EPiC Series in Technology

Volume 3, 2026, Pages 103–107

Proceedings of AI for Accelerated Research Symposium

Design of a Cross-Layer AI Agent for Secure Spectrum-Aware Network Slicing

Guiran Liu, Binrong Zhu, Yang Liu, and Qun Wang

Computer Science Department, San Francisco State University, San Francisco, CA 94132
{gliu,bzhu2, yliu68, qunwang}@sfsu.edu

Abstract

The sixth-generation (6G) network will enable applications from holographic communication and XR to massive IoT and autonomous systems. Network Slicing (NS) is central to this vision, creating isolated networks on shared infrastructure with tailored QoS. Yet current AI-driven NS overlooks the volatile wireless spectrum: network agents lack spectrum awareness, while sensing mechanisms miss network context and security posture. This disconnect causes inefficient resource use, vulnerability to cross-layer attacks, and reactive management. We propose CogniSense-Slice, an AI agent unifying spectrum awareness with secure, application-aware slice management. Its cognitive feedback loop links the physical and network layers, enabling proactive, efficient slicing. The architecture integrates three modules: a Perception Module for fast spectrum sensing, an Orchestration Module with on-device LLMs for holistic management, and a Guardian Module for cross-layer threat mitigation. By bridging intelligence across the stack, CogniSense-Slice shifts from siloed optimization to holistic cognition.

1 Introduction

The sixth-generation (6G) network is envisioned to support diverse services—from XR and holographic communication to massive IoT and autonomous systems—each with unique QoS demands. Network Slicing (NS) has been identified as a foundational technology, enabling multiple virtual, isolated networks on shared infrastructure, customized for eMBB, URLLC, or mMTC [4]. Managing these dynamic slices requires automation, with AI as the key enabler [6].

Yet current AI-driven slicing suffers from a critical "spectrum blind spot." While effective at optimizing high-level KPIs, these systems remain unaware of real-time spectrum conditions, leading to: (i) sub-optimal resource use by missing transient opportunities, (ii) vulnerability to cross-layer threats such as jamming and spoofing [6], and (iii) a reactive paradigm that addresses degradation only after it occurs.

At the network and application layers, AI frameworks like DeepSlice use DLNNs to allocate slices (eMBB, URLLC, mMTC) by analyzing KPIs such as QCI, delay budget, and packet loss [4]. Secure5G adds a security layer, detecting anomalous traffic and isolating malicious devices in a "Quarantine Slice" [1]. While effective in managing traffic, these systems remain blind to spectral conditions. For example, an XR slice experiencing latency cannot distinguish between congestion and interference, leaving responses reactive and incomplete.

J. Masnec, H.R. Karimian, P. Kordjamshidi and Y. Li (eds.), AIAS 2025 (EPiC Series in Technology, vol. 3), pp. 103–107

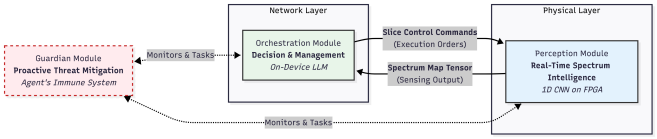

Figure 1: CogniSense-Slice Agent Framework

At the PHY/MAC layers, frameworks like DeepSense achieve sub-ms wideband sensing by mapping spectrum occupancy with CNNs on FPGAs [5]. Agentic architectures, such as LLM-assisted spectrum sharing, further separate high-level reasoning from low-level optimization, ensuring safety and interpretability [2]. Yet these systems face the inverse challenge: they detect spectrum availability but lack awareness of network intent. DeepSense may find a clean 20 MHz channel, but without context, it cannot prioritize whether it serves a URLLC slice for industrial control or an eMBB slice for video. Thus, optimization misaligned with 6G's diverse and mission-critical needs.

This paper introduces the CogniSense-Slice Agent, a cross-layer AI framework that creates a cognitive feedback loop between the physical and network layers, enabling proactive, secure, and efficient slicing. This vision aligns with AI-native 6G, where intelligence is embedded across all layers of the stack [6].

2 The CogniSense-Slice Agent: A Vision for Cross-Layer Cognition

The critical gap in the path toward intelligent 6G networks is the absence of a unified framework that enables seamless information flow between the intelligent systems operating at different layers of the network stack. To bridge this divide, we propose the CogniSense-Slice agent, a unified AI architecture that integrates sensing, orchestration, and security across the physical and network layers as shown in figure 1. The CogniSense-Slice agent is composed of three synergistic modules that form a cognitive loop. The Perception Module senses the physical environment, the Orchestration Module reasons and decides on resource allocation, and the Guardian Module ensures the security and integrity of the entire process.

2.1 The Perception Module: Real-Time Spectrum Intelligence

The Perception Module serves as the agent's sensory interface, delivering a continuous, high-resolution view of the spectral environment. Its design is directly based on the proven DeepSense framework [5]. It employs a lightweight 1D CNN, implemented in hardware (FPGA) at the baseband processing level, to analyze raw I/Q samples directly from the radio front-end. This design enables sub-millisecond inference (as low as 0.61 ms), essential for exploiting transient spectrum opportunities and meeting 6G URLLC latency demands. The CNN is trained to identify not just whether a channel is busy or idle, but also to characterize the spectral environment with high granularity. It can detect subtle interference patterns, identify transient "spectrum holes" that traditional methods would miss, and even learn the unique spectral signatures of different transmission technologies. The module's output is a structured tensor representing the real-time spectrum occupancy map. This format serves as a standardized, machine-readable input for the agent's higher-level cognitive functions.

2.2 The Orchestration Module: Spectrum-Aware Slice Management with ODLLM

The Orchestration Module is the agent's cognitive core responsible for making intelligent and proactive decisions about network slice management. This module leverages an On-Device Large Language Model (ODLLM) running at the network edge (e.g., within a gNodeB) [7]. ODLLMs will be augmented with techniques like Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) prompting to ground their reasoning in domain-specific knowledge [3].

The core innovation of this module is its ability to perform cross-layer information fusion. The ODLLM is designed to process a composite input that combines the flattened spectrum map tensor from the Perception Module with traditional network-layer KPIs (e.g., traffic load, latency metrics, device type, application requirements) sourced from a knowledge base [4]. By reasoning over this cross-layer state, the ODLLM can execute proactive slice management strategies: **Opportunistic Allocation:** Upon receiving a report from the Perception Module about a newly available 100 MHz channel in the 6 GHz band, the Orchestration Module can correlate this with network KPIs showing a pending request for a high-throughput XR slice. It can then reason that this new resource is an ideal match and immediately provision the XR slice on that channel [4]. **Proactive Migration:** If the Perception Module reports rising interference levels on a 3.5 GHz band currently serving a critical URLLC slice, and network KPIs confirm that the slice's latency is approaching its SLA limit, the agent can proactively migrate the entire slice to a more stable licensed band to guarantee its reliability before a service disruption occurs.

2.3 The Guardian Module: Proactive Cross-Layer Threat Mitigation

The Guardian Module provides a security paradigm that transcends individual layers by correlating information across the stack. It combines rule-based systems with AI-driven anomaly detection, which continuously analyzes the outputs and decisions of the Perception and Orchestration modules. The process for identifying such a threat is as follows:

The Orchestration Module utilizes its network-level knowledge to identify a connected device as a low-power "mMTC IoT sensor," with an expected traffic profile characterized by infrequent, low-bandwidth data bursts. The Orchestration Module then observes this same device attempting to transmit a massive amount of data, a clear network-layer behavioral anomaly. The Guardian Module detects this discrepancy and tasks the Perception Module to perform a fine-grained spectral analysis of the specific signal being transmitted by the anomalous device. The Perception Module analyzes the raw I/Q signature of the transmission and reports that it does not match the expected signature of a low-power IoT device. Instead, it exhibits the characteristics of a powerful, wideband jammer or a spoofed high-power user equipment. The Guardian Module correlates these two distinct alerts: a network-layer behavioral anomaly (anomalous traffic pattern) and a physical-layer signature anomaly (anomalous spectral signature). This cross-layer correlation provides a higher-confidence detection of a malicious actor than either layer could achieve in isolation.

Upon high-confidence detection, Guardian Module triggers an automated mitigation response. The malicious device is immediately shunted to the "Quarantine Slice" where its resources are severely restricted, effectively neutralizing the threat before it can propagate and impact the core network or other legitimate slices [1].

2.4 Proof-of-Concept Evaluation

As shown in Table 1, we validate CogniSense-Slice via 300-step simulation across normal/attack/recovery phases. Guardian (Isolation Forest, 40 samples) achieves 98.9% detection with 16.2% false alarms. Local-LLM (Phi-3 mini) achieves 455 ms latency and perfect attack isolation (100%), representing a 2.57× improvement over threshold-based (38.9%). Cloud-LLM (GPT-4), despite the superior model capacity, achieves only 57.8% isolation due to 1734ms latency-induced temporal misalignment ($\chi^2 = 145.7$, $p < 0.00001$).This validates edge deployment as mandatory for time-sensitive orchestration.

Table 1: Performance Summary

Metric	**Rule-based**	**Local-LLM**	**Cloud-LLM**
Decision Latency	0.0003 ms	**455 ms**	1734 ms
Attack Isolation	38.9%	**100%**	57.8%
Decision Accuracy	70.7%	**93.0%**	87.3%

3 Research Roadmap and Open Challenges

While the vision for CogniSense-Slice builds on existing technologies, its realization faces several key challenges.

Cross-Layer Data Fusion and Datasets A major barrier is the lack of synchronized, multimodal datasets linking PHY-layer signals with MAC scheduling, traffic flows, and application KPIs. Large-scale datasets—capturing raw I/Q samples aligned with network events—are essential. High-fidelity simulators (e.g., ns-3) and SDR testbeds must be leveraged to generate them.

Real-Time On-Device Inference The full cognitive loop must meet 6G URLLC sub-ms latency budgets, but deploying CNNs and ODLLMs at the edge raises computational and energy challenges. Research must advance model compression (quantization, pruning, distillation) and explore SoCs integrating FPGAs for PHY tasks with NPUs for AI inference.

Scalability and Multi-Agent Coordination A single-agent model cannot scale to dense multi-cell networks. Coordinating multiple agents for inter-cell allocation, slice handovers, and wide-area threat response—without conflicts or bottlenecks—demands advances in multi-agent reinforcement learning (MARL).

Security and Robustness The agent itself is a new attack surface. CNNs face adversarial perturbations in I/Q signals; ODLLMs risk prompt injection or data poisoning. Future work must enable self-monitoring and safe-mode fallback to ensure resilience.

4 Conclusion

The CogniSense-Slice agent marks a shift from isolated, layer-specific AI to a holistic, cross-layer cognitive architecture. The current divide between network-layer context and physical-layer dynamics fosters inefficiency, security blind spots, and reactive management. By unifying these layers, CogniSense-Slice enables dynamic spectrum use, proactive adaptation to channel conditions, and cross-layer threat detection. This vision calls on the community to move beyond

single-layer optimization and design integrated agents—an essential step toward realizing the intelligent, efficient, and secure potential of 6G networks.

References

[1] Cross-layer security for 5g/6g network slices: An sdn, nfv, and ai-based hybrid framework. *ResearchGate*, 2025.

[2] Lifu Gao, Joshua Sherwood, Nawwaf Aleisa, Andrews Damoah, Yingzhou Lu, and Xiaodong Qu. Human-centered ai agents for healthcare and education: A systematic literature review.

[3] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

[4] Anurag Thantharate, Rahul Paropkari, Vijay Walunj, and Cory Beard. Deepslice: A deep learning approach towards an efficient and reliable network slicing in 5g networks. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0762–0767. IEEE, 2019.

[5] Daniel Uvaydov, Salvatore D'Oro, Francesco Restuccia, and Tommaso Melodia. Deepsense: Fast wideband spectrum sensing through real-time in-the-loop deep learning. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021.

[6] Qun Wang, Haijian Sun, Rose Qingyang Hu, and Arupjyoti Bhuyan. When machine learning meets spectrum sharing security: Methodologies and challenges. *IEEE Open Journal of the Communications Society*, 3:176–208, 2022.

[7] Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. Twin-gpt: Digital twins for clinical trials via large language model. *ACM Trans. Multimedia Comput. Commun. Appl.*, July 2024. Just Accepted.