This paper has been accepted for publication in IEEE Military Communications Conference (MILCOM) 2026. This is the authors' accepted version of the article. The final version will be available in IEEE Xplore.

Neuromorphic Radio Fingerprinting

Roberta Bucchignani, Francesca Meneghello, Francesco Restuccia
Institute for Intelligent Networked Systems, Northeastern University, United States

***Abstract*—Real-time device authentication in multi-user multiple-input multiple-output (MU-MIMO) networks demands continuous, cryptography-free approaches deployable at the network edge. Radio fingerprinting (RFP) addresses this by leveraging tiny device-specific hardware imperfections that affect radio signal generation and processing. Traditional artificial neural networks (ANNs) are extensively used to extract such features from the radio signal collected at a monitoring device or from related metrics such as the channel state information (CSI). However, their dense, multiply-accumulate-heavy processing makes real-time edge deployment prohibitive. In this paper, we tackle this challenge by moving from dense to sparse processing. To our knowledge, we introduce NeuRFP, the first neuromorphic framework for device fingerprinting from compressed MU-MIMO CSI feedback matrices. NeuRFP encodes these matrices into binary spikes and processes them through spiking neural networks (SNNs), replacing costly multiply-accumulate operations with sparse additions. We evaluate the performance–complexity tradeoff of NeuRFP and compare it with a state-of-the-art ANN-based baseline using a publicly available 10-device RFP dataset. Our SNN-based architecture matches the baseline accuracy while reducing estimated energy consumption by more than 98%.**

***Index Terms*—Radio fingerprinting, spiking neural networks (SNNs), channel state information (CSI), multi-user multiple-input multiple-output (MU-MIMO).**

I. Introduction

The rapid expansion of the Internet of Things (IoT) is leading to a significant "spectrum crunch" [1]. As such, modern wireless standards, including Wi-Fi 7 (IEEE 802.11be) and 6G, rely on multi-user (MU)-multiple-input, multiple-output (MIMO) to serve multiple users simultaneously, significantly increasing spectral efficiency [2]. Given their high device density, MU-MIMO networks are more susceptible to adversarial attacks than single-link networks [3]. For this reason, robust device authentication must be used to prevent unauthorized access and protect legitimate users [4].

To perform device authentication without impacting communication performance, physical (PHY) radio fingerprinting (RFP) has emerged as an alternative approach to energy-expensive cryptography [5]. RFP relies on tiny, hardware-level imperfections introduced during the manufacturing process (e.g., I/Q imbalance and power amplifier non-linearities) [6]. These imperfections leave a unique "device fingerprint" that is extremely difficult to mimic or alter and, in turn, can serve as a proxy to identify the device [6]. Recent research has shown that these hardware imperfections percolate in the channel state information (CSI) routinely estimated by any wireless device for communication and sensing purposes. For example, in [7], the authors obtain device fingerprints by estimating the device-specific carrier frequency offset (CFO) from the CSI. More recently, other investigations have shown that channel-independent fingerprints can be extracted directly from CSI feedback matrices [8], [9]. This allows performing authentication using off-the-shelf hardware. *The key challenge, however, is that processing high-dimensional CSI samples in real-time requires substantial computational resources*. Since traditional artificial neural network (ANN)-based approaches for RFP rely on millions of energy-intensive multiply-accumulate (MAC) operations [10], [11], edge devices with strict energy budgets might not be able to achieve real-time fingerprinting.

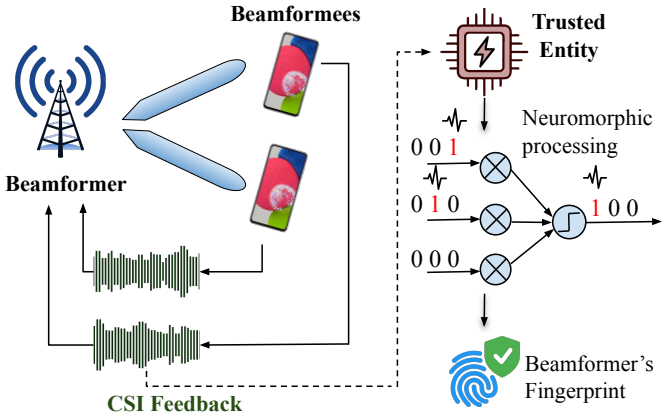

Fig. 1: Workflow of MU-MIMO neuromorphic fingerprinting. The CSI estimates fed back by the beamformees allow fingerprinting the beamformer.

In this paper, we propose using spiking neural networks (SNNs) to address the current limitations. As shown in Figure 1, the CSI feedback transmitted by the beamformees to the beamformer is captured by a Trusted Entity (TE) node and processed through a SNN, running on a commercial-off-the-shelf (COTS) device, to obtain a fingerprint of the beamformer for authentication. While some research efforts have successfully used SNNs for time-domain raw I/Q radio data [12], their application to frequency-domain CSI feedback structures remains to our knowledge unexplored. By encoding CSI feedback into event-driven binary spikes, SNNs replace complex multiplications with simple additions, thus achieving high-accuracy RFP with significantly lower power consumption than ANNs [11].

Summary of Novel Contributions

• We design and implement NeuRFP, the first neuromorphic RFP framework for fingerprinting MU-MIMO beamformers from compressed CSI feedback. As part of the design, we investigate both convolutional (NeuRFP-C) and dense

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

(NeuRFP-D) architectures, and identify the most effective spike encoding mechanism to convert CSI data into spikes;

• We evaluate NeuRFP in comparison with a state-of-the-art ANN strategy [8] using a publicly available dataset collecting CSI associated with 10 different beamformers. Our results show that a dense, convolution-free spiking network matches or closely approaches the ANN baseline's accuracy across different operational scenarios, proving that deep spatial feature extraction is not strictly required for robust RFP;

• We perform a comprehensive analytical energy profiling based on analytical CMOS frameworks to evaluate the neuromorphic computational advantage over the ANN baseline. We reveal that spatial convolutions trigger a severe "memory trap" in the spiking domain and drastically increase energy costs. The purely dense topology bypasses this hardware bottleneck and reduces the estimated energy consumption by nearly two orders of magnitude compared to the ANN baseline. We provide public access to the code of the implementation at https://github.com/Restuccia-Group/NeuRFP-MILCOM.

II. Related Work

Riyaz et al. [13] demonstrated the efficacy of using convolutional neural networks (CNNs) for RFP. Subsequent research by Sankhe et al. [10] introduced a framework named ORACLE to make RFP scalable and robust to environmental changes. Intentional modifications applied to the transmitted signal are used to identify COTS Wi-Fi devices even when exposed to dynamic multipath fading by processing the received signals through CNNs. Recently, more advanced architectures have been proposed in the literature, such as the csi-RFF framework proposed in [14], which leverages hardware-induced micro-signals embedded in the CSI. These approaches still rely on traditional, continuous-valued ANN processing which limits their practical implementation. Indeed, the intensive floating-point matrix multiplications make real-time authentication prohibitive.

To address this bottleneck, recent research has pivoted towards SNN to enable energy-efficient RFP. By leveraging sparse, asynchronous binary events, SNNs drastically reduce power consumption with respect to traditional ANN-based approaches. For example, in [12], the authors target RFP of VHF Data Exchange System (VDES) satellite components from time-domain I/Q samples. They employ a nine-layer CNN as fingerprint extractor together with an LSTM-based channel predictor for data augmentation, both mapped to an adaptive clock/event-driven SNN through an ANN-to-SNN conversion with rate and burst coding. Similarly, the authors in [15] address Automatic Dependent Surveillance–Broadcast (ADS-B) specific emitter identification from time-domain I/Q sequences by introducing a deep spiking complex neural network, where complex-valued convolutional layers precede LIF neurons. Both contributions differ from our proposed NeuRFP framework in terms of task, input representation, and network architecture. Rather than identifying emitters from raw time-domain I/Q signals, NeuRFP focuses on authenticating beamformers in MU-MIMO networks from compressed frequency-domain CSI feedback matrices. While [12] also exploits CSI, it does so for channel prediction within data augmentation, and its fingerprinting input remains the raw I/Q signal. Architecturally, [15] retains complex-valued arithmetic (each complex product expands into four real matrix products) and max-pooling: both are multiply-heavy, non-event-driven operations that neuromorphic hardware is meant to avoid. Instead, we handle the complex-valued CSI data by decoupling the I/Q components into parallel real-valued input channels, as detailed in Section III. This choice allows spiking dynamics to extract phase and amplitude signatures without the computational overhead of complex-valued arithmetic. Lastly, as shown in Section III-B, unlike the convolutional spiking topology employed in [12], [15], NeuRFP leverages a fully connected structure. As we demonstrate in our theoretical analysis, this architectural choice deliberately bypasses the memory bottleneck caused by the continuous time-driven tracking of the neurons' membrane states in deep convolutional SNNs, shrinking the total number of spikes and the overall memory traffic.

III. The NeuRFP Framework

We consider modern MU-MIMO networks, where a beamformer transmits multiple data streams toward different associated beamformees. Our objective is to obtain a fingerprint of the beamformer from the CSI describing the links connecting the beamformer with the beamformees. This information is captured over the air during the MU-MIMO setup procedure, which is summarized next.

A. MU-MIMO Channel Sounding and Precoding

The beamformer relies on the knowledge of the CSI for all the different communication links to effectively multiplex data toward the different beamformees and minimize inter-stream and inter-user interference. This beam shaping operation is referred to as *precoding* while the CSI is obtained through a preliminary procedure named *channel sounding* [8]. The latter is repeatedly executed in wireless networks operating in MU-MIMO mode to assure freshness in the channel estimates available to the beamformer. Indeed, the CSI used for precoding should match the actual channel to allow proper distortion compensation. Channel sounding is triggered by the beamformer that broadcasts a dedicated sounding frame containing pilot signals known by the beamformees. By comparing the symbols in the received frame with the known ones, each beamformee estimate the CSI, which is then fed back to the beamformer for precoding. Beamformees estimate the CSI and feed it back using compression techniques to reduce overhead, such as singular value decomposition (SVD)-derived quantized polar angles (ϕ, ψ) in IEEE 802.11 Wi-Fi, or Type II Codebooks in 5G New Radio [16]. Regardless of the compression strategy, this feedback is typically transmitted in unencrypted frames to minimize latency, allowing any passive wireless observer to collect and decode it without

decryption keys. Crucially, hardware imperfections inherent to the beamformer percolate into the sounding frames and, subsequently, into the compressed CSI feedback [8]. This makes unencrypted CSI an ideal signal for physical-layer authentication. While the unencrypted nature of the CSI feedback allows any device to act as a passive observer, a practical authentication deployment requires a legitimate monitoring node. In this view, we assume the presence of a TE node that continuously captures the CSI feedback matrices broadcasted by multiple beamformees and aggregates them into a unified input tensor, in line with the approach adopted in the DeepCSI framework [8]. In the following, we detail our new neuromorphic-based strategy, named NeuRFP, that operates on the TE to extract such fingerprints directly from CSI, enabling accurate authentication with low complexity.

B. Neuromorphic Fingerprinting Architecture

The NeuRFP algorithm takes as input the in-phase (I) and quadrature (Q) components of the CSI for all the K orthogonal frequency-division multiplexing (OFDM) subcarriers and each pair of receiver–transmitter antennas ($N \times M$), which are extracted and stacked to form a multidimensional input tensors. To evaluate the trade-offs between topological complexity and energy footprint, we design and compare two distinct architectures: a convolutional SNN (NeuRFP-C) and a multi-layer perceptron SNN (NeuRFP-D). The NeuRFP-C serves as the direct spiking equivalent of the DeepCSI baseline [8]. It consists of a convolutional block with five one-dimensional convolutional layers, followed by a dense block. The first two layers use a kernel size of 7 with a stride of 2 to achieve spatial downsampling; the subsequent three layers use a stride of 1 with progressively shrinking kernel sizes (7, 5, and 3, respectively) and explicit padding to preserve spatial resolution. The extracted feature maps are then flattened and passed through a two-layer dense block (with 128 and 64 neurons per layer, respectively), before reaching the classification layer. Conversely, the proposed NeuRFP-D explores a structurally simplified approach to investigate whether deep spatial feature extraction is strictly necessary for CSI data, discarding the convolutional layers and flattening the multidimensional input tensor into a single vector. As shown in Figure 2, this is then processed through a three-layer dense block (with 256, 128, and 64 neurons, respectively) that leads directly into the classifier.

Regardless of the chosen topology, every hidden spatial convolution or dense linear projection is followed by a 1D batch normalization layer and a spiking neuron. We employ the leaky integrate-and-fire (LIF) spiking neuron model, which mimics biological membranes accumulating incoming signals while gradually leaking charge. The sub-threshold membrane potential U for a given LIF is governed by

$$U[t] = \beta \cdot U[t-1] + W \cdot X[t] - S_{\text{out}}[t-1] \cdot \theta \quad (1)$$

where β is the decay rate factor (which we set as a learnable parameter), W denotes the learnable synaptic weights, $X[t]$ represents the input features at time t, and θ is the

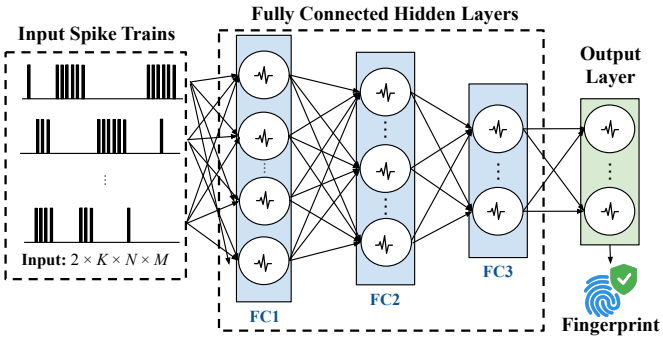

Fig. 2: NeuRFP-D learning algorithm. The I and Q components of the CSI are jointly processed by a multi-layer perceptron SNN that extracts a fingerprint of the beamformer.

firing threshold [17]. A binary spike $S_{\text{out}}[t]$ is emitted if $U[t] > \theta$, followed by a *reset-by-subtraction* mechanism that decreases the value of the membrane potential by θ to prevent excessive voltage accumulation. To generate the final network prediction, the output layer omits batch normalization and utilizes non-resetting, non-leaky LIF neurons ($\beta = 1$) to act as pure temporal integrators for classification. To overcome the non-differentiability of the spiking mechanism, which prevents the use of standard backpropagation during training, we adopt a surrogate gradient learning framework [18]. This method introduces a continuous relaxation of the discrete step function, enabling gradient-based optimization. Specifically, we adopt the arctangent (ATan) function as a surrogate derivative to approximate the spike nonlinearity, ensuring stable gradient flow during backpropagation.

C. Spiking Encoding

To map the CSI feedback matrices into the spiking domain, we evaluated two approaches: *Rate Encoding* and *Direct Encoding*. Traditional Rate Encoding stochastically converts continuous inputs into spike trains over a temporal window [19]. However, this stochastic generation delays network convergence and inflates the required inference latency. Direct Encoding mitigates this by using the normalized I/Q components directly as continuous analog inputs to the first layer, transforming it into a trainable end-to-end spike encoder [11], [20]. Despite this, directly injecting raw analog data into an unscaled layer yields massive accumulated voltages, instantly saturating subsequent LIF neurons and causing uncontrolled over-firing. To maintain a balanced operational regime, we introduce a static current multiplier immediately following the first batch normalization layer. This explicitly scales the input magnitude relative to the neuronal firing threshold, ensuring that the first layer integrates properly over time without saturating the downstream network [19].

D. Considerations on ANN- vs SNN-based Fingerprinting

The DeepCSI architecture proposed in [8] based on ANNs achieves high classification accuracy but relies on operations

that are hostile to event-driven neuromorphic hardware: max-pooling and attention mechanisms. In the spiking domain, as the *maximum* operation is non-linear, spatial max-pooling cannot be trivially computed on a spike-by-spike basis without introducing multi-pass memory buffering and significant hardware overhead [21], [22]. To preserve the asynchronous neuromorphic paradigm, we avoid using max-pooling in NeuRFP designs and achieve spatial downsampling by using strided 1D convolutions (e.g., with stride of 2). Furthermore, the original attention module relies on global feature aggregation, softmax normalization, and continuous-valued matrix multiplications. Executing these operations on neuromorphic hardware forces the system to abandon its multiplier-free binary accumulation logic in favor of costly MAC operations [23]. For these reasons, we omit the attention block entirely.

IV. Experimental Evaluation Setup

To evaluate the performance of the proposed NeuRFP approach, we employ the open-source DeepCSI dataset [8]. The experimental setup consists of an IEEE 802.11ac Wi-Fi access point (the target beamformer) implemented using a Compex WLE1216v5-23 module operating at 5.21 GHz with 80 MHz of bandwidth, and two Netgear Nighthawk X4S Wi-Fi devices acting as stations (the beamformees). Three antennas are enabled at the beamformer while the beamformees operate with one antenna each. The dataset contains CSI data for 10 distinct Compex Wi-Fi modules to be fingerprinted. Each sample collects the beamforming feedback angles for each of the $K = 234$ OFDM subcarriers over the 3×2 MIMO links. For data collection, the beamformees are progressively moved across nine distinct spatial positions to generate spatial diversity. We preprocess the dataset following the same pipeline described in [8] to reverse the IEEE 802.11 CSI compression protocol and reconstruct the complex-valued CSI corresponding to the target transmitter. Hence, we evaluate the performance in two main operational scenarios following the DeepCSI evaluation approach. For the first scenario (S1) all the different positions are considered during training, and testing is performed across all of them. The second scenario (S2) is more challenging as only data collected in positions with odd indices are considered during training while testing is performed on the even positions.

All SNN architectures are implemented in Python using the Pytorch framework alongside the `snnTorch` library for neuromorphic simulations, and trained on an NVIDIA QUADRO RTX6000 GPU. For the spiking dynamics, we simulate the network over an inference window of $T = 15$ discrete timesteps. The models are trained using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 64, minimizing the standard cross-entropy loss function. To ensure robust evaluation and prevent overfitting, we allow a maximum of 250 training epochs but implement an early stopping mechanism with a patience of 25 epochs evaluated on the validation loss. The dataset for each position is split into 64% for training, 16% for validation, and 20% for testing. The analytical energy model we adopt for performance–complexity comparison is detailed next.

A. Energy Consumption Model

To evaluate the efficiency of the proposed SNN against the baseline DeepCSI ANN architecture, we adopt the hardware-agnostic analytical energy framework formulated by Lemaire et al. [11], which decouples the total energy consumption per inference (E_{tot}) into memory access (E_{mem}) and computational/addressing costs ($E_{\text{ops+addr}}$). The model assumes standard 45nm CMOS semiconductor technology (0.9V), with energy costs for 32-bit operations E_{ADD} = 0.1 pJ (addition), E_{MUL} = 3.1 pJ (multiplication), E_{SRAM} = 20 pJ (memory access). The memory energy consumption (E_{mem}) is obtained as the product of E_{SRAM} and the total count of read/write operations for inputs ($RdIn$), parameters ($RdParam$), layer outputs ($WrOut$), and internal membrane potentials ($RdPot, WrPot$):

$$\begin{aligned} E_{\text{mem}} =& (RdIn + RdParam + RdPot) \times E_{\text{SRAM}} \\ &+ (WrOut + WrPot) \times E_{\text{SRAM}} \ . \end{aligned} \tag{2}$$

The compute/addressing energy is modeled as:

$$\begin{aligned} E_{\text{ops+addr}} =& (E_{\text{ADD}} + E_{\text{MUL}}) \times MAC_{\text{ops+addr}} \\ &+ E_{\text{ADD}} \times ACC_{\text{ops+addr}} \ , \end{aligned} \tag{3}$$

where $MAC_{\text{ops+addr}}$ and $ACC_{\text{ops+addr}}$ respectively indicate the number of MAC and accumulate (ACC) operations. These contributions are obtained as follows.

MAC Operations: In ANNs, these capture dense synaptic multiplications executed across all convolutional and fully-connected layers. While ANNs rely on dense MAC operations due to continuous-valued activations, SNNs exploit binary events ($S \in \{0, 1\}$), reducing costly multiplications to sparse ACC operations [23]. Thus, for SNNs, MAC operations are restricted to the first convolutional layer, where continuous analog inputs are encoded into spiking activity [23], and the internal membrane potential leakage (β), where a single multiplication per output neuron is performed at each timestep.

ACC Operations: In ANNs, they primarily represent bias additions. For SNNs, they entail four distinct contributions for any given layer: (i) synaptic weights accumulation triggered by input spikes (θ_{l-1}); (ii) continuous bias integration at each timestep over the layer's output dimension; (iii) membrane potential resets upon spike generation (θ_l); and (iv) hardware addressing overhead for sparse spike routing [23].

Importantly, this analytical framework models all memory accesses as local SRAM operations, creating a "best case" scenario for ANNs, as they often necessitate off-chip DRAM accesses. Conversely, this model represents a "worst case" estimate for SNNs as it does not account for the co-location of memory and computation inherent to actual neuromorphic computing-in-memory (non-Von Neumann) architectures.

To contextualize the energy efficiency of the evaluated SNNs, we track the spike activity rate (SAR) and the total spike volume. The SAR represents the average percentage of

neurons emitting a spike per timestep. However, the energy consumption is governed by the volume of spikes, which is proportional to the product of the network's SAR, the total number of neurons, and the length of the inference window (T). A compact topology can exhibit a high SAR while still generating exponentially fewer spikes than a vast convolutional architecture with a low SAR.

V. Results and Trade-off Analysis

We evaluate the proposed NeuRFP-C and NeuRFP-D approaches by performing a parametric sweep over the simulation timesteps ($T \in \{1, 2, 4, 8, 10, 12, 15\}$). The networks are benchmarked against the original non-spiking DeepCSI framework [8] in terms of classification accuracy, computational operation counts (MACs vs. ACCs) and total theoretical energy consumption (derived via the Lemaire profiling framework described in Section IV-A). To establish a rigorous reference point, we first quantified the algorithmic complexity of this ANN baseline. The DeepCSI network consists of 489k trainable parameters. Based on our architectural analysis, a single forward pass requires approximately 24.3 million MACs to process the convolutional and dense layers, accompanied by nearly 48.7 million memory accesses. When including auxiliary additions (ACCs), the baseline executes roughly 24.45 million core arithmetic operations per inference. Under the Lemaire framework, this intensive operational and memory workload translates to a theoretical energy footprint of 1051.78 μJ.

TABLE I: NeuRFP spike activity and accuracy ('Acc.') across architectures and encoding schemes ('DE' stands for Direct Encoding, 'RE' stands for Rate Encoding).

Architecture	**SAR (%)**	**Spikes**	**Acc.** ($T = 1$)	**Peak Acc.**
NeuRFP-C (DE)	14.19	51,540	91.79%	98.27% ($T = 15$)
NeuRFP-C (RE)	17.72	64,326	87.75%	97.69% ($T = 15$)
NeuRFP-D (DE)	22.73	833	98.05%	98.34% ($T = 2$)
NeuRFP-D (RE)	20.51	751	85.16%	96.97% ($T = 15$)
DeepCSI (ANN)	N/A	N/A	N/A	98.02%

We then evaluate the impact of the encoding strategy on the network's activity and time-to-accuracy trade-off. Table I summarizes the results across the architectures. Although NeuRFP-D with Direct Encoding exhibits a higher overall SAR (22.73%) compared to Rate Encoding (20.51%), such encoding strategy allows NeuRFP-D to achieve 98.05% accuracy at the very first integration step ($T = 1$). In contrast, Rate Encoding starts at 85.16% and requires prolonged temporal integration ($T = 15$) to reach the same accuracy. Moreover, at $T = 8$, NeuRFP-D generates fewer than 900 total spikes per inference. As mathematically defined in Section IV-A, despite having a higher SAR, the compact topology of NeuRFP-D results in a nearly 60-fold reduction in absolute spike volume compared to NeuRFP-C. Consequently, NeuRFP-D coupled with Direct Encoding emerges as the optimal design choice.

The computational complexity of the proposed NeuRFP-C and NeuRFP-D approaches compared to the DeepCSI baseline is reported in Figure 3 and reveals differences spanning multiple orders of magnitude. Although NeuRFP-C successfully converts MACs operations into lighter ACCs, as the inference timesteps increase ($T = 8$), memory accesses exceed 92 million. This phenomenon is due by the continuous hardware requirement to fetch and update membrane potentials across all spatial feature maps iteratively. Consequently, the NeuRFP-C becomes significantly more memory-intensive than the DeepCSI baseline (with only 48.7 million accesses). Conversely, by flattening the spatial dimensions into a dense projection, the proposed NeuRFP-D topology at its optimal operating point ($T = 2$) collapses both arithmetic operations and memory accesses well below the one-million mark, reducing the overall computational and memory footprint by nearly two orders of magnitude compared to the ANN baseline.

Table II provides a detailed layer-by-layer breakdown of the memory accesses, showcasing the hardware bottlenecks of the evaluated models. The results reveal that the NeuRFP-C topology suffers from a severe escalation in memory traffic, especially within its deep convolutional layers. In stark contrast, the proposed NeuRFP-D confines the majority of its memory footprint to the initial encoding phase occurring at Layer 1. By eliminating deep convolutional feature maps, the subsequent dense layers can fully exploit event-driven sparsity. Consequently, the overall memory accesses of the NeuRFP-D plummet to merely 846,000 (at $T = 8$), reducing the total traffic by a factor of over 100 compared to the overwhelming 92.2 million accesses required by the NeuRFP-C topology. To fully comprehend the structural inefficiency of the convolutional approach, it is crucial to isolate the memory traffic dedicated to the LIF state dynamics; specifically, the continuous fetching (*rd_pot*) and updating (*wr_pot*) of membrane potentials. At $T = 8$, where both architectures achieve comparable accuracy ($\sim$98%), the NeuRFP-C requires an overwhelming 92.22 million memory accesses, of which 60.7 million (roughly 66%) are entirely dissipated on *rd_pot* and *wr_pot* cycles to maintain the spatial neurons' internal states. In stark contrast, the NeuRFP-D requires only 846k total accesses, with 163k (19%) dedicated to membrane updates. These results prove that eliminating deep spatial feature maps is mandatory to bypass the state-tracking bottleneck.

The accuracy-energy scaling dynamics across different inference timesteps (T) for NeuRFP-C and NeuRFP-D is reported in Figure 4 for Scenario S1 (left) and Scenario S2 (right), in comparison with DeepCSI. In SNNs, synaptic operations benefit from event-driven sparsity. However, LIF membrane updates are strictly time-driven. The forward step at every timestep requires reading, updating, and writing the membrane potentials of all neurons, regardless of whether they spike or not. To properly visualize the impact of this phenomenon, the total analytically estimated energy is plotted on a logarithmic scale in the figures (right y-axis). In Scenario S1 (Figure 4, left), the proposed NeuRFP-D demonstrates a peak accuracy of 98.34% (at $T = 2$), marginally outperforming the ANN baseline (98.02%), while also residing at the absolute bottom of the energy spectrum ($\sim$14 μJ). Con-

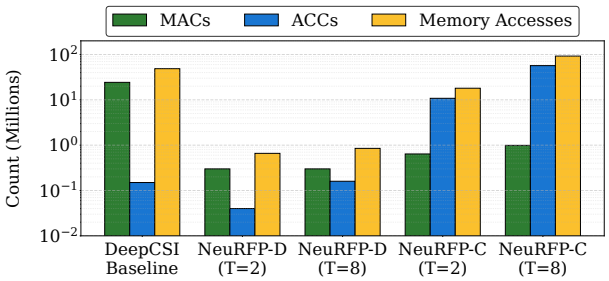

Fig. 3: Comparison of computational operations (MACs, ACCs) and memory accesses among the DeepCSI baseline [8], and our proposed NeuRFP-D and NeuRFP-C architectures at different timesteps (T). The plot is in logarithmic scale.

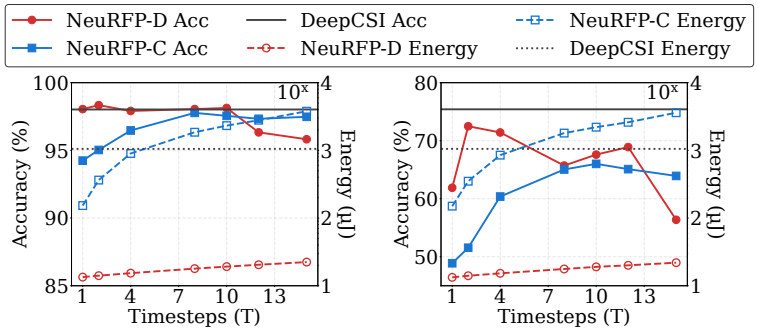

Fig. 4: NeuRFP classification accuracy (left axes) and total analytical energy scaling (right axes, logarithmic) across inference timesteps (T), compared against the DeepCSI baseline in (left) Scenario S1, and (right) Scenario S2.

TABLE II: Layer-by-layer memory accesses comparison across different architectures and timesteps.

Layer	**DeepCSI [8]**	**NeuRFP-D** ($T=2$)	**NeuRFP-D** ($T=8$)	**NeuRFP-C** ($T=2$)	**NeuRFP-C** ($T=8$)
Layer 1 (Conv1 / FC1 / Conv1)	2,201,472	600,943	605,938	1,154,426	1,433,052
Layer 2 (Conv2 / FC2 / Conv2)	26,889,344	43,775	195,112	3,451,957	27,815,390
Layer 3 (Conv3 / FC3 / Conv3)	13,329,792	10,221	41,062	4,981,761	24,632,653
Layer 4 (Conv4 / Out / Conv4)	4,764,288	1,330	4,018	4,625,931	21,816,397
Layer 5 (Conv5 / – / Conv5)	1,382,528	–	–	2,829,443	11,907,195
Attention Block	7,297	–	–	–	–
FC1	115,840	–	–	985,977	4,579,589
FC2	8,448	–	–	9,224	36,898
Classifier Output	724	–	–	1,235	4,112
TOTAL	48,698,883	656,269	846,131	18,039,957	92,225,289

versely, the NeuRFP-C architecture requires longer temporal integration ($T \geq 8$) to converge to comparable accuracy. As T increases, the relentless accumulation of time-driven memory accesses completely overwhelms any energy savings gained from sparse activations. Consequently, the NeuRFP-C's theoretical energy footprint grows exponentially, crossing the ANN baseline's energy budget of 1051.78 μJ at $T \approx 4$. This structural superiority is further validated under the spatial generalization task (Scenario S2, Figure 4, right), where the overall classification accuracy inevitably degrades for all models due to the challenge of interpolating unseen spatial coordinates. Despite this, the NeuRFP-D reaches peak accuracy (72.49%) at $T=2$, closely approaching the non-spiking ANN baseline (75.41%) while significantly outperforming the NeuRFP-C approach and maintaining its low energy profile. One careful consideration concerns the NeuRFP scalability to larger MU-MIMO configurations. As the number of antennas, subcarriers, or beamformees increases, the dimension of the flattened CSI input grows proportionally. Consequently, the number of synaptic weights might increase significantly, potentially leading to higher memory consumption.

VI. Concluding Remarks

In this paper, we introduced a new paradigm for RFP based on neuromorphic processing. Our framework, named NeuRFP, relies on SNNs to process the CSI data and obtain a fingerprint of the transmitter device (beamformer), consuming fewer computation resources than traditional ANN-based approaches. We analyzed two different spiking topologies, demonstrating that deep convolutional SNNs (NeuRFP-C) suffer from a memory bottleneck caused by the continuous time-driven tracking of the neurons' membrane states. We bypass this limitation by flattening the input features into a dense topology (NeuRFP-D), shrinking the total spike volume and memory traffic. We assessed the performance of NeuRFP when fingerprinting 10 MU-MIMO beamformers and show that it matches or closely approaches the same accuracy as a state-of-the-art approach based on ANN while reducing theoretical energy consumption by two orders of magnitude. Despite these promising results, our study has limitations that outline directions for future work. First, the estimated energy savings rely on an analytical hardware model; future studies will focus on deploying NeuRFP on actual neuromorphic platforms, such as SpiNNaker, to validate the efficiency in real-world scenarios. Furthermore, while we evaluated spatial generalization, radio fingerprints often suffer from temporal drift due to, for example, environmental variations. Investigating the temporal robustness of NeuRFP for long-term deployments and exploring continual learning strategies are key objectives for future research avenues. To foster reproducibility and further studies, the implementation code is available at https://github.com/Restuccia-Group/NeuRFP-MILCOM.

ACKNOWLEDGMENT

This work was supported in part by NSF under Grant OAC-2530896, ARO under Grant W911NF261A165, ONR under Grants N00014-23-1-2221 and N00014-26-1-2376, and AFOSR under Grant FA9550-23-1-0261.

REFERENCES

[1] Federal Communications Commission (FCC), "Spectrum Crunch," 2020.

[2] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten Myths and One Critical Question," *arXiv preprint arXiv:1503.06854v2*, August 2015.

[3] F. Meneghello, F. Gringoli, M. Cominelli, M. Rossi, and F. Restuccia, "How to BREAK MU-MIMO Precoding in IEEE 802.11 Wi-Fi Networks," in *Proc. of IEEE INFOCOM*, (London, UK), 2025.

[4] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1727–1765, 2016.

[5] J. Zhang, F. Ardizzon, M. Piana, G. Shen, and S. Tomasin, "Physical layer-based device fingerprinting for wireless security: From theory to practice," *IEEE Transactions on Information Forensics and Security*, 2025.

[6] B. Razavi, *RF Microelectronics*. Upper Saddle River, NJ: Prentice Hall, 2nd ed., 2011.

[7] J. Hua, H. Sun, Z. Shen, Z. Qian, and S. Zhong, "Accurate and efficient wireless device fingerprinting using channel state information," in *Proc. of IEEE INFOCOM*, 2018.

[8] F. Meneghello, M. Rossi, and F. Restuccia, "DeepCSI: Rethinking Wi-Fi Radio Fingerprinting Through MU-MIMO CSI Feedback Deep Learning," in *Proc. of IEEE ICDCS*, 2022.

[9] F. Meneghello, K. F. Haque, and F. Restuccia, "Radio Fingerprinting of Wi-Fi Devices Through MIMO Compressed Channel Feedback," in *IEEE INFOCOM WKSHPS*, 2025.

[10] K. Sankhe, M. Belgiovine, F. Zhou, L. Angioloni, F. Restuccia, S. D'Oro, T. Melodia, S. Ioannidis, and K. Chowdhury, "No Radio Left Behind: Radio Fingerprinting Through Deep Learning of Physical-Layer Hardware Impairments," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, pp. 165–178, Mar. 2020.

[11] E. Lemaire, P.-E. Novac, L. Cordone, J. Courtois, A. Castagnetti, and B. Miramond, "An Analytical Estimation of Spiking Neural Networks Energy Efficiency," in *Proc. of ICONIP*, 2022.

[12] Q. Jiang and J. Sha, "Channel-Resilient and Low-Power Radio Frequency Fingerprint Identification for VDES-SAT," *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, pp. 2161–2174, Aug. 2025.

[13] S. Riyaz, K. Sankhe, S. Ioannidis, and K. Chowdhury, "Deep Learning Convolutional Neural Networks for Radio Identification," *IEEE Communications Magazine*, vol. 56, pp. 146–152, Sept. 2018.

[14] R. Kong and H. Chen, "CSI-RFF: Leveraging micro-signals on CSI for RF fingerprinting of commodity WiFi," *IEEE Transactions on Information Forensics and Security*, 2024.

[15] N. Zhang, Y. Zhang, C. Xu, W. Zhang, H. Zha, Y. Han, Q. Tian, and Z. Tang, "SComplex: Advancing Specific Emitter Identification Through Deep Spiking Complex Neural Networks Beyond Traditional Lightweight Approaches," in *Proc. of IEEE/CIC ICCC*, 2024.

[16] 3GPP, "NR; Physical layer procedures for data (Release 15)," Technical Specification (TS) 38.214, 3rd Generation Partnership Project (3GPP), 2018. Version 15.0.0.

[17] J. K. Eshraghian, M. Ward, E. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proceedings of the IEEE*, vol. 111, no. 9, pp. 1016–1054, 2023.

[18] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019.

[19] P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *2015 International joint conference on neural networks (IJCNN)*, pp. 1–8, ieee, 2015.

[20] N. Rathi and K. Roy, "DIET-SNN: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3752–3763, 2021.

[21] B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu, "Conversion of Continuous-Valued Deep Networks to Efficient Event-Driven Networks for Image Classification," *Frontiers in Neuroscience*, vol. 11, p. 682, 2017.

[22] A. Dorzhigulov and V. Saxena, "Spiking CMOS-NVM mixed-signal neuromorphic ConvNet with circuit-and training-optimized temporal subsampling," *Frontiers in Neuroscience*, vol. 17, p. 1177592, 2023.

[23] M. Yao, J. Hu, Z. Zhou, L. Yuan, Y. Tian, B. Xu, and G. Li, "Spike-driven Transformer," in *Proc. of NeurIPS*, 2023.