

Securing High-Speed Networks: From Programmable Data Planes to Trustworthy AI-Enabled Defense

Sergio Elizalde, Elie Kfoury, Jorge Crichigno

Molinaroli College of Engineering and Computing, University of South Carolina, USA

e-mails: elizalds@email.sc.edu, ekfoury@email.sc.edu, jcrichigno@cec.sc.edu

Abstract—Widespread adoption of encryption has pushed monitoring and security functions toward machine-learning (ML) models that infer behavior from metadata and context rather than payload content. This shift occurs as network speeds continue to increase, leaving less time budget to make a detection. When deployed on hosts, these functions compete with applications for compute, memory, and I/O resources. Offloading them to Smart Network Interface Cards (SmartNICs) reduces host involvement but can shift the bottleneck to a resource-constrained data plane. This paper evaluates this trade-off through a common Data Plane Development Kit (DPDK) and Arm NEON architecture deployed for Internet of Things traffic classification, industrial control system anomaly detection, and encrypted Domain Name System threat detection. Across these studies, compact models provide microsecond-scale inference and high packet-processing rates, while feature cost, per-flow state, and observation-window length determine the resulting performance and detection quality. An adversarial evaluation further exposes vulnerabilities in encrypted-traffic classification and the limitations of feature-space feasibility. These findings characterize the capabilities and limits of SmartNIC-based ML and motivate controlled placement and collaborative inference across network execution tiers.

Index Terms—SmartNIC, programmable data plane, network security, machine learning, adversarial machine learning, network management.

I. INTRODUCTION

Next-generation networks increasingly rely on artificial intelligence (AI) to operate and defend themselves. At the same time, encryption has removed the payload visibility these defenses once assumed: protocols such as Transport Layer Security (TLS) and DNS over HTTPS (DoH) conceal the content conventional inspection relies on. Security functions must therefore infer behavior from metadata and flow history. Machine-learning (ML) models suit this task, but they need observation time, per-flow state, and inference resources. Those needs collide with link rates: at 100 Gbps and beyond, the budget per packet falls to tens of nanoseconds.

On the host, applications compete for compute, memory, and I/O, and packet copies and kernel traversal add overhead [1]. Offloading the detector into the network removes that competition but raises a placement question: where along the path should inference run, and at what cost?

Prior work has explored both ends of this continuum. At the switch tier, Flowrest showed that compact Random Forest (RF) models run in the match-action pipeline at line rate, up to 39% more accurately than earlier in-switch designs [2], though fixed stages and limited tables cap state and model size. At the Central Processing Unit (CPU) tier, P4-DPDK

lifts those caps on general-purpose cores at a higher cost per packet [3]. SmartNICs sit between these extremes: their cores and single-instruction, multiple-data (SIMD) instructions run stateful feature extraction and compact inference near the wire, though cores, caches, and shared accelerators remain bounded [4]. That bound tightens for richer models: DomURLs-BERT [5] and netFound [6] need tokenization and far more computation.

Offloading also extends the security boundary: feature state, model parameters, queues, and updates become security-critical resources whose exhaustion or compromise can invalidate a design that looks sound on average. The National Institute of Standards and Technology (NIST) characterizes such threats by adversary goals, capabilities, and knowledge [7], so near-the-wire ML must be evaluated for robustness, not just performance.

Cost and robustness frame the two questions this paper addresses. **RQ1—Cost of near-the-wire inference:** How do feature extraction, per-flow state, observation-window length, and model size affect detection quality and online performance on a SmartNIC? A common DPDK/Arm NEON pipeline answers this for Internet of Things traffic classification, industrial control system anomaly detection, and encrypted DNS threat detection. **RQ2—Robustness of near-the-wire inference:** How vulnerable is an encrypted-traffic detector to adversarial evasion under feature-feasibility constraints? This paper studies white-box attacks and their transferability against a surrogate model.

The paper is organized as follows. Section II describes the common offline-training and online-inference pipeline. Section III presents the three SmartNIC-based case studies and the adversarial robustness evaluation. Together, they establish the cost and robustness limits of near-the-wire ML on a single tier. Section IV then introduces the next question: how should inference be placed across a P4-programmable switch, a SmartNIC, and a host/GPU?

II. SYSTEM OVERVIEW

The three deployed detectors in Section III share the same two-stage pipeline (Fig. 1).

A. Offline Training and Adversarial Evaluation

Public packet or sensor traces are converted into flow- or sensor-level features over a fixed 8–16-sample observation window. Compact Multilayer Perceptrons (MLPs) with one or two hidden layers are trained on these features, translated into fixed-size weight/bias/scaler parameters, and analyzed with

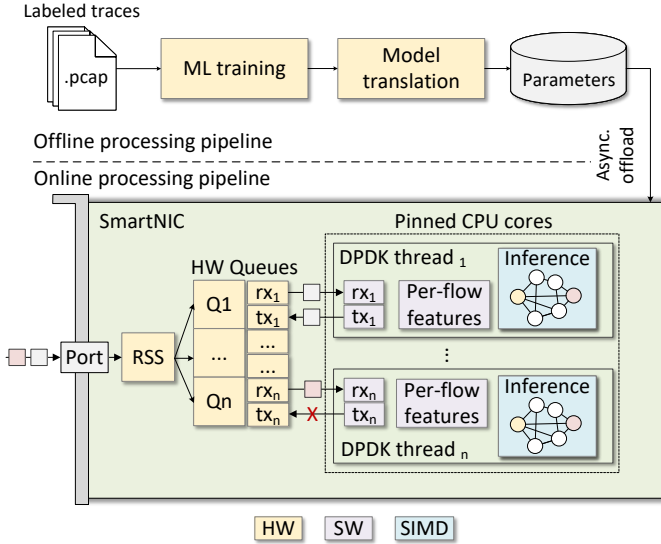


Fig. 1. Common pipeline: an offline stage trains a multilayer perceptron (MLP) on labeled traces and translates it into deployable parameters, deployed to an online DPDK data plane where symmetric receive-side scaling (RSS) steers each flow to a hardware queue, a pinned DPDK thread per queue extracts per-flow features and runs SIMD inference, and a malicious flow (queue n) is dropped rather than transmitted.

precision, recall, F1, confusion matrices, and SHapley Additive exPlanations (SHAP) for feature selection. Layer dimensions are kept to multiples of four so exported parameters map onto 128-bit SIMD registers without scalar fallback.

The same offline loop can also evaluate adversarial robustness before deployment. The Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks are crafted against the trained MLP under an attacker-knowledge model (white-box against the MLP, gray-box via transfer to an RF classifier), optionally constrained through a variational autoencoder (VAE) latent space so perturbations respect inter-feature relationships (e.g., packet- and byte-count consistency) [8]. Attack success rate, transferability, and feasibility-constraint violations are recorded alongside accuracy. Only the encrypted-DNS detector is gated on robustness before deployment (Section III-D); extending this gate to every model is part of the framework (Section IV).

B. Online Data-Plane Inference

The exported parameters are deployed to a DPDK application on NVIDIA BlueField SmartNICs (Fig. 1, bottom). At the SmartNIC's port, symmetric Receive-Side Scaling (RSS) steers every packet of a bidirectional flow to the same hardware queue, and each queue is served by a pinned DPDK thread, giving every core an independent, lock-free path to inference. Each thread keeps flow (or, for the Industrial Control System (ICS) study, per-sensor) state in a hash table, updating size, timing, and directionality statistics until the N -packet window is reached, when the feature vector is normalized and passed to the MLP engine.

Inference is accelerated with Arm NEON 128-bit SIMD: the fused multiply-add `vfmacc_f32()` implements dense-layer accumulation, and `vmaxq_f32()` implements the rectified linear unit (ReLU) activation, updating four output neurons per

instruction. A thread whose flow is classified as malicious drops its transmit path instead of forwarding (Fig. 1, queue n), giving the data plane a detection and enforcement role; otherwise it forwards the flow and marks it finalized so later packets bypass re-classification. Because queues, threads, and per-flow state are independent, throughput scales roughly linearly with cores; we did not isolate the limiting resource. This workflow is common to the three deployed detectors; the feature set, observation window, and MLP size differ among them.

The case studies occupy only the SmartNIC tier of a larger three-tier design spanning a P4-programmable application-specific integrated circuit (ASIC) switch, the SmartNIC, and the host/GPU, extended in Section IV.

III. CASE STUDIES IN SMARTNIC-BASED SECURITY

We apply the pipeline of Section II to three deployed detection problems, spanning high-rate flow classification, mixed physical/cyber anomaly detection, and metadata-rich encrypted-traffic inspection, and extend the encrypted-traffic detector with an offline robustness study. Each workload stresses a different cost in RQ1: IoT stresses packet rate with a minimal feature set, ICS stresses observation-window length, since physical faults develop slowly, and encrypted DNS stresses per-flow state and feature count. Together they bound what one SmartNIC tier sustains.

A. High-Rate IoT Traffic Classification

The Internet of Things (IoT) study trains an 8-feature MLP (packet-size and inter-arrival statistics, total bytes, SYN/ACK ratio) over the first 8 packets, using CIC-IoT-2023 for training and Bot-IoT/TON-IoT for generalization [9]. A two-hidden-layer (64, 32) MLP reaches a weighted F1 of 98.52% in the four-way multiclass setting and 98.59% with attacks merged into one binary class; reconnaissance is the most difficult class to separate from benign traffic. SHAP ranks total bytes and maximum packet size as most influential for tunneling-style command-and-control traffic. Classifying on the first 8 packets lets the detector act before a flow completes, trading away the context a longer window would provide.

Online, the same model runs entirely on the SmartNIC, without host CPU involvement. NEON vectorization yields up to $4.36\times$ speedup over scalar C, and moving from BlueField-2 to BlueField-3 adds a further $1.5\text{--}4.3\times$ latency reduction across model sizes. Throughput scales with RSS-steered cores, reaching 37.5 million packets per second (Mpps) on BlueField-3 with 8 cores, a $2.0\text{--}2.6\times$ gain over BlueField-2 at the same core count (Fig. 2). Detection quality is unaffected by the compact window, while model size and hardware generation determine sustainable throughput.

B. Anomaly Detection in Industrial Control Systems

The ICS study targets Modbus/TCP networks with two paths on the same BlueField-3 pipeline: a per-sensor path using cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) statistics for drift and abrupt shifts, and a per-flow path extracting packet-size, timing, and TCP-flag statistics

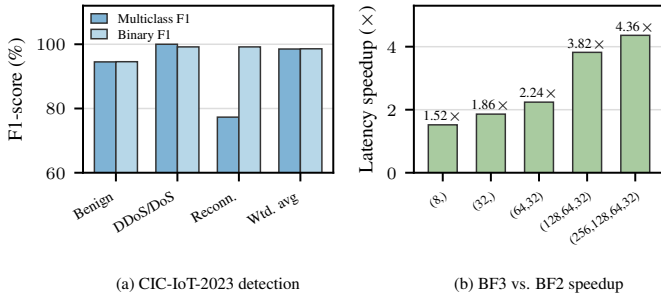


Fig. 2. IoT case study [9]. (a) Detection quality on CIC-IoT-2023. (b) Per-inference latency speedup of BlueField-3 over BlueField-2 across MLP sizes.

TABLE I
ICS ONLINE RESULTS [10]: CUSUM/EWMA IMPROVES FAULT DETECTION ON A 16-PACKET WINDOW.

Class	Statistical only (%)			+CUSUM/EWMA (%)		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Normal	93.13	68.20	78.74	93.20	70.04	79.98
Fault	77.90	95.64	85.89	78.90	95.70	86.47
Wtd. avg	84.91	83.04	82.60	85.49	83.85	83.48

for network-layer attacks (reconnaissance, denial of service, man-in-the-middle, replay) [10], both feeding a compact MLP. Offline, with a 60-second window, the fault MLP separates normal from faulty readings with over 99% accuracy, and the cyberattack MLP reaches over 99% true-negative and 96% true-positive rates.

Online, the constrained 16-packet window reduces fault-detection weighted F1 to 83.48%, below the offline 60-second result, since slow drifts are harder to capture in a short window; CUSUM/EWMA descriptors on that same window recover part of the gap (Table I). Cyberattack detection is unaffected by the shorter window, sustaining 99.985% true-negative and 96.046% true-positive rates at sub-3 μ s inference (Fig. 3). This reflects a trade-off: short windows suit strong per-packet signatures but need drift-aware features for slower physical faults.

C. Encrypted DNS Threat Detection

This study targets DoH traffic and detects domain generation algorithm (DGA) activity and DNS tunneling hidden by encryption [11]. From 29 flow statistics over the first 16 packets, SHAP retains the top 16 to train a four-class MLP (benign DoH, benign non-DoH, DGA, tunneling). Offline, the model exceeds 98% accuracy across all classes (Fig. 4(a)); most confusion is between benign DoH and tunneling, which share volumetric profiles.

Online on BlueField-3, recall against tunneling and DGA traffic stays above 98% across all three evaluated datasets and resolvers (Fig. 4(b)), evidence the compact feature set generalizes across DGA/tunneling tools. NEON reduces inference latency by 1.89–3.65 $\times$, keeping per-flow inference under 5 μ s and sustaining 16 Mpps, lower than IoT’s 37.5 Mpps because DoH flows carry more metadata and longer average durations.

D. Adversarial Robustness Analysis

This study evaluates the encrypted-DNS detector’s robustness using the adversarial procedure of Section II-A [8]. Malicious DGA/tunneling samples are perturbed with FGSM/PGD, under

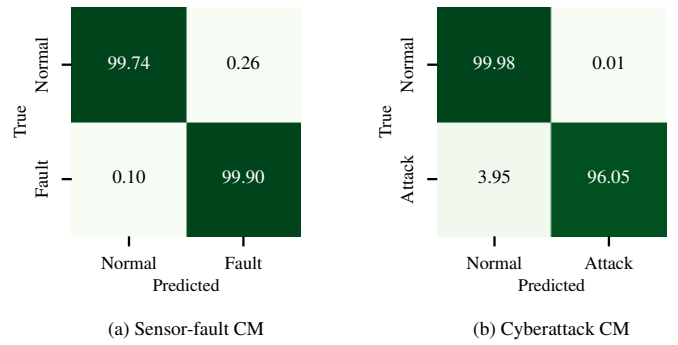


Fig. 3. Online confusion matrices [10]: (a) sensor-fault and (b) Modbus/TCP cyberattack classification, both with the 16-packet online window.

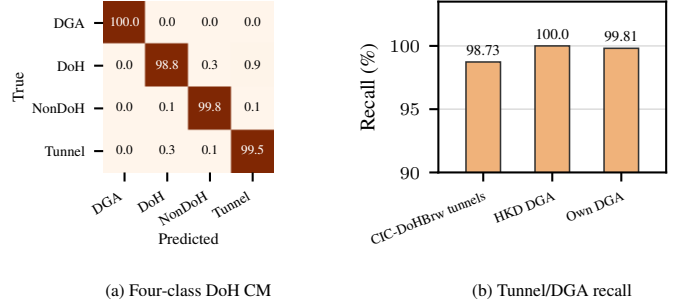


Fig. 4. Encrypted-DNS case study [11]. (a) Offline four-class confusion matrix. (b) Online recall against tunneling/DGA traffic from three datasets.

an L_∞ budget of $\epsilon = 0.3$ on normalized features, to force misclassification as benign, subject to a feasibility model requiring respect for packet-count, byte-count, and ordering constraints from the fixed 16-packet window.

PGD is stronger than FGSM in the white-box setting, reaching 99.9% targeted success against benign non-DoH and 92.3% against benign DoH (Fig. 5(a)). Attacks transfer to a gray-box RF classifier with up to 80% success, so robustness cannot be assumed from model diversity alone. VAE-guided perturbations reduce the strongest attack’s success rate and cut average constraint violations from 5.04 to 3.58 for PGD (Fig. 5(b)); no configuration eliminates violations entirely, so feature-space feasibility is necessary but not sufficient for packet-level realizability.

E. Cross-Study Synthesis

Across the three deployed detectors, inference latency alone does not determine whether offloading is beneficial. IoT and encrypted-DNS both use sub-microsecond inference yet sustain 37.5 and 16 Mpps, respectively; the gap is consistent with DoH’s heavier per-flow state and longer flow duration. Likewise, ICS’s 16-packet window sustains cyberattack detection but is less sensitive to physical faults than its 60-second offline counterpart. Moreover, DPDK performance is inherently constrained by available CPU cores, cache capacity, and memory bandwidth, so the same pipeline can reach different operating points across hardware tiers. With Section III-D’s adversarial results, these are RQ1’s system costs: parsing cost, state size, window length, and robustness, measured on one tier and awaiting cross-tier evaluation.

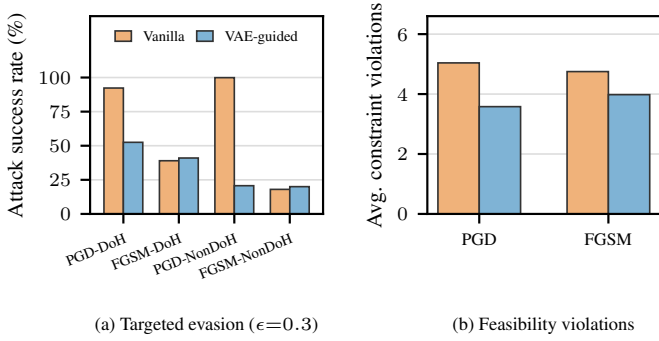


Fig. 5. Adversarial robustness [8]: (a) targeted white-box evasion by attack/target class; (b) average feasibility-constraint violations. VAE guidance reduces but does not eliminate infeasible perturbations.

IV. FUTURE WORK

The case studies answer RQ1’s cost question on a single, pre-chosen tier, not its placement question. Placement spans two timescales: a slow control-plane decision fixes which tier hosts each stage, while a fast, per-flow decision governs whether a flow escalates, trading cross-tier latency and host/GPU load for accuracy and robustness. Fig. 6 shows the proposed extension: packet processing on P4/ASIC, compact inference on the SmartNIC, and heavy compute (training, large-model inference, forensic analysis) on the host/GPU, governed by a management plane, built *incrementally* by integrating the P4 and SmartNIC stages first, then the host/GPU tier and its escalation path.

The mechanism that binds the tiers is *learning to defer*. The compact SmartNIC model must emit not only a label but a signal of when its own decision is unreliable, so only those flows escalate to a larger model. We plan to train the two jointly, so the compact model learns which inputs it can settle locally and which to hand off, treating the deferral rate as a tunable operating point rather than a fixed confidence threshold; that signal must itself be cheap to compute at line rate.

Escalation is not free: each deferred flow costs a cross-tier transfer, queueing, and occupancy of a GPU shared with training and forensics, so a large model earns its keep only if its accuracy gain outweighs that cost. Deferring too often forfeits the latency and host-load advantage that motivated offloading; deferring too rarely forfeits the larger model’s accuracy. We plan to measure the marginal cost of an escalated flow against the accuracy it buys and set the operating point from that curve, which should also expose when a larger model is not worth invoking at all.

Offloading helps only if it remains trustworthy under attack. We plan to broaden RQ2 beyond Section III-D’s evasion study into adversarial training, out-of-distribution detection, and resource-exhaustion attacks on the escalation path.

V. CONCLUSION

This paper evaluated IoT, ICS, and encrypted-DNS detectors on a common DPDK/SIMD SmartNIC pipeline and analyzed the encrypted-DNS model’s adversarial robustness. The results show that feature cost, observation-window length, available compute, and robustness shape the practical value of near-the-wire ML beyond packet rate and inference latency alone.

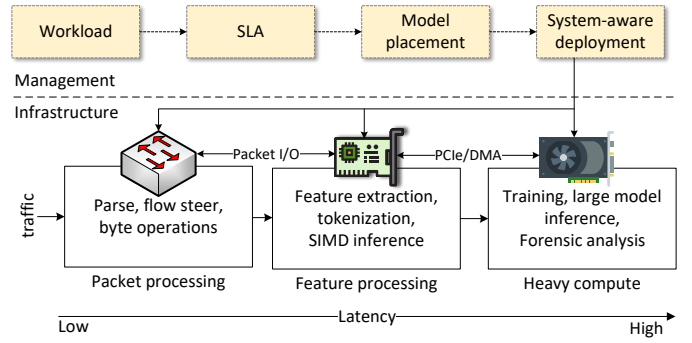


Fig. 6. Target three-tier architecture: a management plane maps workloads to a model-placement decision that is deployed and monitored across packet processing (P4/ASIC), feature processing (SmartNIC), and host/GPU tiers.

These experiments establish the SmartNIC tier’s capabilities and limits, providing the evidence needed for the next stage: assigning each network tier a clear role in AI-enabled defense and evaluating the resulting system as a whole.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) under awards 2346726, 2417823 and 2403360.

REFERENCES

- [1] S. Elizalde, A. Mazloun, S. Choueiri, E. Kfoury, and J. Crichigno, “Performance evaluation of stateless firewalling: Host-based, SmartNIC, and P4 switch,” in *NOMS 2025-2025 IEEE Network Operations and Management Symposium*. IEEE, 2025, pp. 1–5.
- [2] A. T.-J. Akem, M. Gucciardo, and M. Fiore, “Flowrest: Practical flow-level inference in programmable switches with random forests,” in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.
- [3] A. G. S. P. Narasimha, A. Mazloun, S. Choueiri, S. Elizalde, E. Kfoury, and J. Crichigno, “Real-time encrypted traffic classification with p4-dpdk,” in *ICC 2026-IEEE International Conference on Communications*. IEEE, 2026, pp. 1–6.
- [4] S. Elizalde, A. AlSabe, A. Mazloun, S. Choueiri, E. Kfoury, J. Gomez, and J. Crichigno, “A survey on security applications with SmartNICs: Taxonomy, implementations, challenges, and future trends,” *Journal of Network and Computer Applications*, p. 104257, 2025.
- [5] A. El Mahdaoui, S. Lamsiyah, M. Janati Idrissi, H. Alami, Z. Yartaoui, and I. Berrada, “DomURLs_BERT: Pre-trained BERT-based model for malicious domains and URLs detection and classification,” *Journal of Network and Systems Management*, vol. 34, no. 2, p. 36, 2026.
- [6] S. Guthula, R. Beltiukov, N. Battula, W. Guo, A. Gupta, and I. Monga, “netFound: Foundation model for network security,” *arXiv preprint arXiv:2310.17025*, 2023.
- [7] A. Vassilev, A. Oprea, A. Fordyce, and H. Andersen, “Adversarial machine learning: A taxonomy and terminology of attacks and mitigations,” 2024.
- [8] S. Elizalde, A. AlSabe, A. Mazloun, J. Galán-Jiménez, E. Kfoury, and J. Crichigno, “Evaluating white-box adversarial attacks on encrypted DGA and DNS tunneling detection,” in *IEEE/IFIP Network Operations and Management Symposium (NOMS) Workshops*, 2026.
- [9] S. Elizalde, A. AlSabe, A. Mazloun, A. Gspn, E. Kfoury, and J. Crichigno, “Accelerated anomaly detection on IoT traffic using SmartNICs,” in *Proc. of IEEE GlobeCom*, 2025.
- [10] S. Elizalde, S. Choueiri, A. Mazloun, J. Gomez, E. Kfoury, and J. Crichigno, “Accelerating anomaly detection in industrial control systems using SmartNICs and DPDK,” in *2026 IEEE 23rd Consumer Communications & Networking Conference (CCNC)*. IEEE, 2026, pp. 1–6.
- [11] S. Elizalde, A. AlSabe, S. Choueiri, J. Galán-Jiménez, E. Kfoury, and J. Crichigno, “Real-time detection of encrypted DGA and DNS tunneling using DPDK on SmartNICs,” in *2026 IEEE International Black Seas Conference on Communications and Networking (BlackSeaCom)*. IEEE, 2026, pp. 1–6.