Reinforcement Learning-Based Dynamic Multi-Layer Authentication for Resource-Constrained IoMT Systems Under Advanced Persistent Threats

Zainab Issa Ahmad
DAVID Laboratory, Paris-Saclay University, UVSQ, France
Dept. of Mathematics and Computer Science, Univ. of Ngaoundere, Cameroon
znamadina@gmail.com

Yacine Benallouche
Orange Research
University Sorbonne Paris Nord
Paris, France
yacine.benallouche@orange.com

Abdelhak Mourad Gueroui
DAVID Laboratory, Paris-Saclay University, UVSQ
Versailles, France
mourad.gueroui@uvsq.fr

Ado Adamou Abba Ari
DAVID Laboratory, Paris-Saclay University, UVSQ, France
LaRI Laboratory, University of Maroua, Cameroon
ado-adamou.abba-ari@uvsq.fr

***Abstract*—Internet of Medical Things (IoMT) devices operate under strict energy and memory budgets while Advanced Persistent Threats (APTs) compromise them through non-stationary, multi-stage kill chains. Existing authentication schemes either ignore resource constraints, assume a static threat, or reduce authentication to a binary accept/reject decision. We propose DDQN-IoMT, a dynamic multi-layer authentication framework that casts policy optimization as a Constrained Markov Decision Process (CMDP) solved online by a Double Deep Q-Network. Three IoMT-specific adaptations are introduced: resource-aware action masking enforcing hard budgets at every step, kill-chain-aware isolation via a phase-dependent dynamic threshold, and potential-based reward shaping guided by clinical severity. On a heterogeneous population of ten devices over 200 test episodes, DDQN-IoMT attains 81.5% authentication accuracy, a 0.0% false positive rate, and 100% budget satisfaction, uniting adaptive multi-level authentication, reinforcement learning, hard resource constraints, and kill-chain awareness in a single framework.**

***Index Terms*—IoMT; authentication; reinforcement learning; Double DQN; advanced persistent threats; resource constraints; Markov decision process**

I. Introduction

The Internet of Medical Things (IoMT) interconnects sensors, actuators, and gateways that monitor patients and transmit clinical data. These devices are severely resource-constrained yet increasingly targeted by Advanced Persistent Threats (APTs)-stealthy, multi-stage attacks that advance through a kill chain from reconnaissance to physical impact [1], [2], each stage demanding a different defensive posture.

Authentication in this setting faces a threefold challenge. First, authentication consumes energy and memory that constrained devices cannot exhaust, so a budget-agnostic policy is unsafe in practice [3], [4]. Second, the threat is non-stationary: the appropriate authentication strength depends on how far an APT campaign has progressed, making static policies suboptimal. Third, false positives—needlessly disrupting a legitimate medical device—can directly impair patient care and must be avoided. Existing work addresses these only partially: classical authentication fixes a scheme regardless of budget or threat progression [5], [6]; RL-based security [7]–[9] targets intrusion detection rather than graduated authentication and treats resource cost as a soft reward term; kill-chain-aware defenses [10]–[12] model APT stages but do not couple them to a resource-constrained authentication action space. To our knowledge, no prior framework jointly combines all four capabilities.

We propose **DDQN-IoMT** to close this gap. Our contributions are: (i) we formulate authentication policy optimization as a CMDP solved online by a Double Deep Q-Network (DDQN) requiring no prior knowledge of threat dynamics; (ii) we introduce three IoMT-specific adaptations—resource-aware action masking guaranteeing hard budgets by construction, kill-chain-aware forced isolation via a dynamic threshold, and clinically-guided potential-based reward shaping; (iii) we design a two-phase pipeline coupling offline calibration with online DDQN learning; and (iv) we evaluate on ten heterogeneous devices, achieving 81.5% accuracy, 0.0% false positives, and 100% budget satisfaction.

II. Related Work

Authentication for IoMT. Lightweight schemes focus on cryptographic efficiency: AI-assisted device authentication [5] and two-phase dual mechanisms combining ECDH and AES-GCM [6] report strong accuracy, but do not model energy/memory budgets, assume a stationary threat, and collapse the decision into binary accept/reject. Benchmarks of

979-8-3195-0167-7/26/$31.00 ©2026 IEEE

TABLE I
COMPARISON WITH RELATED WORK

Method	Auth.	RL	Resource	Kill-chain
Kelaniki [5]	✓	–	–	–
Asif [6]	✓	–	–	–
Jamshidi [7]	–	✓	–	–
D3O-IIoT [8]	–	✓	–	–
HCLR-IDS [9]	–	✓	–	–
RAPTOR [10]	–	–	–	✓
DeepStage [11]	–	✓	–	✓
DDQN-IoMT (ours)	✓	✓	✓	✓

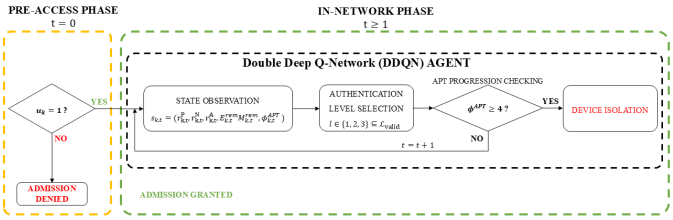


Fig. 1. DDQN-IoMT architecture. A device is admitted only if its initial risk is below τ_0; the DDQN agent then observes the six-dimensional state, selects a feasible authentication level, and enforces isolation once the estimated kill-chain phase reaches lateral movement.

lightweight primitives on constrained platforms [3], [4] quantify the overhead motivating our first-class treatment of resource cost.

RL-based IoT/IoMT security. Deep RL has been applied to IoT intrusion detection [7], dynamic deception for industrial IoT (0.3% FP) [8], and IoMT IDS [9]; edge approaches combine Q-learning with blockchain [13], and energy-aware RL targets offloading/routing [14], [15]. These target detection, not graduated authentication, and treat resources softly rather than as hard guarantees.

Kill-chain-aware defense. RAPTOR [10], DeepStage [11], and DeepXplain [12] model APT progression and identify lateral movement as the critical inflection point [16], but address detection or generic response—not authentication—and never prune the action space by per-device budgets. Table I positions DDQN-IoMT as the only framework combining all four capabilities.

III. System Model and Problem Formulation

We consider devices $\mathcal{D} = \{1, \dots, N\}$ over horizon $t \in \{0, \dots, T\}$ ($t = 0$ pre-access, $t \geq 1$ in-network). At each instant an authentication level is chosen from $\mathcal{L} = \{1, 2, 3\}$: normal operation ($\ell$=1), step-up verification (ℓ=2), isolation/quarantine (ℓ=3). Each ℓ has energy cost e_ℓ, memory cost m_ℓ, and normalized security gain s_ℓ with $s_1 < s_2 < s_3$ [17]; costs are calibrated from constrained-platform benchmarks [3], [4].

The state of device k is $s_{k,t} = (r^P_{k,t}, r^N_{k,t}, r^A_{k,t}, E^{\text{rem}}_{k,t}, M^{\text{rem}}_{k,t}, \varphi^{\text{APT}}_{k,t})$: the perception-, network-, and application-layer risks, remaining energy/memory, and estimated kill-chain phase $\varphi^{\text{APT}}_{k,t} \in \{0, \dots, 6\}$. The aggregated cross-layer risk is

$$r_{k,t} = \alpha\, r^P_{k,t} + \beta\, r^N_{k,t} + \delta\, r^A_{k,t}, \quad \alpha + \beta + \delta = 1, \tag{1}$$

with weights reflecting deployment context [18], [19]. The seven-phase chain (Normal, Reconnaissance, Initial Intrusion, Persistence, Lateral Movement, Exfiltration, Impact) adapts the Cyber Kill Chain [1] and MITRE ATT&CK [2] to IoMT, separating Exfiltration from Impact to distinguish data theft from physical sabotage [10], [16].

With $x^{(\ell)}_{k,t} \in \{0, 1\}$ indicating level ℓ and $u_k \in \{0, 1\}$ admission, the objective trades operational cost against security gain:

$$\min Z = \sum_k \sum_{t=1}^{T} \sum_\ell \big(\lambda_E e_\ell + \lambda_M m_\ell - \lambda_S s_\ell\big) x^{(\ell)}_{k,t}, \tag{2}$$

subject to: admission (C1) $u_k{=}1 \Leftrightarrow r_{k,0} \leq \tau_0$; one level per step (C2) $\sum_\ell x^{(\ell)}_{k,t}{=}u_k$; energy/memory budgets (C3)–(C4) $\sum_{t,\ell} e_\ell x^{(\ell)}_{k,t} \leq E^{\max}_k u_k$ and $\sum_{t,\ell} m_\ell x^{(\ell)}_{k,t} \leq M^{\max}_k u_k$; and forced isolation (C5) $r_{k,t} \geq \tau_{k,t} \Rightarrow x^{(3)}_{k,t}{=}u_k$. This binary program is NP-hard and its risk scores and phase are revealed only sequentially [7], [20], so we recast it as a CMDP solved online by an RL agent.

IV. The DDQN-IoMT Framework

As shown in Fig. 1, a device is admitted at $t{=}0$ iff $r_{k,0} \leq \tau_0$. In-network ($t \geq 1$) the DDQN agent observes $s_{k,t}$, selects a level from the resource-feasible set $\mathcal{L}_{\text{valid}}(s_{k,t})$, escalating as risk grows and isolating whenever $\varphi^{\text{APT}}_{k,t} \geq 4$ or $r_{k,t} \geq \tau_{k,t}$—adapting continuously without exceeding budgets.

CMDP. We model $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, C, b \rangle$ [21] with per-step costs $C_E(s, \ell){=}e_\ell$, $C_M(s, \ell){=}m_\ell$ and budgets $b{=}\{E^{\max}, M^{\max}\}$. Since costs depend only on the action, realized cost equals expected cost, so hard budgets (C3)–(C4) hold exactly rather than in expectation. The transition P is unknown, motivating the model-free RL approach.

Reward. Aligned with (2),

$$R(s_{k,t}, \ell) = \lambda_S s_\ell - \lambda_E e_\ell - \lambda_M m_\ell - \rho\, \mathbf{1}\{\text{FP}(s_{k,t}, \ell)\}, \tag{3}$$

where a false positive is a step-up/isolation ($\ell \geq 2$) on a below-threshold device in the normal phase; $\rho > 0$ penalizes disrupting legitimate devices. Maximizing discounted reward is equivalent to minimizing Z.

Three adaptations. *(1) Resource-aware action masking:* the feasible set $\mathcal{L}_{\text{valid}}(s_{k,t}) = \{\ell \mid e_\ell \leq E^{\text{rem}}_{k,t},\ m_\ell \leq M^{\text{rem}}_{k,t}\}$ is recomputed online, and infeasible actions are masked to $-\infty$ before the arg-max in both selection and the DDQN target, guaranteeing hard budgets without Lagrangian relaxation. *(2) Kill-chain-aware isolation:* the threshold decreases with the phase,

$$\tau_{k,t} = \tau_0\big(1 - \varphi^{\text{APT}}_{k,t}/6\big), \tag{4}$$

Algorithm 1 DDQN-IoMT Online Training Loop

Require: Budgets $E_k^{\max}, M_k^{\max}$; τ_0; weights; potential Φ; φ_{iso}; penalty ρ
Ensure: Trained online network Q_θ
Init Q_θ, target $Q_{\bar{\theta}} \leftarrow Q_\theta$, buffer $\mathcal{B} \leftarrow \emptyset$, $\varepsilon \leftarrow \varepsilon_{\max}$
for ep $= 1$ **to** N_{ep} **do**
  Admit device k iff $r_{k,0} \le \tau_0$ // (C1)
  for $t = 1$ **to** T **do**
    $r_{k,t} \leftarrow \alpha r^P_{k,t} + \beta r^N_{k,t} + \delta r^A_{k,t}$; $\varphi^{\text{APT}}_{k,t} \leftarrow$ EstPhase$(r_{k,t})$
    $\tau_{k,t} \leftarrow \tau_0(1 - \varphi^{\text{APT}}_{k,t}/6)$; compute $\mathcal{L}_{\text{valid}}$ // (C3–C4)
    if $r_{k,t} \ge \tau_{k,t}$ **or** $\varphi^{\text{APT}}_{k,t} \ge \varphi_{\text{iso}}$ **then**
      $\ell_{k,t} \leftarrow 3$ // forced isolation (C5)
    else if $\mathcal{U}(0,1) < \varepsilon$ **then**
      $\ell_{k,t} \sim \text{Unif}(\mathcal{L}_{\text{valid}})$
    else
      $\ell_{k,t} \leftarrow \arg\max_{\ell \in \mathcal{L}_{\text{valid}}} Q_\theta(s_{k,t}, \ell)$
    end if
    Apply $\ell_{k,t}$; observe $s_{k,t+1}$; update budgets
    $\tilde{R}_{k,t} \leftarrow R + \Delta R^{\text{APT}}$; store transition in $\mathcal{B}$
    if $|\mathcal{B}| \ge B$ **then**
      Sample mini-batch; compute Y^{DDQN} (5); gradient step
    end if
    Every C_{upd} steps: $Q_{\bar{\theta}} \leftarrow Q_\theta$
  end for
  $\varepsilon \leftarrow \max(\varepsilon_{\min}, \varepsilon \cdot \varepsilon_{\text{decay}})$
end for
return Q_θ

TABLE II
HYPERPARAMETER CONFIGURATION

Param.	Value	Param.	Value
Buffer $\lvert\mathcal{B}\rvert$	50,000	Learning rate η	5×10^{-4}
Batch B	64	Warm-up steps	500
γ	0.99	Episodes N_{ep}	1,000
$\varepsilon_{\max}/\varepsilon_{\min}$	1.0/0.01	Isolation φ_{iso}	4
$\varepsilon_{\text{decay}}$	0.995	FP penalty ρ	5
Target update C_{upd}	200		

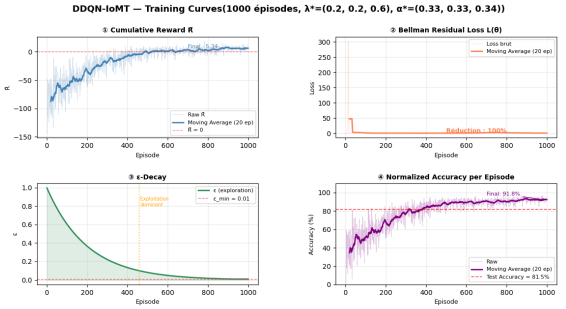


Fig. 2. Training over 1,000 episodes: cumulative reward, Bellman loss, ε-decay, and normalized accuracy (test accuracy 81.5%).

raising sensitivity as the attack advances; isolation is forced when $r_{k,t} \ge \tau_{k,t}$ or $\varphi^{\text{APT}}_{k,t} \ge \varphi_{\text{iso}}$ (C5). *(3) Potential-based reward shaping:* following Ng *et al.* [22], $\Delta R^{\text{APT}} = \gamma\Phi(\varphi^{\text{APT}}_{k,t+1}) - \Phi(\varphi^{\text{APT}}_{k,t})$ with severity potential Φ normalized to $[-1, 0]$ preserves the optimal policy while encouraging early containment; the shaped reward $\tilde{R} = R + \Delta R^{\text{APT}}$ enters the Bellman update.

Double DQN. To avoid Q-value overestimation [23], selection and evaluation are decoupled:

$$Y_t^{\text{DDQN}} = \tilde{R} + \gamma\, Q_{\bar{\theta}}\Big(s_{k,t+1}, \arg\max_{\ell' \in \mathcal{L}_{\text{valid}}} Q_\theta(s_{k,t+1}, \ell')\Big), \quad (5)$$

minimizing $\mathcal{L}(\theta) = \mathbb{E}[(Y_t^{\text{DDQN}} - Q_\theta(s_{k,t}, \ell))^2]$. Since $\mathcal{L}_{\text{valid}}$ appears identically in training and inference, hard constraints hold throughout [24].

Training. An offline phase calibrates τ_0 by ROC/Youden analysis and pre-trains a Random-Forest APT phase classifier (with deterministic fallback) that estimates $\varphi^{\text{APT}}_{k,t}$; the online phase (Algorithm 1) learns the policy with the three adaptations embedded in the loop.

V. Experimental Evaluation

Setup. Experiments run CPU-only (Intel Core i5-8265U, 16 GB), matching edge deployment, in Python/PyTorch with a Gymnasium IoMT environment and scikit-learn classifier. Risk scores are generated via phase-parameterized Beta distributions. We model N=10 heterogeneous devices (glucose/cardiac/BP/temperature sensors, insulin pump, ECG, oximeter, pressure sensor, MQTT gateway, hub) with energy budgets 200–450 mJ and memory 300–500 KB; one device is drawn per episode. The threshold τ_0=0.4437 is set by Youden index [25]; reward weights $(\lambda_S, \lambda_E, \lambda_M)$=$(0.6, 0.2, 0.2)$ and (α, β, δ)=$(0.33, 0.33, 0.34)$ by grid search [26]. Key hyperparameters, following standard DQN practice [27], are in Table II.

Training. Cumulative reward rises monotonically and crosses zero near episode 400; the Bellman loss collapses within the first 50 episodes; exploitation dominates around episode 460. The accuracy moving average reaches 91.8% while out-of-sample test accuracy stabilizes at 81.5% (Fig. 2), the gap reflecting environment stochasticity and device heterogeneity.

Performance. Over 200 unseen-device test episodes (Table III), the agent reaches 81.5% accuracy, 0.0% false positives, and 100% budget satisfaction; (C3)–(C4) hold in every episode, validating action masking. Step-up (31.5%) and isolation (15.4%) rates confirm threat-proportional escalation. Lowering η from the Adam default 10^{-3} [28] to 5×10^{-4} raises accuracy 72.0%→81.5% and triples the step-up correct rate (10.0%→31.5%), while the two hard guarantees (0.0% FP, 100% budget) hold identically—both by construction, independent of η, which governs only escalation quality.

Discussion. Patient safety is preserved (0.0% FP: no legitimate device is disrupted); resource budgets are strictly enforced (100% via masking, absent from all baselines); and early escalation contains APTs before lateral movement. The 81.5% accuracy is below binary IDS frameworks (89–97%), but the comparison is not direct: DDQN-IoMT decides among three levels under hard constraints and kill-chain awareness while keeping FP strictly below the 0.3% of the closest method [8]. The main limitation is the synthetic environment.

TABLE III
PERFORMANCE OVER 200 TEST EPISODES (AND EFFECT OF η)

Metric	η=10^{-3}	η=5×10^{-4}
Overall accuracy	72.0%	**81.5%**
Step-up correct rate	10.0%	**31.5%**
Isolation rate	–	15.4%
False positive rate	0.0%	0.0%
Budget respect (C3/C4)	100.0%	100.0%
Mean cumulative reward	8,104.7	**8,276.0**

VI. CONCLUSION AND FUTURE WORK

We presented DDQN-IoMT, a reinforcement learning-based framework for dynamic multi-layer authentication of resource-constrained IoMT devices under APT conditions. By formulating authentication as a CMDP and introducing resource-aware action masking, kill-chain-aware isolation, and clinically-guided reward shaping, the framework guarantees hard budget satisfaction by construction while adapting authentication strength to threat progression. Experiments on a heterogeneous devices demonstrate 81.5 % accuracy, a 0.0 % false positive rate, and 100 % budget satisfaction, uniquely combining adaptive multi-level authentication, RL, hard resource constraints, and kill-chain awareness.

The main limitation is the use of synthetic simulation data. Future work will replace it with real IoMT network traffic from CICIoMT2024 [29] and APT kill-chain labels from DAPT 2020 [30], introduce a clinical criticality indicator into the admission constraint to keep life-critical devices accessible, and extend the per-device CMDP to a multi-agent setting capturing inter-device APT propagation across the network graph.

ACKNOWLEDGMENT

This work was supported by the French Government Scholarship (BGF), within the cotutelle between Paris-Saclay University (DAVID Laboratory) and the University of Ngaoundere (Dept. of Mathematics and Computer Science).

REFERENCES

[1] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains," Lockheed Martin Corp., Tech. Rep., 2011.
[2] MITRE Corporation, "MITRE ATT&CK: Adversarial tactics, techniques and common knowledge," 2024. [Online]. Available: https://attack.mitre.org
[3] H. Y. Naser *et al.*, "A comparison of lightweight cryptographic protocols for energy-efficient and sustainable IoMT authentication," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 4, pp. 25746–25756, 2025.
[4] M. A. Saare *et al.*, "Evaluating elliptic curve cryptography in constrained environments: A Raspberry Pi-based approach," *Int. J. Innov. Res. Sci. Stud.*, vol. 8, no. 2, 2025.
[5] S. Miri Kelaniki and N. Komninos, "A study on IoT device authentication using artificial intelligence," *Sensors*, vol. 25, no. 18, art. 5809, 2025.
[6] M. Asif *et al.*, "Intelligent two-phase dual authentication framework for Internet of Medical Things," *Sci. Rep.*, vol. 15, art. 1760, 2025.
[7] S. Jamshidi *et al.*, "Application of deep reinforcement learning for intrusion detection in Internet of Things," *Internet of Things*, 2025.
[8] U. Wushishi, A. Hussain *et al.*, "D3O-IIoT: Deep reinforcement learning-driven dynamic deception orchestration for industrial IoT security," *Sci. Rep.*, vol. 15, 2025.
[9] J. A. Shaikh, C. Wang *et al.*, "A deep reinforcement learning-based robust intrusion detection system for securing IoMT healthcare networks," *Front. Med.*, 2025.
[10] A. Kumar and V. L. Thing, "RAPTOR: Advanced persistent threat detection in industrial IoT via attack stage correlation," arXiv:2301.11524, 2023.
[11] T. V. Phan, T. G. Nguyen, and T. Bauschert, "DeepStage: Learning autonomous defense policies against multi-stage APT campaigns," arXiv:2603.16969, 2026.
[12] T. V. Phan and T. Bauschert, "DeepXplain: XAI-guided autonomous defense against multi-stage APT campaigns," arXiv:2603.21296, 2026.
[13] M. Bayat *et al.*, "Enhancing secure IoT data sharing through dynamic Q-learning and blockchain at the edge," *Sci. Rep.*, vol. 15, art. 39153, 2025.
[14] Y. Wang *et al.*, "Robust and energy-efficient RPL optimization algorithm with scalable deep reinforcement learning for IIoT," *Comput. Netw.*, vol. 255, 2024.
[15] X. Jiao, H. Ou *et al.*, "Deep reinforcement learning for time-energy tradeoff online offloading in MEC-enabled industrial IoT," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, pp. 3465–3479, 2023.
[16] D. D. Dau, S. Lee, and H. Kim, "A comprehensive comparison study of ML models for multistage APT detection," *J. Supercomput.*, vol. 80, no. 10, pp. 14143–14179, 2024.
[17] NIST, "SP 800-63B: Digital identity guidelines," Tech. Rep., 2024.
[18] R. Mustafa *et al.*, "A cross-layer secure and energy-efficient framework for the Internet of Things: A comprehensive survey," *Sensors*, 2024.
[19] Z. Almulla, H. Almajed, and M. M. H. Rahman, "A layered security perspective on Internet of Medical Things," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 5, 2025.
[20] F. Wang, Q. He, and S. Li, "Solving combinatorial optimization problems with deep neural network: A survey," *Tsinghua Sci. Technol.*, vol. 29, no. 5, pp. 1266–1282, 2024.
[21] E. Altman, *Constrained Markov Decision Processes*. Routledge, 2021.
[22] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations," in *Proc. ICML*, 1999, pp. 278–287.
[23] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI*, vol. 30, 2016.
[24] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, 1992.
[25] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
[26] M. Aledhari *et al.*, "Optimizing intrusion detection in IoMT networks through interpretable and cost-aware machine learning," *Mathematics*, vol. 13, no. 10, 2025.
[27] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[29] S. Dadkhah *et al.*, "CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT," *Internet of Things*, vol. 28, 2024.
[30] S. Myneni *et al.*, "DAPT 2020—constructing a benchmark dataset for advanced persistent threats," in *Proc. MLHat*, 2020.