

Reinforcement Learning-Based Handover for LEO Non-Terrestrial Networks

Caio Brasil, Cleverson Nahum, Wallace Martins, Moisés V. Ribeiro, Leonardo Ramalho and Aldebaro Klautau

Abstract—In low earth orbit (LEO) non-terrestrial networks (NTN), the high orbital speed of satellites and their limited visibility time over a given user can lead to frequent handovers, increasing signaling overhead, handover failures, outage probability, and service instability. This makes satellite selection a relevant problem for NTN mobility management, especially in dense mega-constellation scenarios. To address this problem, this paper presents a reinforcement learning-based agent for satellite selection in LEO NTN systems. This work uses CAVIAR with ns-3 NTN channel simulation to generate datasets for handover studies. With that, a deep Q-learning (DQN) agent is trained to select serving satellites while considering spectral efficiency, handover attempts, handover failures, outages, and satellite load. Results show that the DQN agent reduces handover failures without great loss of spectral efficiency performance, indicating the potential of reinforcement learning for NTN handover management.

Keywords—Reinforcement learning, non-terrestrial networks, handover

I. INTRODUCTION

Non-terrestrial networks (NTN) have been considered an important component of future wireless systems due to their ability to extend connectivity beyond the coverage of terrestrial infrastructure. By relying on satellite, aerial, or high-altitude platforms, NTN can support service continuity in remote areas, maritime and aeronautical scenarios, disaster recovery situations, and regions where the deployment of dense terrestrial networks is economically or technically challenging. These characteristics make NTN relevant for the evolution of 5G and beyond systems, where ubiquitous and resilient connectivity is a key requirement [1].

Among the different NTN architectures, low Earth orbit (LEO) satellite networks have received significant attention due to their relatively low launch cost compared to geostationary ones [2]. In particular, the deployment of large-scale satellite constellations, such as Starlink and OneWeb, has shown the practical relevance of mega-constellation systems for broadband access and global coverage. Recent surveys on

satellite constellations report a rapid increase in the number of planned and deployed LEO satellites, reinforcing the importance of studying the technical and economic sustainability of these systems [3]. However, the use of LEO satellites also introduces important mobility-related challenges. Since these satellites move at high orbital speeds and remain visible to a given user equipment (UE) for a limited time interval, the serving satellite may need to be changed frequently. As a consequence, LEO-based NTN systems are expected to experience a higher number of handovers than conventional terrestrial networks.

Frequent handovers can degrade network performance by increasing signaling overhead, service interruption time, and the probability of handover failure. These issues are especially relevant in dense LEO mega-constellations, where multiple satellites may be visible, and the best serving satellite can change rapidly due to orbital motion and propagation conditions. As a result, intelligent handover management has become an active research topic in NTN systems. Recent studies have explored both analyzers to implement handover triggers [4] and reinforcement learning (RL) mechanisms for selecting or anticipating the most suitable target satellite. However, some learning formulations rely on rich observation spaces or structured network-side information [5], do not consider the impacts of spectral efficiency [6], or do not have a robust reward function implementation [7].

In this context, we propose a DQN agent for satellite selection and handover management. For this, we use a framework that integrates satellite mobility generation, user mobility, and network-level simulation outputs, enabling the construction of datasets that describe the interaction between UEs and LEO satellites over time. Using this framework, we study an RL agent that selects the serving satellite based on observations derived from the simulated NTN environment. The objective is not only to evaluate the feasibility of RL-based handover decisions but also to provide a reproducible simulation pipeline for analyzing NTN mobility management under controlled yet realistic LEO constellation conditions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a LEO NTN scenario composed of a set of satellites and terrestrial UEs. Let $\mathcal{S} = \{s_1, s_2, \dots, s_S\}$ denotes the set of S satellites in the constellation, and $\mathcal{U} = \{u_1, u_2, \dots, u_U\}$ denotes the set of U UEs. At each simulation timestamp t , each satellite $s \in \mathcal{S}$ has a position vector $\mathbf{p}_s(t)$, while each UE $u \in \mathcal{U}$ has a position vector $\mathbf{p}_u(t)$. The distance between UE u and satellite s at time t is given by $d_{u,s}(t) = \|\mathbf{p}_s(t) - \mathbf{p}_u(t)\|$. For each UE, a subset of

This study was financed in part by the project SOLVE (CAPES); Ericsson Telecomunicações S.A., Brazil; the Brasil 6G project (01245.020548/2021-07); Project Smart 5G Core And MULtiRan Integration SAMURAI (MCTIC/CGI.br/FAPESP under Grant 2020/05127-2); and partially supported by the National Institute of Science and Technology (INCT) STREAM, funded by the Brazilian National Council for Scientific and Technological Development (CNPq), process no. 409179/2024-8.

Caio Brasil, Cleverson Nahum, Leonardo Ramalho and Aldebaro Klautau are with Federal University of Pará, Belém, PA 66075-110, Brazil, (email: caio.brasil@itec.ufpa.br, {cleverson, leonardolr, aldebaro}@ufpa.br)

Moisés V. Ribeiro are with the Electrical Engineering Department, Federal University of Juiz de Fora, Juiz de Fora, MG 36036-900, Brazil, (email: moises.ribeiro@ufjf.br)

Wallace A. Martins is with the Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France (email: wallace.martins@isae-supaero.fr).

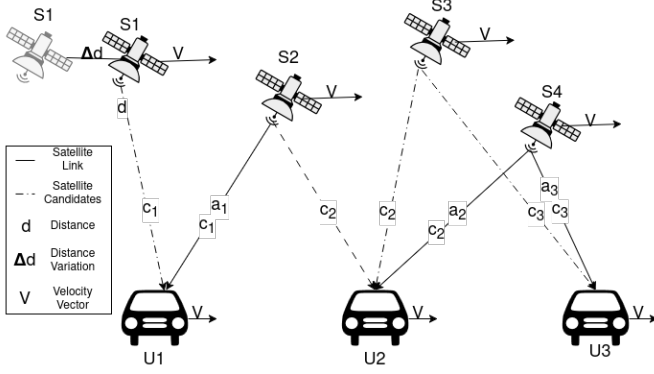


Fig. 1. System Model.

candidate satellites is selected from the full constellation. Let $\mathcal{C}_u(t) \subseteq \mathcal{S}$ denote the candidate set available to UE u at time t , with $|\mathcal{C}_u(t)| = N$ being a maximum number of candidate satellites. The handover decision consists of selecting one serving satellite $a_u(t) \in \mathcal{C}_u(t)$ for each UE at each decision instant.

The wireless channel between a UE and a satellite is characterized by the NTN path loss model, following the assumptions of 3GPP TR 38.811 [8] given by the equation (1). It considers both free-space path loss and additional propagation effects, such as shadow fading (PL_{SF}), outdoor-to-indoor penetration loss (PL_{O2I}), atmospheric absorption (PL_{AA}), rain and cloud attenuation ($\text{PL}_{\text{rain}}, \text{PL}_{\text{cloud}}$), scintillation (PL_{scint}), and fast fading (PL_{fast}) given by the scenario models adopted by 3GPP [8].

$$\begin{aligned} \text{PL}_{u,s}^{\text{NTN}}(d_{u,s}(t), f_c) = & \text{PL}_{\text{FS}} + \text{PL}_{\text{SF}} + \text{PL}_{\text{O2I}} \\ & + \text{PL}_{\text{AA}} + \text{PL}_{\text{rain}} + \text{PL}_{\text{cloud}} \\ & + \text{PL}_{\text{scint}} + \text{PL}_{\text{fast}}, \end{aligned} \quad (1)$$

where, for a UE–satellite link, the free-space path loss component can be written as

$$\text{PL}_{\text{FS}}(d_{u,s}, f_c) = 32.45 + 20 \log_{10}(f_c) + 20 \log_{10}(d_{u,s}), \quad (2)$$

where PL_{FS} is given in dB, f_c is the carrier frequency in MHz, and $d_{u,s}$ is the UE–satellite distance in km.

Given the path loss, the received power associated with the link between UE u and satellite s is computed as

$$P_{u,s}^{\text{rx}}(t) = P_s^{\text{tx}} + G_s^{\text{tx}} + G_u^{\text{rx}} - \text{PL}_{u,s}^{\text{NTN}}(t), \quad (3)$$

where P_s^{tx} is the satellite transmit power in dBm, G_s^{tx} is the satellite antenna gain, and G_u^{rx} is the UE antenna gain. The corresponding SNR is given by

$$\text{SNR}_{u,s}^{\text{dB}}(t) = P_{u,s}^{\text{rx}}(t) - P_{\text{noise}}, \quad (4)$$

where the noise power is

$$P_{\text{noise}} = -174 + 10 \log_{10}(B) + \text{NF}, \quad (5)$$

with B denoting the system bandwidth in Hz and NF the receiver noise figure in dB. The spectral efficiency (SE) of the UE–satellite link is given by

$$\eta_{u,s}(t) = \log_2(1 + \text{SNR}_{u,s}^{\text{lin}}(t)), \quad (6)$$

where $\eta_{u,s}(t)$ is given in bit/s/Hz and $\text{SNR}_{u,s}^{\text{lin}}(t)$ is the linear-scale SNR.

The main objective of the handover decision process is to maintain good link quality over time while avoiding unstable or inefficient associations. In LEO NTN systems, this problem is challenging because the relative motion between UEs and satellites causes rapid variations in distance, elevation angle, visibility, and channel quality. As a result, the satellite providing the best spectral efficiency at a given instant may no longer be the best option in the next time step.

However, frequent handovers are undesirable because they increase signaling overhead, consume control-plane resources, and may introduce service interruptions. In addition, a handover attempt may fail when the selected target satellite is unable to serve the UE. In this work, a handover failure (HOF) occurs when the selected action is invalid, when the target satellite exceeds its maximum serving capacity. An outage occurs when the UE has no serving satellite or when the spectral efficiency of the selected link is lower equal to zero. Therefore, the handover policy must consider not only the instantaneous link quality but also the stability and feasibility of the selected association.

Let $a_u(t)$ denote the satellite selected to serve UE u at time t , and let $a_u(t-1)$ denote its previous serving satellite. The handover indicator $h_u(t)$ is equal to 1 if $a_u(t) \neq a_u(t-1)$. Similarly, let $\phi_u(t)$ and $o_u(t)$ correspond to the handover failure indicator and outage indicator, respectively. $\phi_u(t) = 1$ when there is a HOF given a attempt to attach the UE to a full satellite or when the satellite is not visible at the next simulation step, $o_u(t) = 1$ when there is no satellite for a UE or its SE is equal to 0.

Then, the problem can be formulated as the search for a satellite selection policy that maximizes the accumulated spectral efficiency while minimizing handovers, handover failures, and outages. For a time horizon T , the optimization objective can be written as

$$\begin{aligned} \max_{\pi} \quad & \sum_{t=1}^T \sum_{u \in \mathcal{U}} \left[\eta_{u,a_u(t)}(t) - \lambda_{\text{HO}} h_u(t) \right. \\ & \left. - \lambda_{\text{HOF}} \phi_u(t) - \lambda_{\text{out}} o_u(t) \right], \end{aligned} \quad (7)$$

subject to

$$a_u(t) \in \mathcal{C}_u(t), \quad \forall u \in \mathcal{U}, \forall t, \quad (8)$$

and

$$\sum_{u \in \mathcal{U}} \mathbb{1}\{a_u(t) = s\} \leq C_s, \quad \forall s \in \mathcal{S}, \forall t, \quad (9)$$

where π is the satellite selection policy, λ_{HO} , λ_{HOF} , and λ_{out} are penalty coefficients, and C_s is the maximum number of UEs that satellite s can serve simultaneously. The first constraint ensures that each UE can only select a satellite from its candidate set, while the second constraint limits the number of UEs associated with each satellite.

This formulation captures the trade-off addressed in this work. Maximizing only the instantaneous spectral efficiency may lead to excessive handovers and higher signaling

overhead, while minimizing handovers alone may keep UEs connected to poor or unavailable links. The objective is therefore to learn a policy that performs handovers only when they are beneficial, feasible, and unlikely to result in HOF or outage.

III. REINFORCEMENT LEARNING METHODOLOGY

A. RL-Based Satellite Selection Environment

Based on the system model and optimization objective defined in the previous section, the satellite selection task is mapped into a reinforcement learning problem. The environment is constructed from the dataset generated by the proposed problem formulation and is modeled as a sequential decision process in which the agent selects the serving satellite for one UE at each decision step. Therefore, the agent interacts with an offline environment rather than directly controlling the network simulator during training.

Let k denote the environment decision step. Each step k is associated with a pair (u_k, τ) , where $u_k \in \mathcal{U}$ is the UE being evaluated and τ is the corresponding index of the simulation timestamp. Since each simulation timestamp contains information for multiple UEs, the environment step does not necessarily coincide with the simulation time index. Instead, the relation between times is given by $k = |\mathcal{U}|\tau + u_k$, where the environment iterates over the UEs available at a given simulation timestamp before advancing to the next one.

For each UE u , the environment provides a candidate set

$$\mathcal{C}_{u_k}(k) = \{a_{k,1}, a_{k,2}, \dots, a_{k,N}\}, \quad (10)$$

This set is obtained by analyzing both distance between the UE and the satellites, $d_{u_k, c_{k,i}}(k)$, and the variation of this distance, $\Delta d_{u_k, c_{k,i}}(k)$, using the equation $Score = -w_1 d_{u_k, c_{k,i}}(t) + w_2 \Delta d_{u_k, c_{k,i}}(t)$. In this work, both w_1 and w_2 are equal to 1. After computing the score for each satellite, the N best ones will be selected to create the candidates set.

After that, the observation at step k is represented as a matrix

$$\mathbf{o}_k \in \mathbb{R}^{N \times F}, \quad (11)$$

where each row contains the features associated with one candidate satellite. For candidate satellite $c_{k,i}$, the feature vector is defined as

$$\mathbf{x}_{k,i} \triangleq [d_{u_k, a_{k,i}}(k), \Delta d_{u_k, a_{k,i}}(k), \eta_{u_k, a_{k,i}}(k), b_{u_k, a_{k,i}}(k), \rho_{a_{k,i}}(k)], \quad (12)$$

where $\eta_{u_k, a_{k,i}}(k)$ is the estimated spectral efficiency, $b_{u_k, a_{k,i}}(k)$ is a binary association indicator that equals 1 when the satellite $a_{k,i}$ is serving the UE u_k , and $\rho_{a_{k,i}}(k)$ is the load ratio of the serving satellite, given by $\rho_{a_{k,i}}(k) = \frac{L_{a_{k,i}}(k)}{C}$, where $L_{a_{k,i}}(k)$ is the number of UEs associated with satellite $a_{k,i}$ and C is its maximum capacity.

The action $i \in \{0, \dots, N-1\}$ corresponds to selecting the i -th satellite in the candidate set $\mathcal{C}_{u_k}(k)$. Therefore, the selected satellite is given by $s_k = c_{k, a_k}$. If the selected satellite differs from the current serving satellite, the action is interpreted as a handover attempt.

In this environment, handover execution delay is not explicitly modeled. Thus, if the handover is valid and does not violate the satellite capacity constraint, the new association is assumed to be available in the next decision involving the same UE. If the selected candidate is invalid or the target satellite exceeds its maximum capacity, the handover is considered unsuccessful, and the UE remains associated with its previous serving satellite.

The reward is defined to be consistent with the optimization objective introduced in the problem formulation. It combines the spectral efficiency obtained after the decision with penalties associated with handover attempts, handover failures, and outage. Let τ^+ denote the next simulation timestamp for UE u_k , and let $\tilde{s}_k$ be the satellite effectively serving the UE after applying the action. The immediate reward is

$$r_k = \alpha \eta_{u_k, \tilde{s}_k}(\tau^+) - \lambda_{\text{HO}} h_k - \lambda_{\text{HOF}} \phi_k - \lambda_{\text{out}} o_k, \quad (13)$$

where α is the spectral efficiency scaling factor that was considered equal to 1, ϕ_k is the handover failure indicator, o_k is the outage indicator.

This RL formulation allows the agent to learn a policy that maps the observation of a UE and its candidate satellites into a satellite selection action. Differently from a purely greedy policy based only on instantaneous spectral efficiency, the learned policy is trained to account for future consequences of the association decision, including handover costs, failed handovers, and outage events.

B. Deep Q-learning

The Deep Q-Network (DQN) algorithm extends Q-learning by approximating the action-value function for a given pair of state and action observed with a deep neural network [9], enabling learning in problems with large state spaces and discrete action spaces. At each time step, the agent selects an action, commonly using an ϵ -greedy policy, observes the environment transition, and stores it in an experience replay buffer. During training, mini-batches are sampled from this buffer to reduce correlations between consecutive experiences.

To improve training stability, DQN uses a target network with parameters θ^- , updated separately from the main network with parameters θ . This reduces instability caused by the combination of function approximation, bootstrapping, and off-policy learning. The network is trained by minimizing the mean squared temporal-difference error expressed as

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right], \quad (14)$$

where γ is the discount factor and (s, a, r, s') is a transition sampled from the replay buffer.

IV. EXPERIMENTAL SETUP

The experiments were designed to evaluate the proposed RL-based satellite selection environment under a LEO NTN scenario using the out-loop model of CAVIAR [10], in which external simulators and data-generation modules are coordinated through message exchange. In this approach, the

machine learning model is not directly embedded inside the network simulator. Instead, the simulator is used to generate observations, measurements, and performance indicators that are later consumed by the learning environment.

The simulated scenario considers 20 UEs randomly distributed over an area of 20,000 km². The UE mobility model was configured so that users move from left to right across the simulation area. For the satellite segment, a Starlink-based constellation with 1600 LEO satellites was adopted. At each simulation timestamp, the framework evaluates the satellite positions and selects 5 candidate satellites for each UE.

The communication parameters were selected based on the configuration adopted in [7]. The satellite downlink carrier frequency was set to 30 GHz, with a bandwidth of 250 MHz. The transmit power was defined through a power spectral density of -4 dBW/Hz, and the antenna gain was set to 15.9 dBi. A dataset containing 610,000 samples was generated and temporally divided into 600,000 training samples and 10,000 test samples collected two hours later, ensuring that the DQN was evaluated on previously unseen network states. Table I summarizes the main dataset and scenario parameters.

TABLE I
SCENARIO PARAMETERS.

Parameter	Value
Carrier frequency	30 GHz
Bandwidth	250 MHz
Power spectral density	-4 dBW/Hz
Satellite antenna gain	15.9 dBi
Channel model	NTN-urban
Number of UEs	20
Simulation area	20,000 km ²
UE mobility direction	Left to right
Satellite constellation	Starlink-based
Number of satellites	1600
Candidate satellites per UE	5
Satellite Capability	8 UEs
Total Simulation Time	10 minutes
Handover Penalty	0.1
HOF Penalty	2.0
Outage Penalty	1.0
Training dataset size	600,000 samples
Test dataset size	10,000 samples

The DQN agent was compared with two baseline strategies. The first baseline, referred to as the *Best Spectral Efficiency* (BSE) baseline, always selects the candidate satellite with the highest future spectral efficiency. This strategy maximizes the immediate link quality but does not consider the cost or frequency of handovers. The second baseline, referred to as the *No-Handover* (Holder) baseline, avoids handovers whenever possible. In this case, the UE remains connected to the current serving satellite and only changes its association when there is no feasible alternative. These two baselines represent opposite behaviors: one prioritizes instantaneous link performance, while the other prioritizes connection stability.

The RL agent was trained using the DQN algorithm. The policy was implemented as a multilayer perceptron with two hidden layers of 128 neurons each. The agent was trained for 10,000,000 environment steps. The main hyperparameters used during training are shown in Table II.

Overall, the experiment evaluates whether the DQN agent

TABLE II
DQN TRAINING PARAMETERS.

Parameter	Value
Policy	MLP
Hidden layers	[128, 128]
Learning rate	10 ⁻⁴
Replay buffer size	100,000
Learning starts	2,000 steps
Batch size	64
Discount factor γ	0.98
Training frequency	Every 4 steps
Gradient steps	1
Target update interval	1,000 steps
Exploration initial ϵ	1.0
Exploration final ϵ	0.05
Exploration fraction	0.25
Total training steps	10,000,000

can learn a satellite selection policy that balances spectral efficiency, handover reduction, outage avoidance, and load awareness. While the BSE baseline provides an upper reference for instantaneous link quality, it may lead to excessive handovers. Conversely, the Holder baseline reduces handover frequency but may remain connected to degraded links.

V. RESULTS

This section presents the evaluation of the trained DQN agent in comparison with the BSE and Holder baseline methods. The DQN agent was evaluated using the test dataset and reward formulation described in the previous sections. Table III summarizes the main evaluation metrics. The BSE policy produces the largest number of handovers, with 1400 handovers, since it frequently changes the serving satellite to follow the best instantaneous link. The Hold policy produces the lowest number of handovers, with 675, as expected from its conservative behavior. The DQN agent presents an intermediate value, with 818 handovers, indicating that it does not avoid handovers completely but performs them more selectively than the BSE baseline.

TABLE III
SUMMARY OF THE EVALUATION RESULTS FOR THE TESTED POLICIES.

Policy	Mean SE	Handovers	HOF Rate	Outage Rate
DQN	0.38	818	0.05	0.61
Holder	0.40	675	0.18	0.63
BSE	0.50	1400	0.21	0.51

The most relevant improvement observed with the DQN agent is in the handover failure rate. The DQN agent achieves the lowest HOF rate, equal to 0.05, while the Holder and BSE policies obtain 0.18 and 0.21, respectively. This result shows that the learned policy is more effective in avoiding invalid or unsuccessful associations. Although the DQN agent performs more handovers than the Holder policy, these handovers are more reliable, resulting in a significantly lower HOF rate.

Figure 2 shows the accumulated mean spectral efficiency during the evaluation. The three policies present similar behavior. The mean spectral efficiency increases during the initial part of the evaluation and then gradually decreases, stabilizing close to 0.4 bps/Hz near the end of the test. The BSE policy

obtains a slightly higher final mean spectral efficiency, while DQN remains close to both baselines. This indicates that the DQN agent does not maximize spectral efficiency alone; instead, it sacrifices part of the instantaneous link performance to reduce handover failures and improve the overall reward.

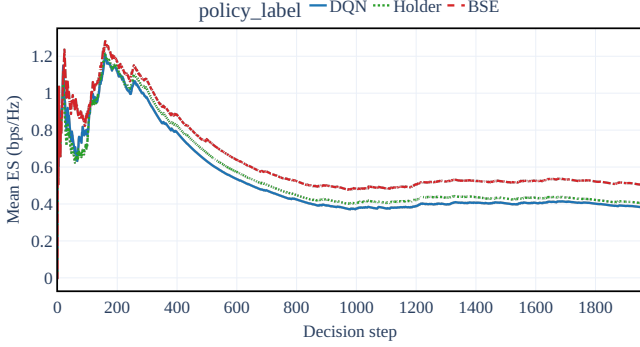


Fig. 2. Accumulated mean spectral efficiency during the evaluation.

The cumulative reward is shown in Figure 3. The DQN agent achieves the best cumulative reward among the evaluated methods, remaining consistently above the two baselines throughout most of the evaluation. The BSE policy obtains the worst cumulative reward, mainly due to its high number of handovers and the highest HOF rate. The Holder policy performs better than BSE, but still accumulates a lower reward than DQN. This behavior is consistent with the proposed reward function, which does not depend only on spectral efficiency, but also penalizes handover attempts, handover failures and outages.

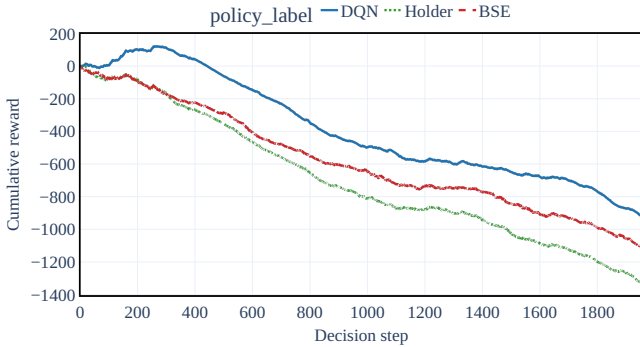


Fig. 3. Cumulative reward during the evaluation.

Overall, the results indicate that the DQN agent learns a more balanced satellite selection policy. It does not behave as aggressively as the BSE policy, which leads to excessive handovers, nor as conservatively as the Holder policy. Instead, the agent performs handovers selectively, achieving the lowest HOF rate and the highest cumulative reward while maintaining spectral efficiency close to the baseline policies.

It is also important to note that the spectral efficiency values observed in this work are lower than those reported in [7]. A possible explanation is the propagation scenario adopted in the simulations. While [7] considers a rural NTN scenario, this work adopts an urban NTN channel model, which is expected to introduce more NLOS conditions, stronger shadowing, and

additional propagation losses. These effects can reduce the received signal quality and, consequently, the resulting spectral efficiency.

VI. CONCLUSION

This paper presents an RL problem formulation that evaluates LEO NTN scenarios for learning-based satellite selection. Using the proposed environment model, a DQN agent was evaluated for this problem and compared with two other baseline policies. The results show that the DQN was able to balance spectral efficiency, handover stability, and handover failure reduction. Compared with the baseline policies, the one provided by the DQN agent achieved the lowest handover failure rate and the highest cumulative reward, indicating that reinforcement learning can be a promising approach for NTN handover management.

Future work includes extending the framework to consider handover execution delays, more detailed traffic models, different NTN scenarios, additional reinforcement learning algorithms, and varying the reward penalty parameters. The integration of more realistic signaling procedures and online interaction between the learning agent and the network simulator may also provide a more complete evaluation of RL-based handover strategies in LEO satellite networks. The dataset created and the agent training code will be available at <https://github.com/CaioBBrasil/RLHandoverNtn-SBRT2026>.

REFERÊNCIAS

- [1] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kotheli, H. Sallouha, A. Colpaert, J. F. M. Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani *et al.*, "Evolution of non-terrestrial networks from 5G to 6G: A survey," *IEEE communications surveys & tutorials*, vol. 24, no. 4, pp. 2633–2672, 2022.
- [2] Z. M. Bakhsh, Y. Omid, G. Chen, F. Kayhan, Y. Ma, and R. Tafazolli, "Multi-satellite MIMO systems for direct satellite-to-device communications: A survey," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 3, pp. 1536–1564, 2024.
- [3] S. Mahboob and L. Liu, "Revolutionizing future connectivity: A contemporary survey on AI-empowered satellite-based non-terrestrial networks in 6G," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 1279–1321, 2024.
- [4] E. Juan, M. Lauridsen, J. Wigard, and P. Mogensen, "Location-Based Handover Triggering for Low-Earth Orbit Satellite Networks," in *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, 2022, pp. 1–6.
- [5] C. Lee, J. Jung, and J.-M. Chung, "Intelligent Dual Active Protocol Stack Handover Based on Double DQN Deep Reinforcement Learning for 5G mmWave Networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7572–7584, 2022.
- [6] S. He, T. Wang, and S. Wang, "Load-Aware Satellite Handover Strategy Based on Multi-Agent Reinforcement Learning," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.
- [7] M. Choi, M. Park, J. Kim, and J.-M. Chung, "Intelligent Handover Scheme for Improved 6G NTN LEO Satellite Network Performance," *IEEE Transactions on Mobile Computing*, 2025.
- [8] 3GPP, "Study on New Radio (NR) to Support Non-Terrestrial Networks," 3rd Generation Partnership Project (3GPP), Technical Report TR 38.811, Sep. 2020, Release 15. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.811/
- [9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," 2013. [Online]. Available: <https://arxiv.org/abs/1312.5602>
- [10] J. Borges, F. Bastos, I. Correa, P. Batista, and A. Klautau, "CAVIAR: Co-simulation of 6G communications, 3-D scenarios, and AI for digital twins," *IEEE Internet of Things Journal*, vol. 11, no. 19, pp. 31 287–31 300, 2024.