation protocol.

L2: Fixed entropy and ARNN thresholds: The thresholds $\tau_s = 0.15$ and $\theta = 0.7$ are calibrated once on the benign portion of the NTP corpus and held fixed at inference time. This static calibration is adequate for the validated scenario but is expected to be *scenario-dependent* in heterogeneous DT ecosystems, where baseline entropy varies with topology, duty cycle and traffic mix, and where the optimal activation gate depends on the asymmetric cost of false negatives versus false positives. We therefore foresee three complementary calibration strategies as future work: (a) *online calibration* of (τ_s, θ) via streaming quantile estimation on rolling benign windows; (b) *adaptive thresholding* through Bayesian optimisation over the simplex $\{\alpha, \beta, \gamma, \tau_s, \theta\}$ for deployment-specific tuning [16]; and (c) *drift-aware learning* that retrains or fine-tunes the ARNN under detected concept drift, possibly co-designed with the MSU update locality so that knowledge-graph rewrites and threshold updates remain mutually consistent.

L3: Adversarial manipulation of entropy and semantic signals: The current threat model does not explicitly cover adversaries that craft traffic to keep ΔH_t below τ_s while still inducing semantic desynchronisation, nor does it cover ontology-poisoning attacks against the RDF store. Hardening Ψ against such evasion—e.g., via robust entropy estimators, signed semantic updates and provenance-aware SPARQL reasoning—remains a critical open challenge.

VI. Conclusion and Future Work

This paper introduced Ψ-Risk-DT, a neurosymbolic framework for adaptive zero-day threat detection in Digital Twin (DT) ecosystems. By coupling entropy-driven anomaly detection, an Associated Random Neural Network (ARNN) and RDF/SPARQL semantic reasoning through the formal entropy-gated operator Ψ, the framework bridges statistical sensitivity and symbolic explainability in resource-constrained IoT–DT environments. The Modular Semantic Update (MSU) mechanism localises RDF rewrites to affected DT components, and the hybrid loss aligns classification accuracy, graph coherence and semantic consistency with Lyapunov-style stability guarantees [11].

Within the validation scope of Section IV-A—a containerised NTP amplification scenario representative of volumetric zero-day-like attacks—Ψ-Risk-DT attains an AUC of 0.993, an FPR of 0.021, an end-to-end pipeline latency of approximately 25.8 ms and an MSU update-load reduction of up to 88% relative to a global-rewrite baseline. Comparative positioning against representative IoT/DT IDS baselines from the literature (Kitsune [9], deep-learning zero-day detectors [2], NeSy-IDS [4] and (H-DIR)^2 [5]) suggests order-of-magnitude latency gains and a 20–40% FPR reduction relative to deep-learning baselines under the conditions reported in those works. These results are presented as directly measured on the validated scenario, and as positional with respect to literature-reported figures; broader empirical generalisation is not yet claimed.

Future work: Immediate priorities are: (1) replication of the protocol on OOD variants of Kitsune [9], WiseML 2024

and Sec4ML 2023, and on multi-vector RPL attack scenarios such as DAO–DIO manipulation [10]; (2) device-level benchmarking on heterogeneous edge platforms (Raspberry Pi 4, Jetson Nano) to confirm the architectural feasibility argument of Section V-C; (3) drift-aware online calibration of (τ_s, θ) via Bayesian optimisation and streaming quantile estimation [16]; (4) extension of MSU with Graph Neural Network (GNN)-based partitioning [7], [9] for data-driven semantic modularisation at scale; and (5) adversarial-robustness analysis against semantic-evasion and ontology-poisoning attacks [4]. An explainable interface, Ψ-UI, for interactive threat tracing and forensic analysis accessible to non-technical DT operators, is currently under development as ongoing future work and is *not* part of the empirical contribution of this paper.

Supplementary Material

Due to space constraints, the extended mathematical formalisation—including complete stability proofs, hybrid loss derivations and operator properties—together with deployment scripts, SPARQL rules and REST API specifications, are publicly available in the companion repository [18].

Appendix A
Mathematical Refinements

This appendix consolidates key formal refinements to the definitions in Section III. Extended proofs, derivations and pseudocode are available in the companion repository [18].

a) Hybrid loss refinements: The normalised entropy deviation used in Eq. (9) is

$$\Delta H_t \;=\; \frac{H(X_t) - H_{\text{base}}}{\max(\varepsilon,\, H_{\text{base}})} \;\in\; [0, 1], \tag{11}$$

ensuring scale-invariant modulation. The embedding function for the semantic loss is $\text{emb} : (s, p, o) \mapsto \mathbb{R}^{128}$, obtained via RDF2Vec with random walks of length 3. The discrepancy between activations and RDF triples is

$$\Delta(a_t, G_t) \;=\; \frac{1}{|T_t|} \sum_{(s,p,o)\in T_t} \big\| a_t - \text{emb}((s, p, o)) \big\|_2. \tag{12}$$

b) Operator Ψ refinements: The extended operator signature is $\Psi : (\Delta H_t, x_t, G_t) \mapsto (a_t, G_{t+1})$, where $x_t = \phi(S_{\Delta t})$ is the entropy-weighted embedding. Symbolic injection maps activations into RDF triples:

$$\begin{aligned}\Psi_{\text{inject}}(a_t) = \{(v_i,\ &\texttt{:hasActivation},\ a_t(i)) \mid \\ &v_i \in V,\ a_t(i) > \theta\}.\end{aligned} \tag{13}$$

Discrete-time stability analysis (cf. Eq. (4)) is grounded in the Lyapunov functional $L(\Delta H_t, a_t) = \|\Delta H_t\|^2 + (1 - a_t)^2$, which decreases when $\Delta H_t \le \tau_s$ and $a_t \ge \theta$, under learning rate $\eta < 2/L$ with L the Lipschitz constant of the gradient [11]. Graph coherence is measured as $\text{Coh}(G_t) = 1 - \frac{\#\text{inconsistent triples}}{\#\text{total triples}}$, with minimum threshold $\kappa = 0.95$.

References

[1] A. Wakili, S. Bakkali, and I. A. Ibrahim, "A digital twin-enhanced cybersecurity framework for IoT in healthcare: Applications in industry 4.0," *Telematics and Informatics Reports*, vol. 20, p. 100254, 2025.

[2] S. I. Berrios Vasquez, P. A. Hermosilla Monckton, D. I. Leiva Muñoz, and H. Allende, "Zero-day threat mitigation via deep learning in cloud environments," *Applied Sciences*, vol. 15, no. 14, 2025.

[3] A. J. Aparcana-Tasayco, X. Deng, and J. H. Park, "A systematic review of anomaly detection in IoT security: towards a quantum machine learning approach," *EPJ Quantum Technology*, vol. 12, no. 1, pp. 1–39, 2025.

[4] A. Bizzarri, B. Jalaian, F. Riguzzi, and N. D. Bastian, "A neuro-symbolic artificial intelligence network intrusion detection system," in *Proc. 33rd Int. Conf. Computer Communications and Networks (ICCCN)*, IEEE, 2024, pp. 1–9.

[5] D. Tosi and R. Pazzi, "(H-DIR)^2: A scalable entropy-based framework for anomaly detection and cybersecurity in cloud IoT data centres," *Sensors*, vol. 25, no. 15, p. 4841, 2025.

[6] D. E. Denning, "An intrusion-detection model," *IEEE Trans. Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.

[7] M. Nakıp and E. Gelenbe, "An associated random neural network detects intrusions and estimates attack graphs," in *Proc. 32nd Int. Conf. Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 2024, pp. 1–4.

[8] F. Ding and C. Luo, "The entropy-based time-domain feature extraction for online concept drift detection," *Entropy*, vol. 21, no. 11, p. 1187, 2019.

[9] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," in *Proc. 25th Annu. Network and Distributed System Security Symp. (NDSS)*, 2018, pp. 1–15.

[10] A. Benmohamed, E. Barka, C. A. Kerrache, Y. Hadjadj-Aoul, and H. Benkraouda, "RPL-based attack detection approaches in IoT networks: review and challenges," *Artificial Intelligence Review*, vol. 57, p. 256, 2024.

[11] A. Samaddar, N. Potteiger, and X. Koutsoukos, "Out-of-distribution detection for neurosymbolic autonomous cyber agents," in *Proc. IEEE 4th Int. Conf. AI in Cybersecurity (ICAIC)*, 2025, pp. 1–9.

[12] S. P. K. Reddy, U. Nagavelli, Y. S. Kiran, C. S. Kondoju, S. Bushmoni, and A. Yashaswi, "Deep learning for zero-day threat detection and mitigation," in *Proc. Int. Conf. IoT Based Control Networks and Intelligent Systems (ICICNIS)*, 2024, pp. 1362–1368.

[13] B. Lourenço, P. Adão, J. F. Ferreira, M. M. Marques, and C. Vaz, "Structuring security: A survey of cybersecurity ontologies, semantic log processing, and LLMs application," *arXiv preprint arXiv:2510.16610*, 2025.

[14] H. Lei, Y. Ge, and Q. Zhu, "ADAPT: A game-theoretic and neuro-symbolic framework for automated distributed adaptive penetration testing," in *Proc. MILCOM 2024 – IEEE Military Communications Conference*, 2024, pp. 7–12.

[15] C. Xiang, S. Zhao, H. Chen, S. Yang, and Z. Zhang, "IoT-IE: An information-entropy-based approach to traffic anomaly detection in IoT," *Wireless Communications and Mobile Computing*, vol. 2021, p. 1828182, 2021.

[16] A. A. Heydari, D. Metivier, and R. Bhatt, "Multi-objective loss balancing for physics-informed deep learning," *Computer Methods in Applied Mechanics and Engineering*, 2025.

[17] A. Gupta, Y. Tian, and F. Zhou, "TWIN-ADAPT: Continuous learning for digital twin-enabled online anomaly classification in IoT-driven smart labs," in *Proc. IEEE Int. Conf. Communications (ICC)*, 2024.

[18] R. Pazzi, D. Facheris, and D. Tosi, "Ψ-Risk-DT — companion repository," 2026. [Online]. Available: https://github.com/DAV-IDE/Alpha123 (accessed Apr. 28, 2026).