Deep Reinforcement Learning for Handover Optimization in 5G Networks

Damianos Diasakos
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
Email: up1084632@upnet.gr

Vasileios Kokkinos
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
Email: kokkinos@cti.gr

Christos Bouras
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
Email: bouras@upatras.gr

Apostolos Gkamas
Department of Chemistry
University of Ioannina
Ioannina, Greece
Email: gkamas@uoi.gr

Philippos Pouyioutas
Computer Science Department
University of Nicosia
Nicosia, Cyprus
Email: pouyioutas.p@unic.ac.cy

***Abstract*— Ultra-dense 5G networks require advanced traffic steering to maintain performance and balance load amid growing user and base station (gNB) densities. Traditional heuristics such as nearest-base-station and Signal-to-Interference-plus-Noise Ratio (SINR)-based selection provide simple solutions but struggle to adapt to dynamic user mobility, diverse traffic, and fluctuating radio conditions at the mobility-control level, often leading to inefficient handovers and degraded network quality. We propose a deep reinforcement learning (DRL) framework to dynamically tune a global handover hysteresis margin that governs handover triggering decisions, optimizing handover success, reducing failures, and enhancing throughput and fairness. Implemented in Python using Stable Baselines3 and NumPy, our custom simulation environment models key mobility-related 5G dynamics at a high level, including user mobility, pathloss-based signal degradation, and interference. We evaluate DRL agents—Deep Q-Network (DQN) and Proximal Policy Optimization (PPO)—against heuristic and hysteresis-based baselines. Results show that DRL-based hysteresis optimization provides strong and robust performance under the considered ultra-dense mobility conditions in handover success rate, average SINR, throughput, and fairness, with PPO demonstrating the most consistent behavior across configurations. This work offers a reproducible simulation framework for further research into adaptive mobility management.**

Keywords—Deep Reinforcement Learning, Traffic Steering, 5G Networks, Load Balancing, Handover Optimization, Self-Organizing Networks

I. INTRODUCTION

The deployment of 5G networks marks a major advancement in wireless communications, offering unprecedented data rates, ultra-low latency, and massive device connectivity. A key enabler of this evolution is the ultra-dense network (UDN) paradigm, where a high density of base stations (gNBs)—ranging from macrocells to small cells—is deployed to meet growing traffic demands and support emerging applications such as autonomous vehicles, augmented reality, and the Internet of Things (IoT).

However, network densification significantly increases the complexity of mobility management. Closely spaced gNBs lead to overlapping coverage regions, increased inter-cell interference, and frequent handover events. In such environments, effective mobility control—particularly the tuning of handover triggering behavior via a network-wide global hysteresis margin—is essential for balancing load, maximizing throughput, and ensuring seamless connectivity.

Traditional heuristic approaches, such as nearest-base-station or Signal-to-Interference-plus-Noise Ratio (SINR)-based selection, are computationally efficient but struggle to cope with dynamic user mobility and rapidly varying radio conditions. As a result, these methods often suffer from excessive handovers, signaling overhead, and degraded user experience, especially in ultra-dense deployments. In this work, the hysteresis margin is explicitly modeled as a single global control parameter, applied uniformly across all users, reflecting practical operator-level mobility configuration rather than per-user decision-making. Unlike DRL approaches that directly decide handover execution for each UE or dynamically assign target gNBs on a per-user basis, our framework operates at the parameter-control level by optimizing the shared hysteresis threshold governing standardized Event A3 triggering. This design significantly reduces the action-space complexity and improves scalability, since the agent selects one network-wide control action instead of managing hundreds of simultaneous user-specific decisions. While per-user DRL schemes may achieve finer local optimization, they typically suffer from high-dimensional state spaces, increased training instability, and substantial signaling complexity. Our global-parameter approach prioritizes operational simplicity, reproducibility, and compatibility with practical 3GPP mobility management procedures while still enabling adaptive traffic steering and load balancing.

In response to these challenges, Deep Reinforcement Learning (DRL) has gained increasing attention as a promising tool for traffic steering and mobility management in 5G networks. By enabling agents to learn control policies through interaction with the network environment, DRL can capture complex dependencies between mobility, interference, and network load that are difficult to model analytically. Prior studies have demonstrated the potential of DRL in this context. Habib et al. [1] employed DRL for traffic steering in multi-Radio Access Technology (multi-RAT) 5G systems, achieving throughput and delay improvements over heuristic and classical Q-learning approaches. Within the Open Radio Access Network (O-RAN) framework, Kavehmadavani et al. [2] leveraged DRL to optimize key performance indicators (KPIs) across representative 5G service scenarios. Similarly, Gu et al. [3] applied Proximal Policy Optimization (PPO) to adaptive handover control, outperforming standardized 5G New Radio (NR) procedures in terms of link stability and data rates. In addition, several studies have investigated learning-based handover and

mobility control in dense and ultra-dense cellular networks, showing that reinforcement learning can effectively adapt handover parameters under dynamic channel and mobility conditions, while also highlighting the importance of scalable hysteresis-aware traffic steering, adaptive mobility optimization, and self-organizing mobility management in dense wireless environments [4],[5],[6].

Despite these advances, existing studies typically focus on a single learning algorithm or adopt heterogeneous assumptions regarding mobility models, reward formulations, and evaluation metrics, making direct comparison across DRL paradigms difficult. Moreover, the lack of unified and reproducible simulation setups limits systematic assessment of stability–performance trade-offs relative to conventional heuristics.

To address these gaps, this paper proposes a DRL-based traffic steering framework evaluated within a custom simulation environment that models key mobility-related 5G dynamics at an abstract yet controlled level, including mobile users, interference-coupled gNBs, and time-varying channel conditions. We investigate two representative DRL paradigms—Deep Q-Network (DQN) and PPO—and benchmark them against strong heuristic and adaptive non-learning baselines commonly used in cellular mobility management. All methods are evaluated under identical environment dynamics, reward formulation, and training conditions. Performance is assessed using multiple KPIs, including handover success rate, handover failures, average SINR, total throughput, and fairness. The results show that PPO achieves consistent performance gains and stable convergence, while DQN exhibits higher sensitivity to non-stationarity and reward configuration. The proposed framework enables reproducible evaluation and supports further research on learning-based mobility management in ultra-dense 5G networks.

Unlike prior works that primarily optimize direct per-user handover execution or evaluate a single DRL paradigm, this work focuses on operator-level global hysteresis optimization and provides a unified comparison of DQN, PPO, heuristic baselines, and standardized hysteresis-based baselines under identical simulation assumptions.

The remainder of this paper is organized as follows. Section II summarizes the DRL algorithms considered in this study. Section III describes the model architectures, baseline methods, and reward design. Section IV presents the simulation environment and testbed configuration. Section V reports the experimental results and discussion, and Section VI concludes the paper and outlines future research directions.

II. Deep Reinforcement Learning Algorithms

DRL enables agents to learn optimal policies in complex environments through trial-and-error interactions, using deep neural networks to approximate value functions or policies [7]. This section outlines two DRL algorithms—DQN and PPO—used for traffic steering in 5G networks, highlighting their mechanisms, strengths, and limitations.

A. Deep Q-Network

DQN, introduced by Mnih et al. [7], combines Q-learning with deep neural networks to approximate the action-value function $Q(s,a;\theta)$, where s is the state, a is the action, and θ denotes network parameters. DQN minimizes the temporal-difference (TD) error using the loss:

$$L(\theta) = E[(r + \gamma \max_{\acute{a}} Q(\acute{s}, \acute{a}; \theta^{-}) - Q(s,a;\theta))^2] \quad (1)$$

Here, r is the immediate reward, γ is the discount factor balancing immediate and future rewards, $\acute{s}$ is the next state, and θ^{-} represents the parameters of a target network. To further improve stability and sample efficiency, DQN employs experience replay, where transitions are stored in a replay buffer and randomly sampled during training to reduce temporal correlations [8]. Exploration is typically achieved via an ϵ-greedy policy, balancing exploration and exploitation. Extensions such as Double DQN mitigate overestimation bias by decoupling action selection and evaluation [9]. In our implementation, both DQN and PPO select a single global hysteresis action at each decision step, which is uniformly applied across all users, while individual user mobility and radio conditions determine whether a handover is triggered.

DQN is well-suited for problems with discrete action spaces and benefits from off-policy learning, allowing efficient reuse of collected experience. However, its performance can be sensitive to hyperparameter selection and environmental non-stationarity, particularly in dynamic wireless environments [10].

B. Proximal Policy Optimization

PPO, proposed by Schulman et al. [11], is a policy-gradient method that optimizes a stochastic policy $\pi(a|s;\theta)$, within an actor–critic framework. The actor represents the policy, while the critic estimates the state-value function $V(s;\varphi)$. PPO improves training stability by constraining policy updates using a clipped surrogate objective:

$$L^{CLIP}(\theta) = \mathrm{E}\Big[\min(r_t(\theta)\,\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)\Big] \quad (2)$$

Where $r_t(\theta)$ is the probability ratio between the new and old policies:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \quad (3)$$

and ϵ is a clipping parameter (typically 0.2). The advantage estimate $\hat{A}_t$ is computed using Generalized Advantage Estimation (GAE):

$$\hat{A}_t = \sum_{i=0}^{T-t-1} (\gamma\lambda)^i \delta_{t+i} \quad (4)$$

with the temporal difference (TD) error defined as:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (5)$$

Here, r_t denotes the reward at time step t, γ is the discount factor controlling the importance of future rewards, λ is the smoothing parameter used in GAE, and $V(s;\varphi)$ represents the state-value function parameterized by φ. PPO typically performs multiple optimization epochs over collected trajectories and includes entropy regularization to encourage exploration. Its clipped objective prevents excessively large policy updates, yielding stable learning behavior across a wide range of environments [11]. While PPO is more computationally intensive than DQN due to its on-policy nature and multi-epoch updates, its actor–critic structure and clipped objective provide robust convergence and effective

handling of stochastic and non-stationary dynamics [12]. These characteristics make PPO a suitable candidate for stable handover control in dynamic ultra-dense 5G networks.

DQN and PPO are selected due to their complementary characteristics and widespread use in networking and control applications. DQN represents value-based, off-policy learning with discrete control actions and low inference complexity, while PPO represents policy-gradient learning with improved training stability through constrained updates. Evaluating both algorithms under identical environment dynamics and reward formulation enables a systematic comparison of performance, stability, and computational overhead. More complex DRL methods such as DDPG or Soft Actor-Critic (SAC) are left for future work, as they introduce additional architectural and hyperparameter complexity beyond the scope of this study.

III. DRL Model Architecture and Implementation

This section summarizes the DRL architectures, baseline methods, reward formulation, and hysteresis-based control policies used in the proposed handover optimization framework. The implementation uses Python, Stable Baselines3, PyTorch, and NumPy.

A. Deep Q-Network

The DQN agent uses a multi-layer perceptron to estimate $Q(s, a; \theta)$, where a is one of five discrete global hysteresis actions. The input layer has 600 units, corresponding to 100 users with 6 features per UE: normalized 3D position (x,y,z), normalized speed magnitude, normalized serving-gNB assignment, and one reserved placeholder feature for future extensions. No explicit target-gNB load or queue-length information is included in the current state representation. A UE-level representation was selected to preserve fine-grained mobility dynamics and local SINR variation during handover triggering, while BS-centric aggregation is reserved for future scalability-oriented extensions. The network uses two hidden layers with 256 ReLU neurons each and an output layer with 5 units. At each decision step, the selected hysteresis margin is applied uniformly to all users, while individual handovers are triggered locally according to user-specific SINR conditions. Exploration follows an ϵ-greedy policy with ϵ decaying from 1.0 to 0.05.

B. Proximal Policy Optimization

PPO is implemented using Stable Baselines3 with an actor–critic architecture. Both networks use the same 600-dimensional state vector and two hidden layers of 64 ReLU neurons. The actor outputs a categorical distribution over the five global hysteresis actions, while the critic estimates the state value. Compared with DQN, PPO has higher training complexity due to its on-policy updates, but its clipped objective improves stability under stochastic mobility and interference conditions.

C. Baseline Methods

We compare against two heuristic baselines: Nearest-BS and SINR-based association. Nearest-BS assigns each UE to the geographically closest gNB, providing a simple distance-driven strategy but ignoring interference and load balancing. SINR-based association selects the gNB with the highest instantaneous SINR, improving link quality but potentially concentrating users around strong cells. These baselines represent common distance-driven and signal-driven mobility strategies and provide non-learning benchmarks for evaluating DRL-based control.

D. Reward Function Design

The reward function guides the DRL agent toward maximizing handover success and network quality. At each time step, the reward r_t is computed as:

$$r_t = a * SuccessRate - b * HandoverFailures + \gamma * SINR + \delta * Throughput + \eta * Fairness \quad (6)$$

with normalized metric values per episode and weights set as: a = 1.0, b = 0.625, γ = 0.3, δ = 0.4, η=0.2. The handover success rate is defined as the ratio of successful mobility events to total decision steps, while handover failures occur when the received SINR falls below the predefined signal threshold required for reliable connectivity. Throughput is estimated using the Shannon-inspired expression $\log_2(1 + SINR)$, and fairness is computed using Jain's fairness index over the user distribution across all gNBs.

To examine the sensitivity of DRL behavior to reward prioritization, multiple reward-weight configurations (S0–S5) were evaluated. S0 represents the default balanced setting, while S1 increases the penalty on handover failures (b = 0.625), giving stronger emphasis to mobility robustness and service continuity. Additional configurations vary the relative importance of failures, throughput, and SINR to study policy stability under different operator objectives. Among these settings, S1 produced the most consistent and practically meaningful behavior, with both PPO and DQN converging to stable low-hysteresis policies and improved reliability. For this reason, S1 is used for the final comparative evaluation reported in Table IV.

Although explicit target-gNB load indicators (e.g., number of connected users or queue lengths) are not included in the observation space, load balancing is learned indirectly through the fairness component of the reward. Since handover decisions modify user-to-gNB assignments, the selected global hysteresis margin directly influences the resulting load distribution. This allows the agent to improve traffic steering and balance load without requiring explicit queue-state modeling, while keeping the observation space compact and scalable.

For computational stability, each metric is normalized to lie within the range [0,1] during training. In future work, we plan to extend the reward formulation with explicit gNB-side load indicators, queue-length awareness, and QoS-class-based weighting to better support heterogeneous traffic demands.

E. Hysteresis-Based Handover Policies

In cellular mobility management, hysteresis is a key control parameter that regulates handover triggering by requiring a minimum signal quality advantage of a target gNB over the serving gNB. By tuning this margin, networks can balance responsiveness against stability, mitigating ping-pong effects while avoiding delayed handovers. In practical 3GPP systems, hysteresis margins correspond directly to Event A3 offset and hysteresis parameters.

In this study, the hysteresis margin constitutes the sole control variable optimized by both DRL agents. At each decision step, the agent selects a single global hysteresis level, which is uniformly applied across all users. Individual

handover decisions are then made locally by UEs based on their SINR measurements and the selected global margin. This design mirrors operational 5G mobility control, where shared threshold parameters govern handover behavior rather than explicit per-user commands.

To benchmark DRL performance, several hysteresis-based non-learning baselines are evaluated. The Fixed-hyst(k) schemes apply static hysteresis margins from a discrete set $\{H_0,\ldots,H_4\}$, which also defines the discrete action space of both DQN and PPO. Lower indices correspond to aggressive handover behavior, while higher indices impose increasingly conservative decisions. An Adaptive-Hyst baseline dynamically adjusts the hysteresis margin based on recent handover outcomes, introducing limited rule-based adaptiveness without learning. Finally, the A3-Smoothed-Hyst baseline approximates standardized 3GPP Event A3 triggering by applying temporal smoothing to signal quality differences, reducing sensitivity to short-term fluctuations.

By optimizing the same hysteresis parameter used by these baselines, the proposed DRL agents operate at the parameter-control level, enabling fair and interpretable comparison with standardized and heuristic mobility strategies under identical decision semantics.

IV. Simulation Environment and Testbed Configuration

A Gymnasium-based simulation environment, HandoverEnv, was developed to evaluate DRL-based mobility control in a dense 5G setting. The environment contains five fixed gNBs with overlapping coverage and 100 mobile UEs in a bounded 3D area. It abstracts the full 3GPP protocol stack while preserving key mobility-control dynamics, including user movement, pathloss-based signal degradation, interference, and handover triggering. The main simulation parameters are shown in Table I.

TABLE I. General Experiment Parameters

Parameter	***Value***
# Base stations	5
# Users	100
Episode length	500 steps
User Speed Range	[-1, 1]
User Position Range	0-100 units
Training Time steps	500,000

Users are initialized randomly and move continuously with uniformly sampled speeds. Boundary wrapping ensures persistent mobility and recurring handover opportunities. Each gNB has a 50-unit reception range, and received power follows a distance-dependent $1/d^2$ pathloss model with log-normal fading. Overlapping gNB coverage creates interference-coupled handover decisions typical of ultra-dense deployments.

The SINR is calculated as:

$$SINR = \frac{P_{signal}}{P_{interference} + N_0} \quad (7)$$

where P_{signal} is the received power from the serving gNB under normalized equal transmit power, distance-dependent pathloss, and log-normal fading. $P_{interference}$ is the sum of received powers from neighboring gNBs within reception range, and $N_0 = 0.01$ is the normalized thermal noise floor.

Fig. 1 illustrates the topology with five gNBs, 50-unit coverage regions, and 100 UEs.

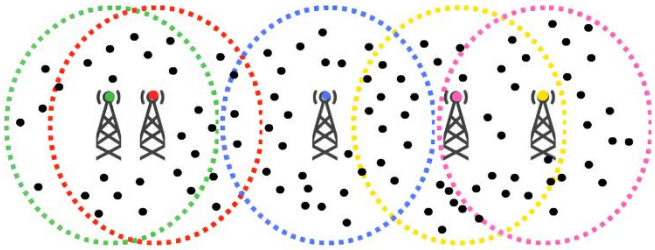

Fig. 1. Network Topology Environment

We run each DRL agent (DQN and PPO) for up to 500,000 timesteps using Stable-Baselines3 implementations. PPO leverages GAE and entropy regularization, while DQN employs experience replay, target networks, and ϵ-greedy exploration with linear decay. The hyperparameters used for DQN and PPO are listed separately in Table II and Table III, respectively. Performance metrics include handover success rate, number of handover failures, average SINR, total throughput, and fairness.

TABLE II. DQN Parameters

Parameter	***Value***
Learning Rate	2.5e-4
Replay Buffer Size	100,000
Learning Starts	5,000
Batch Size	64
Discount Factor	0.99
Target Update Interval	500
Exploration Fraction	0.2
Final Epsilon	0.05

TABLE III. PPO Parameters

Parameter	***Value***
# Steps	2048
Batch Size	256
# Epochs	10
Learning Rate	3e-4
Discount Factor	0.99
GAE Lambda	0.95
Clip Range	0.2
Entropy Coefficient	0.01

V. Experimental Results and Discussion

In this section, we evaluate the effectiveness of the proposed DRL approaches—DQN and PPO—for handover optimization and compare them against heuristic and hysteresis-based baselines, namely including Fixed-hyst policies, Adaptive-Hyst, A3-Smoothed-Hyst, Nearest-BS, and SINR-based assignment. The evaluation is conducted using five KPIs: handover success rate, number of handover failures, average SINR, total throughput and fairness. These metrics capture both user-level mobility reliability and overall network efficiency and are consistent with commonly adopted 3GPP mobility performance indicators.

The handover success rate and handover failures directly reflect mobility robustness. High success rates ensure session continuity during mobility, which is critical for delay-sensitive services such as Voice over IP (VoIP), live streaming, and Ultra-Reliable Low-Latency Communication

(URLLC) applications. In contrast, frequent handover failures may lead to dropped connections and service degradation. SINR quantifies radio link quality and interference handling, while throughput reflects aggregate system capacity and effective load distribution across gNBs.

Table IV summarizes the final performance of all evaluated methods at 500k training steps for the DRL agents and after fixed runs for the heuristic and hysteresis-based baselines. While intermediate checkpoints were monitored during training, we report final results to ensure fair comparison and avoid misinterpretation caused by transient learning instabilities, particularly for DQN.

TABLE IV. PERFORMANCE AT 500K TRAINING STEPS AND FIXED-RUN EVALUATION (BASELINES)

Method	*Success Rate*	*Handover Failures*	*Avg. SINR*	*Total Throughput*	*Fairness*
PPO(500k steps)	0.7509	124.55	0.1841	98.33	0.8918
DQN(500k steps)	0.7509	124.55	0.1841	98.33	0.8918
Fixed-hyst(0)	0.7423	128.85	0.1852	97.58	0.8973
Fixed-hyst(1)	0.3171	341.45	0.1383	65.59	0.9643
Fixed-hyst(2)	0.2415	379.25	0.1178	52.71	0.9663
Fixed-hyst(3)	0.2149	392.55	0.1050	45.27	0.9679
Fixed-hyst(4)	0.2029	398.55	0.0966	40.74	0.9679
Adaptive-Hyst	0.5496	225.20	0.1571	82.01	0.8989
A3-Smoothed-Hyst	0.6190	190.50	0.1674	85.77	0.9061
Nearest - BS	0.7284	135.80	0.1895	92.87	0.8897
SINR-based	0.7325	133.75	0.2018	96.57	0.8913

To complement the numerical results, Fig. 2 illustrates the learning curves of PPO and DQN. It is important to note that the reward function is a weighted composite objective jointly capturing handover success rate, failures, SINR, throughput, fairness, handover cost, and ping-pong penalties. As such, reward trajectories reflect learning stability and policy improvement rather than direct convergence of any single KPI reported in Table IV.

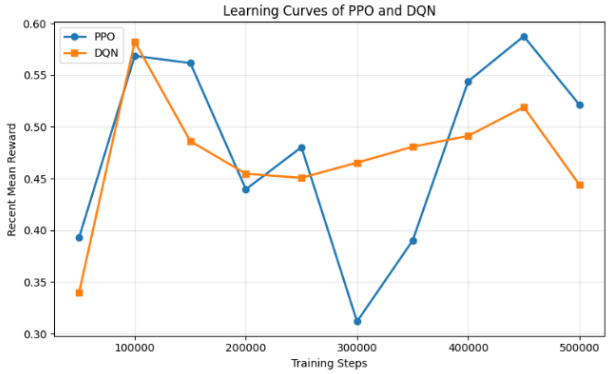

Fig. 2. Learning Curves DQN & PPO

The environment includes continuous user mobility, stochastic fading, interference coupling, and dynamically changing user-to-gNB assignments, which naturally introduce reward fluctuations even after policy stabilization. Therefore, perfectly monotonic convergence is not expected. As shown in Fig. 2, DQN exhibits a smoother and more gradual reward progression, while PPO shows larger fluctuations during intermediate training stages due to the stochastic mobility environment and interference dynamics. Despite this variability, PPO ultimately reaches higher final training reward and stabilizes toward the same final policy.

In addition to Nearest-BS and SINR-based heuristics, several hysteresis-based baselines are evaluated to reflect standardized mobility control mechanisms commonly used in cellular networks. In these baselines, a handover is triggered only when the signal quality difference between a candidate gNB and the serving gNB exceeds a predefined hysteresis margin. This mechanism aims to suppress unnecessary handovers caused by short-term signal fluctuations, at the cost of reduced responsiveness.

The Fixed-hyst(k) baselines correspond to static hysteresis policies with discrete margin levels, where smaller values (e.g., Fixed-hyst(0)) allow aggressive handover behavior, while larger values impose increasingly conservative handover decisions. As shown in Table IV, increasing the hysteresis margin monotonically reduces the handover rate but leads to degraded success rate, SINR, and throughput due to delayed handovers and reduced adaptability.

The Adaptive-Hyst baseline dynamically adjusts the hysteresis level based on recent handover outcomes, aiming to balance stability and responsiveness without explicit learning. The A3-Smoothed-Hyst baseline approximates the 3GPP Event A3 triggering mechanism by applying temporal smoothing to the signal difference before initiating handovers, thereby reducing ping-pong effects caused by fast fading.

The strong performance of Fixed-hyst(0) highlights an important characteristic of the evaluated scenario. Since the simulation assumes a static gNB topology, homogeneous traffic demand, and time-invariant channel statistics within each episode, an aggressively tuned static hysteresis rule can approximate near-optimal behavior. In such quasi-stationary conditions, Fixed-hyst(0) effectively emulates strongest-signal selection with minimal handover suppression. However, unlike PPO, this static policy lacks adaptability and degrades rapidly when mobility patterns, interference conditions, or traffic distributions deviate from those assumed during tuning. PPO, by contrast, learns a policy that remains robust across stochastic transitions and noisy observations, as reflected in its stable reward convergence.

Training required approximately 609 s for PPO and 867 s for DQN. During deployment, inference overhead was negligible because each agent selects only one global hysteresis value per decision step, with sub-millisecond inference times of 0.41 ms for PPO and 0.28 ms for DQN.

A key observation is that under the reliability-oriented S1 configuration, both DQN and PPO converge to the same aggressive low-hysteresis policy, which explains their identical final KPI values in Table IV, including success rate, handover failures, SINR, throughput, and fairness. However, their learning behavior during training differs.

These differences arise from the underlying learning mechanisms of the two DRL methods. DQN relies on off-policy value iteration with bootstrapped targets, which can

produce stable short-term reward behavior but may also encourage convergence toward fixed local policies. In contrast, PPO performs constrained on-policy updates using a clipped surrogate objective, allowing stronger policy adaptation under non-stationary mobility conditions. As a result, although DQN appears smoother during training, PPO demonstrates better robustness across configurations and stronger final policy quality under the reliability-oriented setting.

VI. Conclusion & Future Work

This study evaluated DRL for handover optimization in 5G networks using a custom simulation environment. We compared two DRL algorithms—PPO and DQN—against multiple heuristic and hysteresis-based baselines, including Fixed-hyst policies, Adaptive-Hyst, A3-Smoothed-Hyst, Nearest-BS, and SINR-based association, across key metrics: success rate, handover failures, SINR, throughput, and fairness. While the current simulation employs a distance-based pathloss model to ensure tractable and interpretable learning dynamics, more complex channel effects such as log-normal shadowing and fast fading are not explicitly modeled. Incorporating standardized 3GPP channel models (e.g., UMa/UMi) is a natural extension and will be explored in future work to further validate robustness under realistic radio uncertainty.

Results show that under the reliability-oriented S1 configuration, both PPO and DQN converge to the same aggressive low-hysteresis policy, achieving identical final KPI values across all evaluated metrics, including success rate (0.7509), handover failures (124.55), average SINR (0.1841), total throughput (98.33), and fairness (0.8918). Although their final performance is identical, their learning behavior during training differs. DQN exhibits a smoother and more gradual reward progression, while PPO shows larger fluctuations during intermediate training stages due to the stochastic mobility environment and interference dynamics. Despite this variability, PPO reaches higher final training reward and demonstrates greater robustness across different reward configurations and non-stationary conditions. Although the static Fixed-hyst(0) baseline performs competitively under the quasi-stationary evaluated scenario, both DRL agents provide stronger adaptability under dynamic mobility conditions and maintain stable policy improvement throughout training. Among the heuristic baselines, Nearest-BS and SINR-based strategies provide simple, low-complexity benchmarks; however, the DRL-based approaches generally outperform them in both reliability and overall capacity. Notably, while SINR-based association aligns with strongest-signal selection rules, its performance remains limited by interference coupling and reduced adaptability, further highlighting the value of learning-based mobility control.

Advanced ML-based handover schemes often rely on distinct state definitions, reward formulations, or simulator assumptions, which complicates fair comparison. For this reason, we focus on standardized heuristic and hysteresis-based baselines that reflect widely deployed handover strategies. The comparison highlights the performance gains achievable when transitioning from rule-based policies to learning-based control under identical conditions. The reported success rates reflect a highly challenging ultra-dense mobility scenario with frequent handovers and strong interference. These values are intended for relative comparison across methods, rather than absolute compliance with 3GPP KPIs. All evaluated methods operate under identical conditions, ensuring fair performance comparison.

Future work will focus on extending the environment to include dynamic user populations, heterogeneous gNB types, and energy-aware metrics. In addition, we plan to evaluate the proposed framework under multiple deployment scenarios with varying user densities, different numbers of gNBs, diverse mobility patterns, and heterogeneous traffic conditions in order to better assess generalization under realistic UDN behavior. Incorporating macro–small-cell coexistence, QoS-aware traffic classes, and fairness-aware reward extensions may further improve scalability, robustness, and practical applicability. BS-centric state representations using aggregated cell-level features will also be investigated as a potential approach for improving scalability and accommodating variable user populations in large-scale deployments.

References

[1] M. A. Habib et al., "Traffic Steering for 5G Multi-RAT Deployments using Deep Reinforcement Learning," 2023 IEEE 20th Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2023, pp. 164-169

[2] Fatemeh Kavehmadavani, Van-Dinh Nguyen, Thang X Vu, and Symeon Chatzinotas. 2023. On Deep Reinforcement Learning for Traffic Steering Intelligent ORAN. In IEEE Globecom Workshops (GC Wkshps). 565–570.

[3] J. Voigt, P. J. Gu and P. M. Rost, "A Deep Reinforcement Learning-Based Approach for Adaptive Handover Protocols," 2025 14th International ITG Conference on Systems, Communications and Coding (SCC), Karlsruhe, Germany, 2025, pp. 1-6, doi: 10.1109/IEEECONF62907.2025.10949111.

[4] Tanveer J, Haider A, Ali R, Kim A. “An Overview of Reinforcement Learning Algorithms for Handover Management in 5G Ultra-Dense Small Cell Networks”. Applied Sciences. 2022; 12(1):426. https://doi.org/10.3390/app12010426

[5] Q. Liu, C. F. Kwong, W. Sun, L. Li, and H. Zhao, “Reinforcement learning based adaptive handover in ultra-dense cellular networks with small cells,” Proc. SPIE, vol. 11574, pp. 124–131, Oct. 2020

[6] L. A. Grieco, G. Boggia, G. Piro, Y. Jararweh, and C. Campolo (Eds.), *Ad-Hoc, Mobile, and Wireless Networks: 19th International Conference, AdHoc-Now 2020, Bari, Italy, October 19–21, 2020, Proceedings*, Lecture Notes in Computer Science, vol. 12338, Springer, 2020.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, Feb. 2015,

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: MIT Press, 2018

[9] H. van Hasselt, A. Guez, and D. Silver, “Deep Reinforcement Learning with Double Q-Learning,” in Proc. AAAI Conf. Artificial Intelligence, 2016.

[10] M. Chen, U. Challita, W. Saad, C. Yin and M. Debbah, "Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3039-3071, Fourthquarter 2019, doi: 10.1109/COMST.2019.2926625.

[11] J. Schulman et al., “Proximal policy optimization algorithms,” arXiv preprint arXiv:1707.06347, 2017.

[12] N. D. L. Fuente and D. A. V. Guerra, “A comparative study of deep reinforcement learning models: DQN vs PPO vs A2C,” 2024, *arXiv:2407.1415*