A Bio-Inspired Hierarchical Temporal Defense for Securing Spiking Neural Networks Against Physical and Adversarial Perturbations

Sylvester Kaczmarek
Department of Computing
Imperial College London
research@sylvesterkaczmarek.com

Abstract

Spiking Neural Networks (SNNs) offer an energy-efficient paradigm for processing temporal data but are critically vulnerable to perturbations in spike timing, a common issue in physical systems where jitter arises from sensor noise or radiation-induced effects. Traditional defenses adapted from static neural networks often fail to address these unique temporal dynamics. We introduce the Hierarchical Temporal Defense (HTD), a novel, bio-inspired architecture integrating defenses across the input, neuronal, and synaptic levels. Key innovations include probabilistic encoding for jitter tolerance, adaptive thresholds for stability, and gated plasticity for secure learning. Theoretical analysis provides a robustness bound on information loss for the input encoding under jitter. Empirical evaluation on high-fidelity simulated physical data shows the HTD framework reduces the success rate of a strong PGD attack from 82.1% to 18.7% and maintains high performance (F1 > 0.72) under environmental stress, demonstrating a principled methodology for designing robust neuromorphic systems for safety-critical applications.

1 Introduction

Spiking Neural Networks (SNNs) are a compelling computational model for processing data from physical systems, offering unparalleled energy efficiency by leveraging sparse, event-driven computation. This makes them ideal for edge AI applications in resource-constrained environments, with analogous challenges appearing in domains from high-energy physics event detection to seismic sensing. However, the SNN paradigm, which encodes information in the precise timing of discrete spike events, introduces a critical vulnerability: sensitivity to temporal perturbations. In real-world physical systems, such perturbations are common, arising from sensor noise, thermal fluctuations, or high-energy particle strikes causing single-event upsets (SEUs). This sensitivity can also be exploited by intelligent adversaries through gradient-based manipulations.

Existing defenses for neural networks, such as adversarial training or input denoising, are often ill-suited for SNNs, as they do not account for the rich temporal dynamics and the non-differentiable nature of spike generation. A robust SNN requires defenses that are themselves temporally aware and deeply integrated into the network's dynamics. Drawing inspiration from biological neural circuits, where mechanisms like metaplasticity [1] regulate learning based on activity history, we propose the Hierarchical Temporal Defense (HTD), a novel framework for designing SNNs that are inherently resilient to a wide spectrum of perturbations.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences.

2 Related Work

Adversarial robustness in machine learning has focused on static networks, with methods like projected gradient descent (PGD) attacks [3] and defenses such as adversarial training. For SNNs, recent work explores temporal vulnerabilities [6] and basic defenses like noise injection or fixed thresholds [4], but these lack integration for multi-layered protection. A mini taxonomy of SNN defenses includes: encoding-based (e.g., probabilistic for noise, [2]), threshold-based (e.g., adaptive for stability), and plasticity-based (e.g., metaplasticity for homeostasis, [1]). Prior approaches like [5] address SNN attacks but focus on weight perturbations, not temporal jitter. Our HTD unifies these into a hierarchical framework, addressing gaps in securing SNNs for dynamic physical systems.

3 The Hierarchical Temporal Defense (HTD) Framework

HTD secures SNNs through three integrated layers. The full theoretical derivations are provided in our extended work [8].

3.1 Input Layer: Bayesian Spike Pattern Superposition (BSPS)

To make the input representation robust to jitter, BSPS encodes data as a probability distribution over temporal patterns. For a time window of size W, the state is a probability vector $\mathbf{p}(t) \in \mathbb{R}^{2^W}$. While the state space scales exponentially with W, we find that for physical sensor streams, temporal correlations decay rapidly; thus, we limit $W \leq 10$. This maintains computational tractability on edge hardware while capturing sufficient temporal context to distinguish signal from noise. *Principle: We model input uncertainty probabilistically, inspired by measurement principles in physics, to create a representation resilient to discrete perturbations.* The state updates via a smoothed Bayesian rule:

$$p_i(t+1) = (1-\eta)p_i(t) + \eta \left(\frac{P(o(t) \mid s_i)p_i(t)}{\sum_j P(o(t) \mid s_j)p_j(t)} \right) \tag{1}$$

where $P(o(t) \mid s_i)$ is a softened Bernoulli likelihood obtained by clipping empirical spike probabilities to $[\epsilon, 1-\epsilon]$. This clipping prevents zero-probability collapse in the presence of unexpected noise, providing a finite KL-divergence bound on information loss.

Proposition 1 (KL Bound Under Jitter): Under a jitter process that independently flips each spike in the window with probability q and with likelihoods clipped to $[\epsilon, 1-\epsilon]$, the expected KL divergence between consecutive BSPS states is bounded: $\mathbb{E}\big[D_{\mathrm{KL}}(p(t+1) \,\|\, p(t))\big] \leq \log(1/\epsilon) + H(q)$, where $H(q)$ is the Shannon entropy of the jitter distribution. Sketch: we bound the likelihood ratio under jitter using the clipping parameter ϵ and apply standard KL and entropy inequalities; see [8] for the full derivation.

3.2 Neuronal Layer: Homeostatic Adaptive Thresholds

To regulate neuronal excitability against noise bursts, we use adaptive thresholds inspired by spike-frequency adaptation: $\theta(t) = \theta_0 + \alpha \exp(-\beta \cdot \mathrm{ISI})$. We utilize current-based Leaky Integrate-and-Fire (LIF) neurons for the core architecture, with membrane update $u(t+1) = \lambda u(t) + w^\top s_{\mathrm{in}}(t) - \theta(t)\, s(t)$ and a surrogate gradient for the Heaviside spike nonlinearity. Thresholds are clipped to $[\theta_{\min}, \theta_{\max}]$ to guarantee bounded excitability. We set θ_0, α, β from the validation set via grid search. *Principle: We model neuronal homeostasis, a key biophysical mechanism for stability, to dynamically regulate excitability and reject noise bursts, preventing runaway excitation during high-flux radiation events.*

3.3 Synaptic Layer: Volatility-Gated Metaplasticity

To secure online learning, we introduce a metaplasticity rule that gates STDP. The weight update Δw is scaled by $M(t)$, a logistic function of the synapse's recent volatility, $\bar{\Delta w}$:

$$\Delta w' = M(t) \cdot \Delta w \quad \text{where} \quad M(t) = 1/(1 + \exp(\kappa(\bar{\Delta w}(t) - \phi))) \tag{2}$$

We set the EMA parameter such that $\bar{\Delta w}$ integrates over tens to hundreds of updates, so the gate evolves on a slower timescale than individual STDP events. *Principle: We model synaptic metaplasticity, a biological process governing the history-dependence of learning, to create a security gate that stabilizes online adaptation by freezing weights when volatility exceeds nominal bounds.*

4 Experiments and Results

Anomaly detection serves as a proxy for broader temporal tasks like event detection in physics experiments. We evaluated the HTD framework on an anomaly detection task using our high-fidelity, physics-based dataset, which we term CARD (Cislunar Anomaly and Risk Dataset).

Dataset and Protocol: The CARD dataset comprises $> 10^6$ samples split 70/15/15 (train/val/test). Each example is a length-T multivariate time series with C channels and a binary label (nominal vs anomaly). We normalize each channel to $[0, 1]$ and convert amplitudes to Bernoulli spikes with rate equal to the normalized value, shared across HTD and baseline models. The simulator is calibrated and augmented with anonymized telemetry from cislunar mission archives. The simulation incorporates environmental effects such as radiation-induced SEUs (at rates up to 10^{-4}/bit-day). Attacks utilized the SuperSpike surrogate gradient [7] for backpropagation through time. For PGD, the adversary perturbs the continuous input x within an ℓ_∞ ball $\{\tilde{x} : \|\tilde{x} - x\|_\infty \leq \epsilon\}$ with $\epsilon = 0.10$ and step size $\alpha = \epsilon/10$. Temporal jitter operates on the spike train: each spike at time t may be shifted to any $t' \in [t - J, t + J]$ with $J = 3$ ms; collisions are resolved by logical OR. We use a single PGD run per example without multi-start restarts or explicit EOT; this makes our ASR estimates a lower bound on the worst-case. Inference is executed on a BrainChip Akida neuromorphic board; training runs on a single GPU server. The dataset and code will be made publicly available upon publication. To isolate the contribution of our framework, we compare our proposed SNN (with HTD) to an identical **Baseline SNN** (standard LIF neurons with fixed thresholds and unmodified STDP). All hyperparameters were optimized via 5-fold cross-validation.

4.1 Adversarial Robustness

We tested against white-box PGD (L_∞ norm, 20 iterations, step size $\epsilon/10$) and Temporal Jitter attacks. As shown in Table 1, the HTD framework dramatically reduces the Adversarial Success Rate (ASR). All reported mean ± standard deviation values are computed over the five folds.

Table 1: Adversarial Success Rate (%) on the CARD Dataset.

Attack Type	Perturbation	Proposed SNN (HTD)	Baseline SNN
PGD	$\epsilon = 0.10$	18.7 ± 3.2	82.1 ± 4.9
Temporal Jitter	$J = 3$ ms	25.1 ± 4.5	75.8 ± 6.3

4.2 Ablation Study

To dissect the contribution of each defensive layer, we conducted an ablation study under the PGD attack ($\epsilon = 0.1$). Table 2 shows the increase in ASR when each core component is removed. Ablation results confirm that all three levels contribute substantially: removing any single component increases ASR by 13–27%. This supports the view that HTD distributes the defense across input, neuronal, and synaptic levels instead of relying on a single dominant mechanism. All performance drops were found to be statistically significant ($p < 0.01$).

Table 2: Increase in ASR upon Ablation of Core HTD Components.

Ablated Component	Increase in ASR (%)
BSPS Encoding	+27.2
Metaplasticity	+22.5
Adaptive Thresholds	+13.4

4.3 Robustness to Environmental Perturbations

We also evaluated resilience to simulated environmental stress. We defined "High Stress" as a combined condition with radiation-induced SEU rates of 10^{-4}/bit-day and a signal-to-noise ratio (SNR) degraded to 10 dB. Table 3 shows the F1-score under increasing perturbation.

Table 3: F1-Score Performance under Environmental Stress.

Model	No Perturbation	High Stress
Proposed SNN (HTD)	0.87 ± 0.03	**0.72 ± 0.06**
Baseline SNN	0.86 ± 0.04	0.48 ± 0.09

The HTD framework exhibits graceful degradation, maintaining critical functionality where the baseline model collapses. The baseline failure is largely driven by noise-induced spurious spiking, which overwhelms the fixed thresholds; HTD mitigates this via adaptive regulation. This unified defense against both adversarial and natural physical perturbations is essential for reliable deployment in noisy environments like high-energy physics detectors or space robotics.

5 Discussion and Conclusion

This research has introduced and validated the Hierarchical Temporal Defense (HTD), a novel, bio-inspired framework for securing SNNs. Our results demonstrate that by integrating defenses at the input, neuronal, and synaptic levels, it is possible to build neuromorphic systems that are highly resilient to both physical and adversarial perturbations. The ablation results suggest a hierarchical dependency: input-level filtering by BSPS is necessary to prevent adaptive thresholds from tracking high-frequency noise, while stable thresholds are required for the metaplasticity rule to correctly estimate volatility baselines. This "Physics in ML" approach, where principles from biophysics inform the design of a more robust algorithm, provides a significant advancement over traditional, static defense mechanisms and opens new avenues for efficient, secure AI in scientific discovery. While validated on commercial neuromorphic hardware, future work must extend this validation to space-qualified, radiation-hardened platforms and address the computational scaling of the BSPS encoding for larger windows.

References

[1] Abraham, W. C. (2008). Metaplasticity: tuning synapses and circuits for plasticity. *Nature Reviews Neuroscience*.

[2] Bohte, S. M., et al. (2001). Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing*.

[3] Madry, A., et al. (2017). Towards deep learning models resistant to adversarial attacks. *ICLR*.

[4] Marchisio, A., et al. (2020). Is spiking secure? A comparative study on the security vulnerabilities of spiking and deep neural networks. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1-8. IEEE.

[5] Liang, L., et al. (2022). Toward robust spiking neural network against adversarial perturbation. *arXiv preprint arXiv:2205.01625*.

[6] Sharmin, S., et al. (2020). Inherent adversarial robustness of deep spiking neural networks. *arXiv*.

[7] Zenke, F., and Ganguli, S. (2018). SuperSpike: Supervised learning in multilayer spiking neural networks. *Neural Computation*, 30(6), 1514–1541.

[8] Kaczmarek, S. (2025). *A Principled Neuromorphic Framework for Secure, Adaptive, and Interpretable Anomaly Detection in Autonomous Cislunar Robotics*. Unpublished manuscript, Department of Computing, Imperial College London.