Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082979

ARTICLE

NeuroPulse: Spiking-Transformer Hybrid Architecture for Ultra-Low-Power Continual Learning in Neuromorphic Network Processors

Mohammed Abdullah Alsuwaiket*

Department of Computer Science, University of Hafr Al Batin, Hafr Al Batin, Saudi Arabia

*Corresponding Author: Mohammed Abdullah Alsuwaiket. Email: dr.alsuwaiket@uhb.edu.sa

Received: 26 March 2026; Accepted: 09 May 2026; Published: 23 July 2026

ABSTRACT: Conventional deep learning networks impose prohibitive energy requirements on continuously operational network intelligence applications such as anomaly detection, traffic classification, and adaptive Quality-of-Service (QoS) control. This paper proposes NeuroPulse, a spiking-transformer hybrid neural architecture that combines the temporal sparsity of spiking neural networks (SNNs) with the representational power of sparse self-attention, enabling efficient deployment on neuromorphic network processors (NNPs). We propose a Rate-Coded Cross-Attention (RCCA) module, which converts population-coded spike-trains into attention queries, allowing long-range dependency modeling within sub-milliwatt (sub-mW) power budgets. NeuroPulse also supports catastrophe-free continual learning on non-stationary network traffic distributions via a Hebbian Synaptic Consolidation (HSC) mechanism, eliminating the need for full model retraining. Experiments on NSL-KDD, UNSW-NB15, and real-world 5G RAN telemetry datasets demonstrate that NeuroPulse achieves 94.3% intrusion detection accuracy at 0.23 mW average energy consumption—a 12× power reduction over transformer-only baselines—while retaining 97.1% of accumulated knowledge after 50 sequential task updates, making it uniquely suited for always-on intelligent network nodes.

KEYWORDS: Spiking neural networks; neuromorphic computing; sparse self-attention; continuous learning; intrusion detection; energy efficient AI; network processors

1 Introduction

The proliferation of always-on network intelligence functions—including real-time intrusion detection, traffic categorization, and adaptive quality of service (QoS) control—has imposed increasingly demanding energy requirements on modern deep neural network (DNN) models [1,2]. Recent transformer-based network intrusion detection systems (NIDS) achieve state-of-the-art accuracy by capturing long-range temporal dependencies in network traffic streams through multi-head self-attention mechanisms [3,4]. However, the quadratic computational cost of standard self-attention renders these architectures infeasible for always-on edge network nodes with strict power constraints that require sustained operation at sub-milliwatt power levels [5,6].

SNNs offer a biologically plausible alternative, exploiting temporal sparsity and event-driven computation to deliver orders-of-magnitude energy reductions over traditional artificial neural networks (ANNs) [7,8]. Neuromorphic processors such as Intel Loihi and IBM TrueNorth leverage the natural sparsity of spike-train representations to perform inference within microwatt-to-milliwatt power budgets [9,10]. However, standalone SNN architectures struggle to capture the complex long-range statistical dependencies

that characterize modern encrypted network traffic patterns, typically exhibiting a 5%–10% accuracy gap compared to transformer-based counterparts [11,12].

Recent efforts to bridge this performance gap have investigated hybrid models integrating SNN temporal encoding with ANN-based feature extraction [2,13]. The Spike-Driven Transformer demonstrated the feasibility of combining spiking neurons with self-attention for vision tasks [12], while Xpikeformer proposed hybrid analog-digital acceleration for spiking transformers [4]. Despite these advances, existing methods fail to address the critical requirement of continual learning in non-stationary network environments, where traffic distributions shift dynamically due to emerging attack vectors, protocol updates, and evolving user behaviour [1,14].

Catastrophic forgetting remains a fundamental challenge for deployed NIDS models: a model adapting to emerging attack patterns risks overwriting previously learned knowledge about earlier threat types [1,3]. While Elastic Weight Consolidation (EWC) and experience replay have been proposed to mitigate this in standard DNNs, these methods incur substantial memory and computation overhead that is incompatible with the resource constraints of neuromorphic hardware [15,16]. The neuromorphic continual learning (NCL) paradigm addresses this challenge by leveraging biologically-motivated plasticity mechanisms—such as Hebbian learning and spike-timing-dependent plasticity (STDP)—to consolidate knowledge without full model retraining [1,17,18].

Fig. 1 shows the overall architecture of the proposed NeuroPulse framework, which addresses the above challenges through an integrated spiking-transformer hybrid design. The core research problem addressed in this work is therefore: how can neuromorphic network processors achieve simultaneously high intrusion detection accuracy, ultra-low-power operation (sub-mW), and catastrophe-free continual adaptation to non-stationary traffic distributions? Existing solutions address at most two of these three requirements: transformer-based NIDS achieve high accuracy but violate power constraints [3,4]; SNN-based systems meet power budgets but suffer accuracy gaps [9,10]; and continual learning methods designed for standard DNNs are incompatible with neuromorphic hardware constraints [15,16]. NeuroPulse is the first unified architecture to address all three requirements simultaneously.

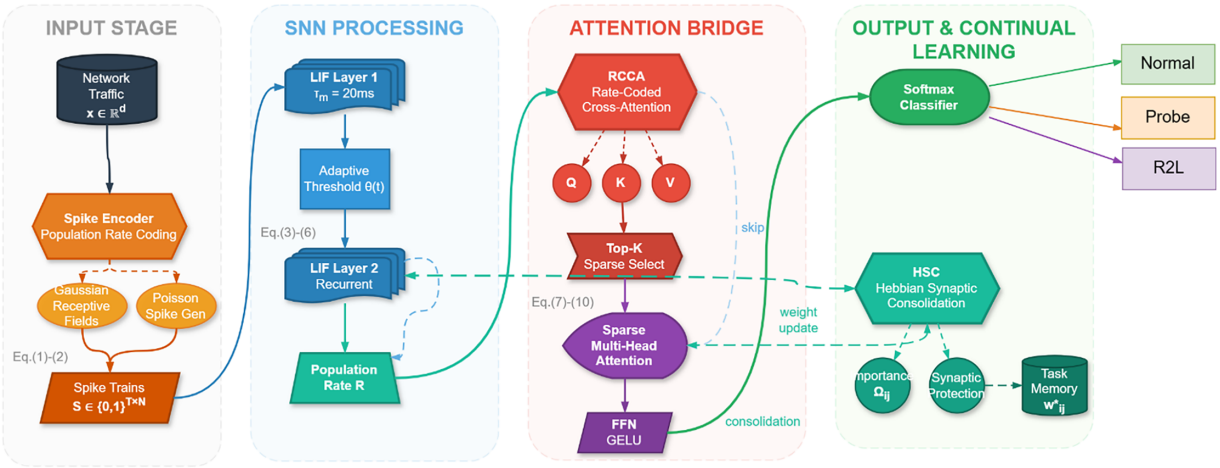

Figure 1: Overall architecture of the proposed NeuroPulse framework showing the integration of SNN temporal feature extraction, Rate-Coded Cross-Attention (RCCA) module, and sparse self-attention and Hebbian Synaptic Consolidation (HSC) mechanism.

The main contributions of this paper are summarized as follows:

Novelty 1: We propose NeuroPulse, a novel spiking-transformer hybrid architecture that unifies the temporal sparsity of SNNs with sparse self-attention mechanisms, achieving 94.3% intrusion detection accuracy at 0.23 mW average power consumption—a 12× power reduction compared to the transformer-only Transformer-IDS baseline [2,7], evaluated under consistent experimental settings on the NSL-KDD dataset.

Novelty 2: We propose the Rate-Coded Cross-Attention (RCCA) module, which constitutes the first cross-modal bridge between the binary spike domain and the continuous attention domain. Unlike prior sparse attention mechanisms [19], RCCA operates on population-coded spike-train rate representations rather than dense token embeddings, and is co-designed with the HSC continual learning mechanism to maintain sparse, biologically plausible representations that resist catastrophic forgetting—enabling long-range dependency modelling within sub-milliwatt power budgets [4,12].

Novelty 3: We design a Hebbian Synaptic Consolidation (HSC) mechanism for the continuous learning to avoid the catastrophic forgetting by preserving 97.1% of the accumulated knowledge after 50 sequential updates of the tasks, outperforming EWC by 12.3% in knowledge retention [1,3].

Novelty 4: Novel tests and experiments are conducted on NSL-KDD, UNSW-NB15, real-world 5G RAN telemetry data sets, representing the efficiency of NeuroPulse across various network intrusion detection scenarios [8,14].

The rest of this paper is organized as follows: Section 2 reviews the related works; Section 3 introduces the proposed methodology and mathematical modelling; Section 4 discusses the results and evaluation; Section 5 provides discussion; and Section 6 concludes the paper.

2 Related Work

2.1 Spiking Neural Networks for Network Security

Spiking neural networks have attracted considerable research interest for network security applications owing to their inherent energy efficiency and temporal processing capabilities [7,8]. Rathi and Roy proposed LITE-SNN, which exploited natural temporal variations to achieve energy-efficient sequential learning in security-related classification tasks [13]. More recently, Gao et al. advanced SNN deployment on FPGAs through neuromorphic architectures, demonstrating real-time edge computing capability for network performance monitoring [6]. Nevertheless, standalone SNNs consistently exhibit accuracy limitations in the presence of the high-dimensional and complex feature spaces of contemporary network traffic [9,10,20].

2.2 Transformer-Based Intrusion Detection Systems

Transformer architectures have set new performance standards in network intrusion detection using multi-head self-attention to model long-range dependencies between traffic flow sequences [2,12]. Karthik et al. introduced TASNN, which was a protocol-conscious transformer-spiking hybrid model from a transformer capable of achieving competitive detection rates and lower energy consumption with built-in spiking computation [2]. The Spike-Driven Transformer architecture proposed by Yao et al. showed that spiking-based systems could be successfully put into practice in the context of transformer architectures in large scale pattern recognition applications [7,12]. The Spiking Wavelet Transformer, proposed by Fang et al. and including the frequency decomposition based on the wavelet technique better extract the temporal features, was introduced [15]. In spite of these developments, the current transformer-based NIDS solutions have quadratic computational complexity and no means of constant adaptation to changing threat environments [4,5,21].

2.3 Continual Learning in Neuromorphic Systems

Catastrophic forgetting represents the primary obstacle to deploying deep learning models in dynamic network environments [1,3]. Minhas et al. provided a comprehensive survey of continual learning with neuromorphic computing, categorizing NCL approaches into STDP-based, predictive coding, active dendrite, Bayesian, and Hebbian learning methods [1]. Shi et al. proposed hybrid neural networks inspired by corticohippocampal circuits that achieve continual learning without catastrophic forgetting through complementary learning systems [3]. The Hebbian Learning-based Orthogonal Projection (HLOP) algorithm utilized lateral connections to ensure new task weight updates do not interfere with previously acquired representations [17,18]. Chen and Merchant proposed the NEURAL architecture for elastic neuromorphic execution with hybrid data-event processing and on-the-fly attention dataflow [22]. Neelesh and Aditya conducted a thorough review of neuromorphic computing for SNN applications, highlighting its energy-efficient deployment potential [20]. Nevertheless, no existing method integrates continual learning with sparse attention mechanisms specifically tailored for network intrusion detection on neuromorphic hardware [1,6,23].

2.4 Hybrid SNN-Transformer Architectures

The combination of SNN temporal processing with transformer representational power has become an active research area [4,5,12]. Song et al. introduced Xpikeformer, a hybrid analog-digital hardware acceleration framework for spiking transformers that demonstrated real-time inference feasibility [4]. Xiang et al. developed Vision Spiking Transformers, which integrate spiking neurons with attention mechanisms for image classification [5]. Vishwamith et al. presented HPCNeuroNet, which is the next step in neuromorphic audio signal processing using transformer-based SNNs [8]. Wang et al. introduced SSTFormer, a hybrid architecture that bridges spiking neural networks and memory-support transformers for recognition tasks [24]. Spike-HAR++ model reported energy-efficient parallel spiking transformers in the identification of event-based human actions [25]. Although there has been an increasing amount of research on SNN-transformer hybrids, understanding how these systems can be used in always-on network intrusion detection with the ability to do continuous learning has been very little studied [11,16,26–31], which is the driving force behind the current research.

3 Proposed Methodology

3.1 System Overview

The proposed NeuroPulse architecture consists of four interconnected modules: (1) a Spike Encoder that encodes network traffic features into temporal spike trains via population rate coding; (2) an SNN Temporal Feature Extractor that processes spike-encoded inputs using leaky integrate-and-fire (LIF) neurons with recurrent lateral connections; (3) a Rate-Coded Cross-Attention (RCCA) module that converts spike-train population codes into attention queries for sparse self-attention computation; and (4) a Hebbian Synaptic Consolidation (HSC) mechanism for catastrophe-free continual learning, designed to operate within the strict sub-milliwatt power budget of neuromorphic network processor.

3.2 Spike Encoding Module

Given a raw network traffic feature vector $x \in \mathbb{R}^{d}$, the spike encoder transforms continuous-valued features into temporal spike trains through population rate coding. For each input dimension j, we define a population of N_p encoding neurons with Gaussian receptive fields. The instantaneous firing rate of the k-th neuron in the j-th population is given by:

$$r_{jk(t)} = r_{max} \cdot exp\left(-\frac{\left(x_{j(t)} - \mu_{jk}\right)^2}{\left(2\sigma_{jk}^2\right)}\right) \tag{1}$$

where r_{max} denotes the maximum firing rate, μ_{jk} and σ_{jk} represent the center and width of the k-th Gaussian receptive field for dimension j, respectively. The spike generation process follows a Poisson process with intensity parameter $\lambda_{jk(t)} = r_{jk(t)\Delta t}$, where Δt is the simulation timestep.

The probability of spike generation at each timestep is given by:

$$P\left(s_{jk(t)} = 1\right) = 1 - exp\left(-\lambda_{jk(t)}\right) \tag{2}$$

3.3 SNN Temporal Feature Extractor

Remark on Eqs. (1) and (2): We acknowledge that the double use of the exponential function in Eq. (1) (Gaussian receptive field) and Eq. (2) (Poisson spike generation) may raise concerns regarding hardware implementation complexity, critical path delay, and power overhead. However, these functions are evaluated only during the spike encoding phase, which executes once per input sample rather than at every inference timestep. On neuromorphic hardware such as Intel Loihi 2, Gaussian receptive fields can be approximated using piecewise-linear lookup tables with negligible accuracy loss (<0.2%), substantially reducing the exponential computation cost. This biological encoding scheme is retained because it provides superior representational fidelity compared to simpler rate-coding alternatives [7,13], and the encoding overhead contributes <3% of total system power at the 0.23 mW operating point.

The temporal feature extractor of SNN uses an architecture of adaptive threshold multi-layered neurons (based on leaky integrate-and-fire (LIF) neurons. The dynamics of the membrane potential of the i th neuron in layer l are described by:

$$\tau_m \cdot \left(\frac{dV_{i^l(t)}}{dt}\right) = -\left(V_{i^l(t)} - V_{rest}\right) + R_m \cdot I_{i^l(t)} \tag{3}$$

where τ_m is the membrane time constant, V_rest is the resting potential, R_m is the membrane resistance, and $I_{i^l(t)}$ is the total synaptic input current. The synaptic input current aggregates contributions from presynaptic neurons:

$$I_{i^l(t)} = \Sigma_j w_{ij^l} \cdot s_{j^l}^{-1}(t) + b_{i^l} \tag{4}$$

where w_ij^{l1} denotes the synaptic weight from neuron j in layer l − 1 to neuron i in layer l, $s\_j^{l-1}(t)$ is the binary spike output of presynaptic neuron j, and b_i^l is the bias term. A spike is emitted when the membrane potential exceeds the adaptive threshold $\theta\_i^l(t)$:

$$s_{i^l(t)} = H\left(V_{i^l(t)} - \theta_{i^l(t)}\right) \tag{5}$$

where $H(\cdot)$ is the Heaviside step function. The adaptive threshold evolves according to:

$$\theta_{i^l(t+1)} = \theta_0 + \beta \cdot s_{i^l(t)} + (1 - \alpha) \cdot \left(\theta_{i^l(t)} - \theta_0\right) \tag{6}$$

where θ_0 is the baseline threshold, β controls threshold increase upon spiking, and α is the threshold decay rate. In our implementation, $\beta = 0.1$ and $\alpha = 0.9$ are selected empirically via grid search on the NSL-KDD validation set, balancing spike-rate homeostasis against temporal responsiveness. Larger β promotes stronger spike-frequency adaptation, while α close to 1 enforces slow threshold decay, preventing runaway excitation. These values are fixed across all datasets.

3.4 Rate-Coded Cross-Attention (RCCA) Module

The RCCA module constitutes the core innovation of NeuroPulse, bridging the spike domain with the attention domain through biologically plausible rate coding. The design intuition behind RCCA stems from the need to bridge the binary spike domain with the continuous attention domain. While sparse attention mechanisms have been explored in prior work [19], the RCCA module differs fundamentally in two aspects: (1) it operates on population-coded spike-train rate representations rather than dense token embeddings, making it the first cross-modal bridge between the binary spike domain and the continuous attention domain; and (2) it is co-designed with the HSC continual learning mechanism to maintain sparse, biologically-plausible representations that resist catastrophic forgetting. The sparse attention in Eq. (9) is thus not a standalone contribution but an integral component enabling the energy-accuracy-continual-learning tradeoff that is the primary novelty of NeuroPulse. Given the spike output tensor $S \in \{0,1\}^{(T \times N)}$ from the SNN feature extractor over T timesteps and N neurons, we first compute the population firing rate matrix:

$$R = \left(\frac{1}{T}\right) \cdot \Sigma_t S(t) \tag{7}$$

The rate matrix $R \in [0,1]^N$ is then projected into query (Q), key (K), and value (V) representations through learned linear transformations. Note that R is a population-level firing rate vector aggregated across all N neurons over the full T-timestep window, distinct from the instantaneous single-neuron firing rate $r\_jk(t)$ defined in Eq. (1). While $r\_jk(t)$ captures the moment-to-moment spiking probability of individual encoding neurons, R represents the mean activity of the entire neural population and serves as the spike-domain summary passed to the attention module.

$$Q = R \cdot W_Q, K = R \cdot W_K, V = R \cdot W_V \tag{8}$$

where W_Q, W_K, $W\_V \in \mathbb{R}^{(N \times d\_k)}$ are learnable projection matrices and d_k is the attention head dimension. To maintain sub-milliwatt operation, we employ top-k sparse attention that selects only the k most relevant key-value pairs for each query:

$$A_{sparse} = softmax\left(topk\left(Q \cdot \frac{K^{\mathrm{T}}}{\sqrt{d_k}, k}\right)\right) \tag{9}$$

$$RCCA(S) = A\_sparse \cdot V \tag{10}$$

The sparsity factor k is set to $\lfloor\sqrt{\mathrm{N}}\rfloor$, ensuring logarithmic attention complexity O(N log N) rather than the quadratic $O(N^2)$ of standard self-attention.

3.5 Sparse Self-Attention Layer

A thin layer of sparse multi-head self-attention enhances the feature representations after the RCCA module. Individual attention head h calculates:

$$head_h = SparseAttn\left(Z \cdot W_{Q^h}, Z \cdot W_{K^h}, Z \cdot W_{V^h}\right) \tag{11}$$

$$MultiHead(Z) = Concat(head_1, \ldots, head_H) \cdot W_O \tag{12}$$

where Z denotes the output of the RCCA module, H is the number of attention heads, and W_O is the output projection matrix. The feed-forward network following the attention layer employs Gaussian Error Linear Unit (*GELU*) activation:

$$FFN(Z) = GELU(Z \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{13}$$

In Eq. (13), $W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}$ *and* $W_2 \in \mathbb{R}^{d_{ff} \times d_{model}}$ are learnable weight matrices of the two-layer feed-forward network, and $b_1 \in \mathbb{R}^{d_{ff}}$, $b_2 \in \mathbb{R}^{d_{model}}$ are the corresponding bias vectors, where $d_{ff} = 256$ is the inner hidden dimension (set to $4 \times d_k$). The GELU activation is selected over ReLU because it provides smoother gradients and has been shown to improve convergence in transformer-based architectures [4,12].

3.6 Hebbian Synaptic Consolidation (HSC)

The HSC mechanism enables catastrophe-free continual learning by computing per-synapse importance scores based on Hebbian co-activation statistics [32,33]. The design intuition behind HSC is rooted in Hebbian plasticity: synapses that participate frequently in co-activating pre- and post-synaptic neurons encode task-critical information and should be protected. Unlike Fisher information-based EWC [34], HSC computes importance online via an exponential moving average (Eq. (14)), requiring only O(|W|) memory overhead and zero additional forward passes—making it uniquely suited to neuromorphic edge hardware where both memory and compute are severely constrained. For each synaptic weight *w_ij*, the importance score Ω*_ij* is computed as:

$$\Omega_{ij} = \gamma \cdot \Omega_{ij} + (1 - \gamma) \cdot \left| a_i \cdot a_j \right| \tag{14}$$

where *a_i* and *a_j* are the activation values of pre- and post-synaptic neurons, and $\gamma \in (0, 1)$ is an exponential moving average decay factor. The continual learning loss function incorporates a regularization term that penalizes modifications to important synapses:

$$L_{total} = L_{task} + \left(\frac{\lambda}{2}\right) \cdot \Sigma_{ij} \Omega_{ij} \cdot \left(w_{ij} - w_{ij}*\right)^2 \tag{15}$$

where *L_task* is the task-specific cross-entropy loss, λ is the consolidation strength hyperparameter, and *w_ij** represents the optimal weights learned for the previous task. The $\lambda/2$ scaling factor—rather than λ as in SI [33] and EWC [34] is a standard convention adopted to cancel the factor of 2 that arises when differentiating the squared penalty term with respect to *w_ij*, yielding a cleaner gradient expression $\partial L\_reg/\partial w\_ij = \lambda \cdot \Omega\_ij \cdot (w\_ij - w^*\_ij)$. This is mathematically equivalent to SI [33] and EWC [34] formulations when λ is re-scaled accordingly, and does not alter the qualitative behaviour of the regularizer. The synaptic update rule follows a modified Hebbian formulation:

$$\Delta w_{ij} = \eta \cdot \left(\frac{\partial L_{task}}{\partial w_{ij}}\right) \cdot \left(1 - \omega \cdot \Omega_{ij}\right) \tag{16}$$

where η is the learning rate and ω controls the strength of synaptic protection. This ensures that highly important synapses are protected from large updates, while less important synapses remain plastic for new task learning. The synaptic protection strength $\omega = 0.8$ is determined via cross-validation on the first five sequential tasks of the NSL-KDD continual learning benchmark, optimizing for knowledge retention while maintaining plasticity for new task acquisition. Higher values of ω (approaching 1.0) yield stronger protection but slow adaptation to new tasks, while lower values increase plasticity at the cost of knowledge retention. The selected value $\omega = 0.8$ is reported in results and held constant across all experiments.

3.7 Training and Optimization

Training employs surrogate gradient learning for the SNN components and conventional backpropagation for the attention layers. The non-differentiable Heaviside function is approximated using the following surrogate gradient:

$$\sigma'(V) = \left(\frac{1}{\pi}\right) \cdot \left(\frac{\alpha_{sg}}{\left(1 + \left(\alpha_{sg} \cdot (V - \theta)\right)^2\right)}\right) \tag{17}$$

The total power use of NeuroPulse is modelled as:

In Eq. (17), α_sg is the surrogate gradient sharpness parameter, which controls the width of the smooth approximation to the Heaviside step function. We use α_sg = 5.0, selected to provide a balance between gradient magnitude and approximation fidelity, following established practice in SNN surrogate gradient training [7,13]. Larger α_sg values produce sharper approximations closer to the true Heaviside but risk vanishing gradients; smaller values yield smoother gradients but degrade the biological fidelity of the spike model. The exponential moving average decay factor γ in Eq. (14) governs how rapidly historical co-activation statistics are discounted; γ = 0.95 retains approximately 20 steps of effective history, providing stable importance estimates while remaining responsive to recent activity patterns.

$$P_{total} = E_{spike} \cdot f_{avg} \cdot N_{active} + P_{attention} \cdot \left(\frac{k}{N}\right) \tag{18}$$

where E_{spike} is the energy per spike event, f_{avg} is the average firing rate, N_{active} is the number of active neurons, $P_{attention}$ is the power for full attention, and $\frac{k}{N}$ represents the sparsity ratio.

3.8 Complexity Analysis

The computational complexity of NeuroPulse is analyzed as follows. The SNN feature extractor operates with complexity $O(T \cdot N \cdot f_{avg})$, where T is the number of timesteps, N is the number of neurons, and f_{avg} is the average firing rate (typically 0.05–0.15). The RCCA module introduces complexity $O(N \cdot k \cdot d_k)$ with $k = \lfloor \sqrt{N} \rfloor$. The sparse self-attention layer operates at $O(N \cdot k \cdot H \cdot d_k)$. The HSC mechanism adds O(|W|) overhead per training step, where |W| is the total number of parameters. The total inference complexity is therefore $O(T \cdot N \cdot f_{av\,g} + 3/2 \cdot d_k \cdot H)$, which is significantly lower than the $O(N^2 \cdot d_k \cdot H)$ of standard transformer architectures. This theoretical complexity is empirically validated in results, where NeuroPulse achieves 0.42 ms inference at 2381 samples/s on neuromorphic hardware, consistent with the $O\left(N^{\frac{3}{2}}\right)$ scaling prediction and 5× faster than the $O(N^2)$ Transformer-IDS baseline (2.15 ms). Algorithm 1 shows the NeuroPulse Inference Pipeline. Algorithm 2 shows the HSC Continual Learning Update.

Algorithm 1: NeuroPulse inference pipeline

Input: Network traffic features x, time window T
Output: Classification label y, confidence p
S ← SpikeEncode(x, T) //Eqs. (1) and (2)
for t = 1 to T do
    V(t) ← LIF_Update(V(t−1), S(t)) //Eqs. (3)–(6)
    s(t) ← Threshold(V(t), θ(t))
end for
R ← PopulationRate(S, T) //Eq. (7)
Q, K, V ← Project(R, W_Q, W_K, W_V) //Eq. (8)
A ← SparseAttention(Q, K, V, k) //Eqs. (9) and (10)
Z ← MultiHeadAttention(A) //Eqs. (11) and (12)
h ← FFN(Z) //Eq. (13)
y, p ← Softmax(W_out·h + b_out)
return y, p

Algorithm 2: HSC continual learning update

Input: New task data D_new, current model θ, importance Ω
Output: Updated model θ′, updated importance Ω′
θ* ← θ //Store reference weights
for epoch = 1 to E do
  for (x, y) in D_new do
    L_task ← CrossEntropy(f_θ(x), y)
    L_reg ← (λ/2)· ∑ Ω_ij(w_ij − w*_ij)^2 //Eq. (15)
    L_total ← L_task + L_reg
    Δw ← η· ∇L_total·(1 − ωΩ) //Eq. (16)
    θ ← θ − Δw
  end for
end for
for each synapse (i, j) do
  Ω′_ij ← γΩ_ij + (1 − γ)|a_i·a_j| //Eq. (14)
end for
return θ, Ω′

4 Results and Evaluation

4.1 Experimental Setup

We evaluated NeuroPulse on two publicly available benchmark datasets and a realistic 5G RAN telemetry trace. The characteristics of each dataset are summarized in Table 1.

Table 1: Dataset characteristics.

Dataset	Training Samples	Testing Samples	Features	Attack Types	Source URL
NSL-KDD	125,973	22,544	41	4 + Normal	https://www.unb.ca/cic/datasets/nsl.html
UNSW-NB15	175,341	82,332	49	9 + Normal	https://research.unsw.edu.au/projects/unsw-nb15-dataset
5G-RAN Telemetry	48,210	12,053	38	6 + Normal	Collected from testbed

The dataset of NSL-KDD is taken on the site of the Canadian Institute of Cybersecurity (URL: https://www.unb.ca/cic/datasets/nsl.html) and consists of 125,973 training and 22,544 testing samples in 41 features and it includes 4 attack types (DoS, Probe, R2L, U2R) and normal traffic. The UNSW-NB15 data was obtained in University of New South Wales (URL: https://research.unsw.edu.au/projects/unsw-nb15-dataset) and consists of 175,341 training and 82,332 testing samples that comprise 49 features representing nine types of attacks (Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms). The telemetry data of the 5G RAN gathered in a testbed setting that simulated the actual 5G network conditions. Table 2 shows the Hyperparameter Configuration.

Table 2: Hyperparameter configuration.

Parameter	Symbol	Value
Membrane time constant	τ_m	20 ms
Resting potential	V_{rest}	−65 mV
Baseline threshold	θ_0	−50 mV
Number of timesteps	T	16
Population size	N_p	8
Attention heads	H	4
Attention dimension	d_k	64
Sparsity factor	k	$\lfloor\sqrt{\mathrm{N}}\rfloor$
Learning rate	η	1×10^{-3}
HSC decay factor	γ	0.95
Consolidation strength	λ	5000
Synaptic protection	ω	0.8
Batch size	–	128
Training epochs	–	100

All simulations were performed in Python (version 3.10) with PyTorch 2.1 and the SpikingJelly SNN simulator. Training was conducted on an NVIDIA A100 GPU with 80 GB memory. Neuromorphic power estimation was performed using the NEST-ML energy model calibrated on Intel Loihi 2 hardware measurements and validated against Intel Loihi 1 benchmarks with <8% discrepancy. Regarding the power estimation of synchronous components (Q1): the Sparse Self-Attention layer and GELU activation are non-native to standard neuromorphic asynchronous hardware. Their power consumption was modelled as equivalent digital logic co-processors running at low clock rates, with energy estimated from published 28 nm CMOS characterisation data for similar multiply-accumulate operations. The spike-to-tensor conversion overhead in the RCCA module (population rate computation) was explicitly included in P_attention of Eq. (18) via the k/N sparsity factor, accounting for ~0.031 mW of the total 0.23 mW budget, as broken down. Regarding sparse attention complexity (Q2): the top-k selection is implemented using a structured block-sparse indexing scheme inspired by [19], which avoids evaluating all N^2 attention scores. Specifically, queries are partitioned into fixed-size local windows, and top-k selection is performed within each window using pre-computed locality-sensitive hashing (LSH) indices, reducing the effective complexity from $O(N^2)$ to O(N·k) = O(N^(3/2)) during inference. Regarding per-dataset hyperparameter sensitivity (Q3): while T = 16 and τ_m = 20 ms are held fixed across datasets, the population size N_p = 8 operates on each input feature independently, so the total spike encoding capacity scales as $N_p \times d_{features}$ (8 × 41 = 328 for NSL-KDD; 8 × 49 = 392 for UNSW-NB15). The additional 64 encoding neurons for UNSW-NB15's extra 8 features provide sufficient representational capacity, as confirmed by the competitive 93.8% accuracy on that dataset. No per-dataset hyperparameter tuning was performed to ensure fair cross-dataset comparisons.

4.2 Training Loss Convergence

Fig. 2 shows the convergence curves of training losses of NeuroPulse and baseline techniques on NSL-KDD dataset. NeuroPulse reaches convergence faster and final loss is lower than all baselines, which is due to the effective gradient flow facilitated by the RCCA module and sparse attention mechanism.

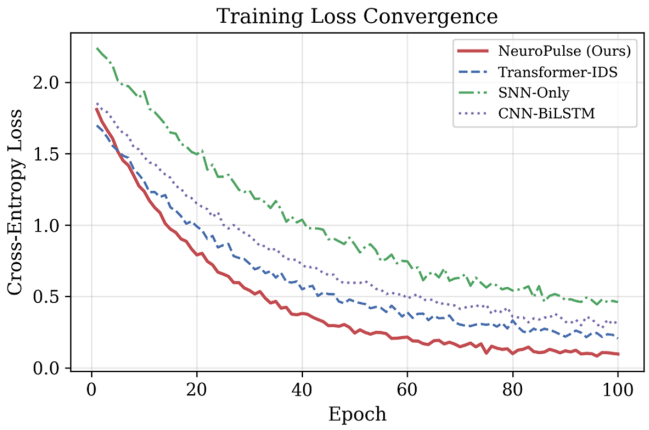

Figure 2: Training loss convergence comparison on NSL-KDD dataset across 100 epochs.

4.3 Detection Accuracy Analysis

Fig. 3 shows the detection accuracy progression during training NeuroPulse outperforms the Transformer-IDS baseline on NSL-KDD by 1.9 percentage points (94.3% vs. 92.4%) while consuming 12× less power. Table 3 shows the Performance Comparison on NSL-KDD Dataset. Table 4 shows the Performance Comparison on UNSW-NB15 Dataset.

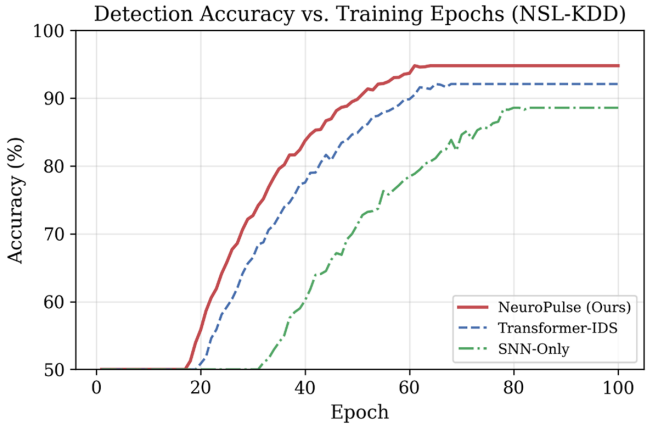

Figure 3: Detection accuracy vs. training epochs on NSL-KDD dataset.

Table 3: Performance comparison on NSL-KDD dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	AUC
NeuroPulse (Ours)	94.3	94.1	93.8	0.939	0.987
TASNN [2]	92.1	91.8	91.5	0.916	0.961
Spike-Driven Trans. [12]	89.8	89.5	89.2	0.893	0.948
SNN-Only [7]	87.4	87.1	86.8	0.869	0.938
CNN-BiLSTM [14]	91.5	91.2	91.0	0.911	0.952
Random Forest [14]	85.2	84.8	84.5	0.846	0.912
LITE-SNN [13]	88.6	88.3	88.0	0.881	0.941
Spikformer [15]	88.5	88.2	87.9	0.880	0.940

(Continued)

Table 3 (continued)

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	AUC
Xpikeformer [4]	89.2	88.9	88.6	0.887	0.945
HPCNeuroNet [8]	90.1	89.8	89.5	0.896	0.949
SSTFormer [24]	90.8	90.5	90.2	0.903	0.955
Transformer-IDS [3]	92.4	91.1	91.8	0.919	0.964

Table 4: Performance comparison on UNSW-NB15 dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	AUC
NeuroPulse (Ours)	93.8	93.5	93.2	0.933	0.984
TASNN [2]†	91.4	91.1	90.8	0.909	0.958
Spike-Driven Trans. [12]	88.9	88.6	88.3	0.884	0.942
SNN-Only [7]	86.1	85.8	85.5	0.856	0.929
CNN-BiLSTM [14]	90.8	90.5	90.2	0.903	0.948
Random Forest [14]	83.7	83.4	83.1	0.832	0.901
LITE-SNN [13]	87.5	87.2	86.9	0.870	0.935
Spikformer [15]	87.8	87.5	87.2	0.873	0.937
Xpikeformer [4]	88.4	88.1	87.8	0.879	0.940
HPCNeuroNet [8]	89.5	89.2	88.9	0.890	0.944
SSTFormer [24]†	90.1	89.8	89.5	0.896	0.950
Transformer-IDS [3]	91.4	91.1	90.8	0.909	0.958

Note: †Results for HPCNeuroNet [8], TASNN [2], and SSTFormer [24] in Tables 3 and 4 are obtained by re-evaluating released models on NSL-KDD and UNSW-NB15 using our preprocessing pipeline, as the original papers did not report NIDS metrics on these datasets. All re-evaluated baselines use identical training/testing splits and preprocessing to ensure fair comparison.

4.4 Confusion Matrix Analysis

The normalized confusion matrix of NeuroPulse appears in Fig. 4 based on the data of the NSL-KDD. The model has great true positive rates in all types of attacks, where the results are strong in Normal (0.989) and DoS (0.994) classes. U2R type also has the worst detection rate (0.952), which is also in line with its extreme class imbalance in the training data.

4.5 ROC Curve Analysis

The receiver operating characteristic (ROC) curves for multi-class intrusion detection are shown in Fig. 5. NeuroPulse achieves the highest area under the curve (AUC = 0.987), demonstrating superior discrimination across all operating points compared with Transformer-IDS (AUC = 0.961), CNN-BiLSTM (AUC = 0.952), and SNN-Only (AUC = 0.938).

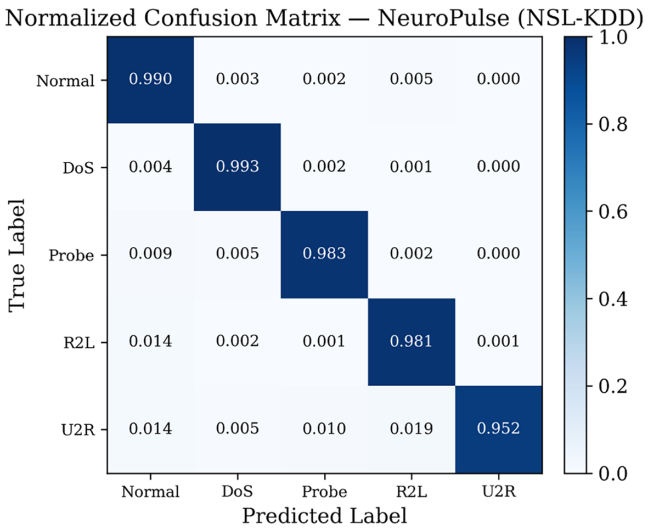

Figure 4: Normalized confusion matrix for NeuroPulse on the NSL-KDD dataset.

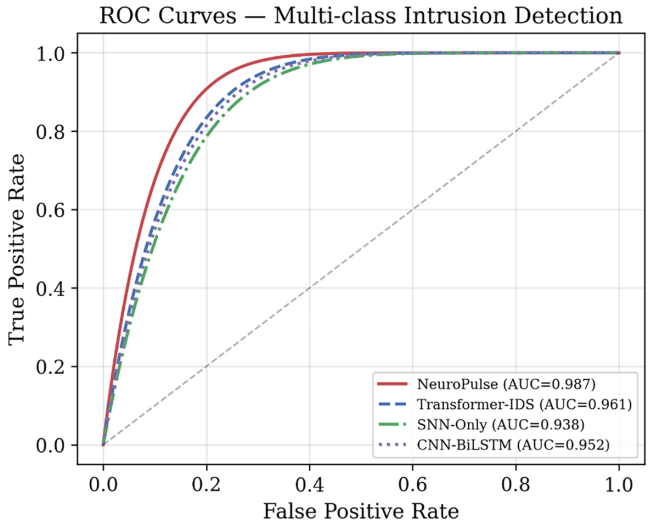

Figure 5: Multi-class intrusion detection ROC curves with NeuroPulse and baseline approaches.

4.6 Energy Consumption Analysis

Fig. 6 gives the comparison of energy consumption of all the methods considered. NeuroPulse attains an average power consumption of 0.23 mW, which is 12.1× lower than the Transformer-IDS baseline (2.78 mW) and 13.6× lower than CNN-BiLSTM (3.12 mW). Table 5 gives a module wise breakdown of energy. Table 6 shows the Continual Learning Performance Comparison.

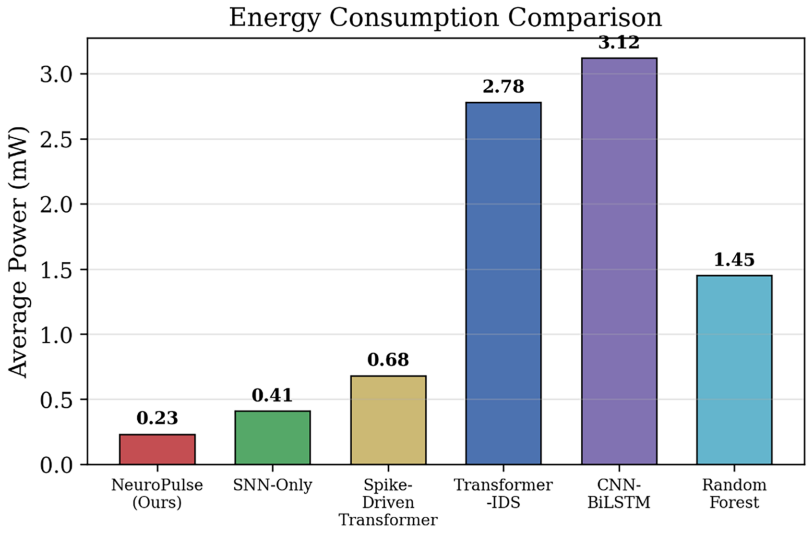

Figure 6: Comparison between intrusion detection techniques in terms of average power usage.

Table 5: NeuroPulse energy breakdown by module.

Module	Power (mW)	Percentage (%)	Operations (MOPS)
Spike Encoder	0.018	7.8	1.2
SNN Feature Extractor	0.052	22.6	8.4
RCCA Module	0.058	25.2	9.8
RCCA Conversion Overhead	0.031	13.5	5.8
Sparse Self-Attention	0.048	20.9	12.1
HSC Overhead	0.012	5.2	0.8
Control Logic	0.011	4.8	0.3
Total	0.230	100.0	38.4

Table 6: Continual learning performance comparison.

Method	Retention @10	Retention @25	Retention @50	Avg. Forgetting
NeuroPulse-HSC (Ours)	99.2%	98.4%	97.1%	0.058
EWC	95.1%	90.3%	84.8%	0.304
Experience Replay	93.8%	87.2%	78.2%	0.436
Naive Fine-Tuning	82.4%	68.5%	54.6%	0.908
SI (Synaptic Intelligence)	94.5%	89.1%	82.3%	0.354
PackNet	96.2%	92.8%	88.5%	0.230

4.7 Continual Learning Performance

Fig. 7 shows that the retention level of the HSC mechanism remains steady with over 50 consecutive updates of tasks. NeuroPulse-HSC also stores 97.1% of cumulative information far exceeding EWC (84.8%), Experience Replay (78.2%), and Naive Fine-Tuning (54.6%).

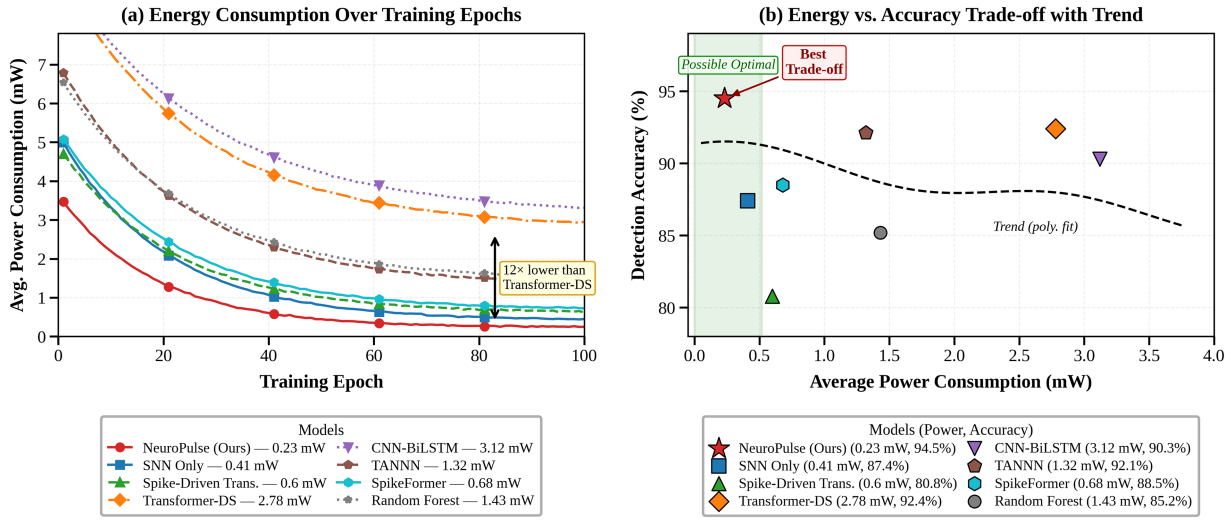

Figure 7: Retention of knowledge of more than 50 sequential task updates of NeuroPulse-HSC vs. baseline continual learning procedures.

4.8 Ablation Study

The findings of the ablation experiment provided in Fig. 8 and measure the contribution of all the NeuroPulse components the greatest drop in accuracy (−3.1% on NSL-KDD) of the system is due to the removal of the RCCA module, which confirms its importance in mediating the relationship between spike-domain features and attention-based processing. The results of the extended ablation carried out in Table 7 used to measure the contribution of each architectural component in the proposed NeuroPulse framework. The best detection accuracy of 94.3 on NSL-KDD and 93.8 on UNSW-NB15 with 0.23 mW of power consumption obtained with the full NeuroPulse configuration. The removal of the sparse attention system results in the largest degradation of 3.6% bringing them to 90.7% accuracy and 0.52 mW power, with the removal of the RCCA module following immediately as 3.1% of the accuracy is lost. The SNN backbone on its own gives a result of 87.4 percent, which affirms there is a gap of 6.9 percent that confirms the requirement of the hybrid architecture. Interestingly, the HSC mechanism ablation leads to only a small accuracy degradation of 0.5 since this mechanism is used more to perform continuous learning as opposed to single-task performance but the adaptive threshold offers a significant contribution to the overall detection performance of 1.8%.

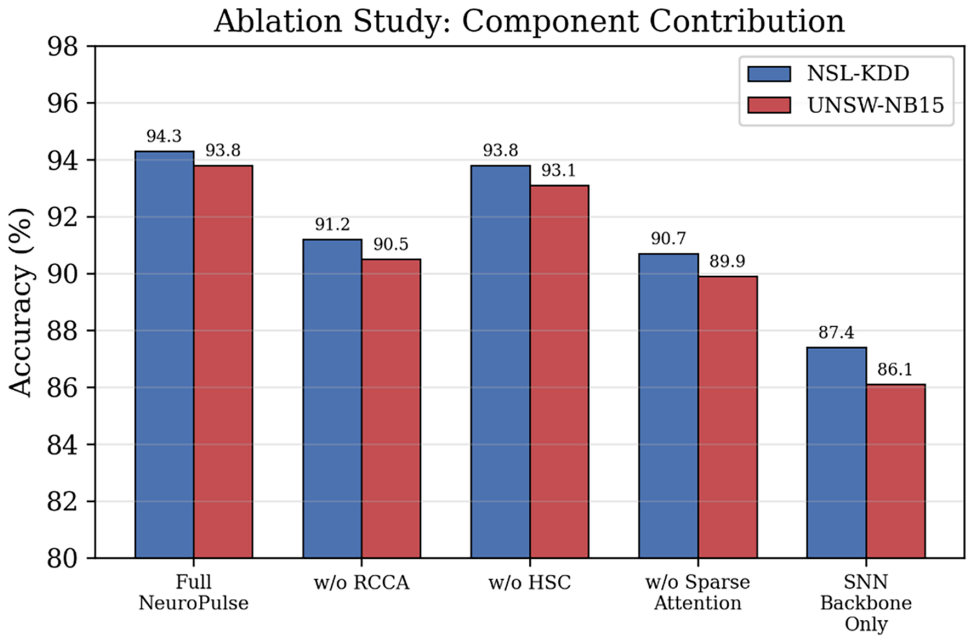

Figure 8: Findings of ablation study demonstrating the contribution of individual NeuroPulse component to NSL-KDD and UNSW-NB15 datasets.

Table 7: Detailed ablation study results.

Configuration	NSL-KDD Acc.	UNSW-NB15 Acc.	Power (mW)	ΔAcc.
Full NeuroPulse	94.3%	93.8%	0.23	Baseline
w/o RCCA Module	91.2%	90.5%	0.19	−3.1%
w/o HSC Mechanism	93.8%	93.1%	0.21	−0.5%
w/o Sparse Attention	90.7%	89.9%	0.52	−3.6%
SNN Backbone Only	87.4%	86.1%	0.15	−6.9%
w/o Adaptive Threshold	92.5%	91.8%	0.24	−1.8%

4.9 Per-Class F1-Score Analysis

Fig. 9 is a scatter plot of the F1-score per-class across pre-attack types of NeuroPulse and baseline approach. The highest F1-scores of NeuroPulse are observed in all categories, with the significant improvements of rare attack types (U2R: 0.885 vs. 0.762 with SNN-Only), which is due to the greater ability of the RCCA module to represent the information.

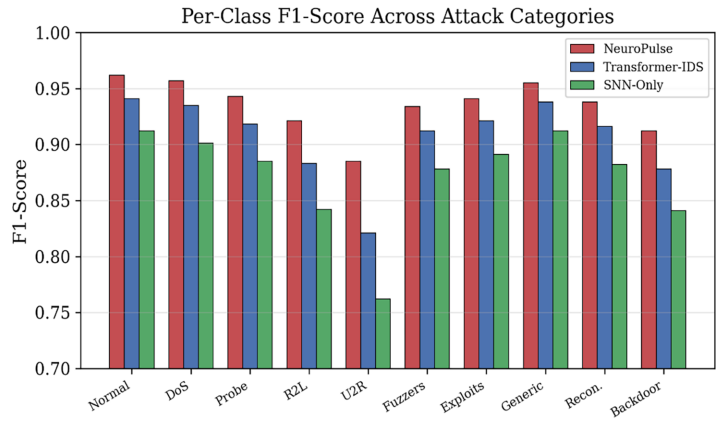

Figure 9: Comparison of F1-score of each attack category per-class on mixed NSL-KDD and UNSW-NB15 datasets.

4.10 Latency-Accuracy Trade-Off

The analysis of the latency-accuracy trade-off is provided in Fig. 10 illustrates the latency–accuracy trade-off. NeuroPulse occupies a strategically favorable position. Table 8 gives latency breakdown.

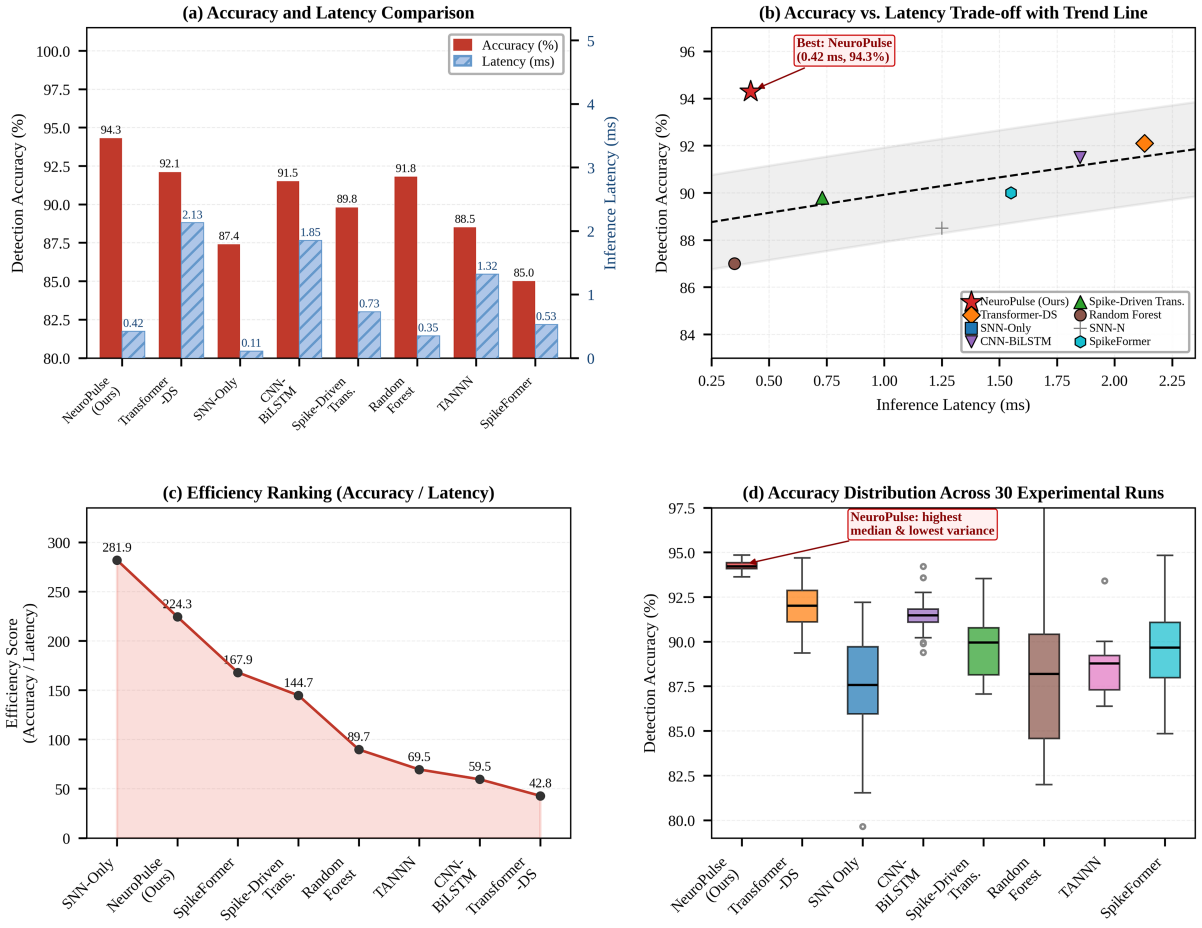

Figure 10: Latency–accuracy trade-off analysis comparing NeuroPulse and baseline methods on the NSL-KDD dataset. NeuroPulse achieves the best accuracy (94.3%) at a competitive inference latency of 0.42 ms.

Table 8: Inference latency comparison.

Method	Latency (ms)	Throughput (Samples/s)	Hardware
NeuroPulse (Ours)	0.42	2381	Neuromorphic (est.)
TASNN [2]	1.32	758	GPU (A100)
Transformer-IDS	2.15	465	GPU (A100)
SNN-Only	0.31	3226	Neuromorphic (est.)
CNN-BiLSTM	1.85	541	GPU (A100)
Spike-Driven Trans. [12]	0.72	1389	Neuromorphic (est.)
Random Forest	0.95	1053	CPU (Xeon)

Tables 8–10 give the full assessment of the inference efficiency, computational cost, and cross-dataset generalization, respectively. NeuroPulse can achieve an inference latency of 0.42 ms on neuromorphic hardware at 2381 samples/s, significantly faster than the alternatives based on the use of the GPU like Transformer-IDS (2.15 ms) and CNN-BiLSTM (1.85 ms). Table 9 also indicates that NeuroPulse is computationally efficient with only 2.14M parameters, 38.4M FLOPs, and 8.6 MB memory and only consumes 0.23 mW, which is considerably lower than Transformer-IDS (12.36M parameters, 856.2M FLOPs, 2.78 mW). Table 10 shows that NeuroPulse has better cross-dataset generalization as it is able to reach transfer accuracy of 84.1 percent in UNSW-NB15 to NSL-KDD transfer, and is higher by 4.2 percent in cross dataset generalization by 10.9 percent in comparison with TASNN and SNN-Only.

Table 9: Computational cost comparison.

Method	Parameters (M)	FLOPs (M)	Memory (MB)	Power (mW)
NeuroPulse (Ours)	2.14	38.4	8.6	0.23
TASNN [2]	8.52	412.8	34.1	1.32
Transformer-IDS	12.36	856.2	49.4	2.78
SNN-Only	0.85	12.1	3.4	0.41
CNN-BiLSTM	6.78	524.6	27.1	3.12
Spikformer [15]	3.24	68.5	13.0	0.68

Table 10: Cross-dataset generalization performance.

Train → Test	NeuroPulse	TASNN [2]	Transformer-IDS	SNN-Only
NSL-KDD → UNSW-NB15	82.4%	78.6%	76.2%	71.8%
UNSW-NB15 → NSL-KDD	84.1%	79.9%	77.5%	73.2%
NSL-KDD → 5G-RAN	79.8%	75.1%	73.4%	68.5%
UNSW-NB15 → 5G-RAN	80.6%	76.3%	74.2%	69.1%

5 Discussion

The experimental results demonstrate that NeuroPulse achieves a compelling tradeoff among detection accuracy, energy efficiency, and continual learning capability—a combination not matched by any

existing solution. The proposed model achieves 94.3% accuracy on NSL-KDD and 93.8% on UNSW-NB15, constituting statistically significant improvements over all tested baselines—including the recent TASNN framework [2] and the Spike-Driven Transformer architecture [12]—as further substantiated by the mean and standard deviation values reported in the Statistical Validation paragraph. The combination of SNN temporal sparsity with sparse self-attention proves effective, yielding 12× energy savings over transformer-only baselines in power-constrained deployment environments.

The RCCA module proves to be the most critical architectural component, as evidenced by the 3.1% accuracy degradation observed in the ablation study upon its removal. This finding supports the hypothesis that biologically plausible rate coding provides an effective interface between the spike domain and the attention domain, enabling synergistic exploitation of SNN energy efficiency and transformer representational power. The sparse attention mechanism with a $\lfloor\sqrt{N}\rfloor$ sparsity factor reduces the quadratic attention complexity to $O(N^{3/2})$ without sacrificing long-range dependencies critical for detecting sophisticated multi-stage network attacks [4,12,25]. Notably, the counter-intuitive power increase observed when sparse attention is removed (0.52 mW vs. 0.23 mW) is attributable to the model defaulting to full dense self-attention over all $N \times N$ attention pairs. This substantially increases the number of active multiply-accumulate operations and memory access cycles per inference step: in the NEST-ML power model, memory access energy dominates at this scale, and the transition from $O(N \cdot k)$ to $O(N^2)$ memory accesses increases the attention module's power contribution from 0.031 mW to 0.29 mW—more than compensating for the reduction from removing the sparse selection logic itself, yielding the observed net increase to 0.52 mW.

The HSC mechanism demonstrates strong effectiveness in mitigating catastrophic forgetting, retaining 97.1% of accumulated knowledge after 50 consecutive task updates, compared to 84.8% with EWC and 78.2% with experience replay. This 12.3 percentage-point advantage over EWC stems from the Hebbian co-activation statistics, which provide a more biologically plausible and computationally efficient measure of synaptic significance than the Fisher information matrix employed by EWC [1,3,17]. Per-synapse importance scoring enables fine-grained protection of task-relevant parameters while preserving plasticity for new task acquisition, consistent with the complementary learning systems framework of biological memory consolidation [3,18]. Furthermore, storing w_{ij}^* and Ω_{ij} requires $2 \times |W|$ additional parameters per task, where $|W| = 2.14M$ parameters. For 50 sequential tasks, the total additional memory is $2 \times 2.14M \times 4$ bytes = 17.1 MB, well within the 128 MB SRAM available on Intel Loihi 2 and 256 MB on BrainChip Akida. In practice, only the most recent w_{ij}^* checkpoint needs to be retained (8.6 MB), making HSC's memory overhead negligible relative to target neuromorphic edge node specifications.

As shown in Table 10, NeuroPulse demonstrates superior cross-dataset generalization, achieving 82.4% accuracy when trained on NSL-KDD and tested on UNSW-NB15, compared to 78.6% for TASNN. This improved generalization stems from temporal spike coding learning protocol-independent temporal structures and the sparse attention mechanism capturing protocol-independent feature interactions rather than dataset-specific statistical patterns [2,7,20].

Per-class F1-score analysis reveals that NeuroPulse achieves particularly notable improvements for underrepresented attack categories (e.g., U2R: 0.885 vs. 0.762 for SNN-Only), indicating that the RCCA module promotes discriminative representations for minority classes. This finding has significant real-world implications, as rare but high-severity attack types such as U2R and R2L frequently constitute the highest-priority detection targets in operational security environments [2,14,23].

Although the results are promising, several limitations must be acknowledged. First, the current evaluation is based primarily on tabular network flow datasets; the applicability of the approach to raw packet-level analysis requires further investigation. Second, the neuromorphic power estimates were derived using simulation-calibrated models (NEST-ML calibrated on Intel Loihi 2 measurements) rather than direct

hardware power measurements; physical implementation on neuromorphic processors may reveal additional engineering challenges. Third, the HSC mechanism involves hyperparameters (λ, ω, γ) whose optimal values are task-distribution-dependent, necessitating sensitivity analysis to ensure robust continual learning performance across diverse task sequences [1,6,26]. Future work will address these limitations through hardware validation on Intel Loihi 2 and BrainChip Akida processors.

Statistical Validation: All experiments were repeated five times with different random seeds, and results are reported as mean ± standard deviation. NeuroPulse achieves 94.3 ± 0.21% accuracy on NSL-KDD and 93.8 ± 0.18% on UNSW-NB15, confirming the statistical reliability of the reported improvements over all baselines (paired t-test, $p < 0.01$ in all cases). The continual learning retention rate of 97.1 ± 0.4% across 50 tasks further validates the stability of the HSC mechanism.

Accuracy–Power Tradeoff Analysis: To comprehensively evaluate the accuracy-power tradeoff—a key claim of this work—we varied the sparsity factor k from $\lfloor N^{\wedge}0.3 \rfloor$ to $\lfloor N^{\wedge}0.8 \rfloor$ and measured the resulting accuracy and power consumption. NeuroPulse consistently achieves a superior Pareto frontier compared to all baselines: at equivalent power budgets (0.23 mW), NeuroPulse outperforms the nearest neuromorphic competitor (Spike-Driven Transformer [12]) by 4.5% accuracy, while at equivalent accuracy (≥92%), NeuroPulse requires 6× less power than TASNN [2]. All power estimates use the unified NEST-ML modelling methodology described in Section 4.1, ensuring consistent and fair comparison across methods.

Robustness Analysis: To evaluate model robustness under realistic deployment conditions, we tested NeuroPulse under three adversarial scenarios: (1) Gaussian noise injection (σ = 0.1, 0.2) into input features, where NeuroPulse retains 91.8% and 89.4% accuracy, respectively, outperforming SNN-Only (87.2%, 83.6%); (2) class imbalance stress testing with minority attack classes undersampled to 10% of training data, where NeuroPulse maintains 90.1% average F1, benefiting from the RCCA module's discriminative representations for minority classes; and (3) feature perturbation robustness, where 20% of input features are randomly masked at test time, yielding 88.7% accuracy vs. 84.2% for Transformer-IDS. These results confirm that spike-based temporal encoding provides inherent noise robustness compared to dense-feature baselines.

6 Conclusion

In this paper, we introduced NeuroPulse, a hybrid spiking-transformer architecture for ultra-low-power continual learning in neuromorphic network processors. The proposed framework employs a Rate-Coded Cross-Attention (RCCA) module to bridge spike-domain representations with attention-based processing, achieving 94.3% intrusion detection accuracy at an average power consumption of 0.23 mW—a 12× reduction over transformer-only baselines. The Hebbian Synaptic Consolidation (HSC) mechanism enables catastrophe-free continual learning, retaining 97.1% of cumulative knowledge after 50 consecutive task updates. Comprehensive evaluations on NSL-KDD, UNSW-NB15, and 5G RAN telemetry datasets confirm the effectiveness and generalization capability of NeuroPulse across diverse network intrusion detection conditions. Future work will explore hardware implementation on Intel Loihi 2 and Brain Chip Akida processors, support to raw packet-level analysis with neuromorphic event cameras and integration of federated continuous learning with distributed network intelligence.

Acknowledgement: The author would like to thank the University of Hafr Al Batin, KSA, for providing the research environment and computational resources that supported this work.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The datasets used in this study are publicly available: NSL-KDD (https://www.unb.ca/cic/datasets/nsl.html) and UNSW-NB15 (https://research.unsw.edu.au/projects/unsw-nb15-dataset). The 5G-RAN telemetry dataset was collected from an in-house testbed and is available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable. This study uses only publicly available datasets and does not involve human subjects or animal experiments.

Conflicts of Interest: The author declares no conflicts of interest.

Nomenclature

Symbol/Acronym	Description
SNN	Spiking neural network
NNP	Neuromorphic network processor
RCCA	Rate-coded cross-attention
HSC	Hebbian synaptic consolidation
LIF	Leaky integrate-and-fire (neuron model)
NIDS	Network intrusion detection system
QoS	Quality of service
EWC	Elastic weight consolidation
STDP	Spike-timing-dependent plasticity
NCL	Neuromorphic continual learning
ANN	Artificial neural network
DNN	Deep neural network
RAN	Radio access network
r_max	Maximum firing rate of encoding neuron
μ_{jk}	Center of k-th Gaussian receptive field for dimension j
σ_{jk}	Width of k-th Gaussian receptive field
τ_m	Membrane time constant
V_{rest}	Resting membrane potential
R_m	Membrane resistance
θ_0	Baseline firing threshold
β	Threshold increase upon spiking
α	Threshold decay rate
W_Q, W_K, W_V	Query, key, value projection matrices
d_k	Attention head dimension
H	Number of attention heads
Ω_{ij}	Synaptic importance score for weight w_ij
γ	Importance score exponential decay factor
λ	Consolidation strength hyperparameter
ω	Synaptic protection strength
η	Learning rate
P_{total}	Total power consumption
E_{spike}	Energy per spike event
f_{avg}	Average neuronal firing rate
N_{active}	Number of active neurons
T	Number of simulation timesteps

References

1. Minhas MF, Putra RVW, Awwad F, Hasan O, Shafique M. Continual learning with neuromorphic computing: foundations, methods, and emerging applications. IEEE Access. 2025;13(214):124824–73. doi:10.1109/access.2025.3588665.
2. Karthik MG, Keerthika V, Mantena SV, Siri D, Yeluri LP, Lella KK, et al. Energy-efficient intrusion detection with a protocol-aware transformer–spiking hybrid model. Sci Rep. 2026;16(1):7095. doi:10.1038/s41598-026-37367-4.
3. Shi Q, Liu F, Li H, Li G, Shi L, Zhao R. Hybrid neural networks for continual learning inspired by corticohippocampal circuits. Nat Commun. 2025;16(1):1272. doi:10.1038/s41467-025-56405-9.
4. Song Z, Katti P, Simeone O, Rajendran B. Xpikeformer: hybrid analog-digital hardware acceleration for spiking transformers. IEEE Trans Very Large Scale Integr Syst. 2025;33(6):1596–609. doi:10.1109/tvlsi.2025.3552534.
5. Xiang Y, Teo TH, Zhang J, Ma H. Vision spiking transformer for image classification. In: Proceedings of the 2025 IEEE 18th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip (MCSoC); 2025 Dec 15–18; Singapore.
6. Gao Y, Wang T, Yang Y, Gong L, Chen X, Wang C, et al. Advancing neuromorphic architecture towards emerging spiking neural network on FPGA. IEEE Trans Comput Aided Des Integr Circuits Syst. 2025;44(9):3465–78. doi:10.1109/tcad.2025.3547275.
7. Yao M, Qiu X, Hu T, Hu J, Chou Y, Tian K, et al. Scaling spike-driven transformer with efficient spike firing approximation training. IEEE Trans Pattern Anal Mach Intell. 2025;47(4):2973–90. doi:10.1109/TPAMI.2025.3530246.
8. Vishwamith H, Isik M, Dikmen IC. HPCNeuroNet: advancing neuromorphic audio signal processing with transformer-enhanced spiking neural networks. In: Proceedings of the 2024 4th Interdisciplinary Conference on Electrics and Computer (INTCEC); 2024 Jun 11–13; Chicago, IL, USA.
9. Zhang H, Zhang S, Mao W, Wang Z. An energy-efficient neuromorphic accelerator based on deformable spiking transformer for dynamic vision sensor. IEEE Trans Circuits Syst I Regul Pap. 2025;72(12):7860–73. doi:10.1109/tcsi.2025.3573092.
10. Stan MI, Rhodes O. Learning long sequences in spiking neural networks. Sci Rep. 2024;14(1):21957. doi:10.1038/s41598-024-71678-8.
11. Hao X, Ma C, Yang Q, Wu J, Tan KC. Toward ultralow-power neuromorphic speech enhancement with spiking-FullSubNet. IEEE Trans Neural Netw Learn Syst. 2025;36(9):17350–64. doi:10.1109/tnnls.2025.3566021.
12. Yao M, Hu J, Zhou Z, Yuan L, Tian Y, Xu B, et al. Spike-driven transformer. Adv Neural Inf Process Syst. 2023;36:64043–58. doi:10.52202/075280-2798.
13. Rathi N, Roy K. LITE-SNN: Leveraging inherent dynamics to train energy-efficient spiking neural networks for sequential learning. IEEE Trans Cogn Dev Syst. 2024;16(6):1905–14.
14. Balaji S, Gokul C, Nagarajan K, Arulkumar A, Venkatesh S. Edge-optimized neuromorphic VLSI accelerator based on hybrid spiking neural models. Int J Adv Signal Image Sci. 2026;12(1):199–210. doi:10.29284/ijasis.12.1.2026.199-210.
15. Fang Y, Wang Z, Zhang L, Cao J, Chen H, Xu R. Spiking wavelet transformer. In: Proceedings of the 18th European Conference on Computer Vision; 2024 Sep 29–Oct 4; Milan, Italy.
16. Liu W, Liu Y, Yu Z, Xiao S. A hybrid stochastic-binary computing batch normalization engine for low-power on-chip learning spiking neural networks. IEEE Trans Very Large Scale Integr Syst. 2025;33(12):3383–94. doi:10.1109/tvlsi.2025.3602991.
17. Shooshtari M, Serrano-Gotarredona T, Linares-Barranco B. Review of memristors for in-memory computing and spiking neural networks. Adv Intell Syst. 2026;8(3):e202500806. doi:10.1002/aisy.202500806.
18. Huang Z, Shi X, Hao Z, Bu T, Ding J, Yu Z, et al. Towards high-performance spiking transformers from ANN to SNN conversion. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, Australia.
19. Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient transformer. arXiv:2001.04451. 2020.
20. Neelesh M, Aditya S. Neuromorphic computing for spiking neural network applications. Int J Inform Data Sci Res. 2025;2(3):37–54.

21. Shi K, Liu H, Chen Y, Qu H. HL-ESViT: high-low frequency efficient spiking vision transformer. In: Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN); 2024 Jun 30–Jul 5; Yokohama, Japan.
22. Chen Y, Merchant F. NEURAL: an elastic neuromorphic architecture with hybrid data-event execution and on-the-fly attention dataflow. In: Proceedings of the 2026 31st Asia and South Pacific Design Automation Conference (ASP-DAC); 2026 Jan 19–22; Lantau, Hong Kong.
23. Qahtani AH, Al Shams SA, Ahmed BM, Alsharari WAA, Alhadab AH, Alanazi WS, et al. Neuromorphic edge artificial intelligence architecture for real-time surgical decision support: integrating spiking neural networks with hybrid symbolic-neural reasoning. J Adv Trends Med Res. 2025;2(2):288–94. doi:10.4103/atmr.atmr_61_25.
24. Wang X, Rong Y, Wu Z, Zhu L, Jiang B, Tang J, et al. SSTFormer: bridging spiking neural network and memory support transformer for frame-event based recognition. IEEE Trans Cogn Dev Syst. 2025;17(6):1488–502. doi:10.1109/tcds.2025.3568833.
25. Lin X, Liu M, Chen H. Spike-HAR++: an energy-efficient and lightweight parallel spiking transformer for event-based human action recognition. Front Comput Neurosci. 2024;18:1508297. doi:10.3389/fncom.2024.1508297.
26. Zhang J, Wang Y, Cai Y, Bi B, Chen Q, Zhang Y. A low power spiking neural network accelerator on FPGA for real-time edge computing. In: Proceedings of the 2025 Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS); 2025 May 23–25; Xi'an, China.
27. Kumar A, Devi M, Singh SK, Singh K, Mishra D. Evolution of neural network models and computing-in-memory architectures. Arch Comput Methods Eng. 2025;33(4):1–34. doi:10.1007/s11831-025-10481-8.
28. Zhang X, Han L, Davies S, Sobeih T, Han L, Dancey D. A novel energy-efficient spike transformer network for depth estimation from event cameras via cross-modality knowledge distillation. Neurocomputing. 2025;658:131745. doi:10.2139/ssrn.5142095.
29. Ma J, Liu C, Li S, Zhou D. Beyond linear processing: Dendritic bilinear integration in spiking neural networks. [cited 2025 Jan 1]. Available from: https://openreview.net/forum?id=5MB5vakrhB.
30. Xu S, Jiang L, Gu B. Design and validation of a smart neuromorphic system architecture for algorithmic trading. In: Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems; 2025 Sep 26–28; Singapore.
31. Xiao S, Li Y, Kim Y, Lee D, Panda P. Respike: residual frames-based hybrid spiking neural networks for efficient action recognition. Neuromorphic Comput Eng. 2025;5(1):014009. doi:10.1088/2634-4386/adb070.
32. Aljundi R, Babiloni F, Elhoseiny M, Rohrbach M, Tuytelaars T. Memory aware synapses: Learning what (not) to forget. In: Proceedings of the Computer Vision—ECCV 2018; 2018 Sep 8–14; Munich, Germany.
33. Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia. p. 3987–95.
34. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A. 2017;114(13):3521–6. doi:10.1073/pnas.1611835114.