2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare | 979-8-3315-9746-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/SENNET64220.2025.11136080

# Autonomous NextG Network Control Using Proximal Policy Optimization: Reducing Latency and Improving Signal Stability

Gourav Anand
*School of Electronics Engineering*
Vellore Institute of Technology
Vellore, India
gourav.anand2021@vitstudent.ac.in

Tanishq Chakravarty
*School of Electronics Engineering*
Vellore Institute of Technology
Vellore, India
tanishq.chakravarty2021@vitstudent.ac.in

Him Raj
*School of Electronics Engineering*
Vellore Institute of Technology
Vellore, India
himraj.2021@vitstudent.ac.in

Saranya Karattupalayam Chidambaram
*School of Electronics Engineering*
Vellore Institute of Technology
Vellore, India
saranya.kc@vit.ac.in

Sudhansu Arya
*School of Electronics Engineering*
Vellore Institute of Technology
Vellore, India
sudhanshu.arya@vit.ac.in

Yogesh Kumar Choukiker
*School of Electronics Engineering*
Vellore Institute of Technology
Vellore, India
yogesh.kumar@vit.ac.in

Abhijit Bhowmick
*School of Electronics Engineering*
Vellore Institute of Technology
Vellore, India
abhijit.bhowmick@vit.ac.in

***Abstract*—The integration of next-generation (NextG) network allows for ultra-high-speed connectivity and seamless merging of devices while introducing new opportunities and challenges for resource distribution, network management, and traffic optimization. Reinforcement Learning is vital in developing self-optimizing systems that respond to dynamic environments. This work investigates implementing Proximal Policy Optimization (PPO) for managing NextG networks, paying particular attention to stabilizing latency, signal strength. The proposed PPO framework achieves optimal traffic network control by policy execution which adjusts network parameters to maintain set limits on power and latency to ensure optimal data transmission. This work demonstrates the usefulness of implementation of PPO in optimization of NextG networks by providing a self-adaptive and scalable network performance enhancement solution enabling higher autonomous control. In this connection, an algorithm named PPO for Post Handover Management is proposed to optimize both parameters: reference signal received power (RSRP) and latency. The described approach enhances the performance of autonomous network management systems and marks the advancement towards more responsive next-generation wireless communication systems.**

***Keywords*—Low-Latency**, **Proximal Policy Optimization (PPO)**, **Reinforcement Learning (RL)**, **Reference Signal Received Power**, **Traffic Optimization**.

## I. Introduction

With the rapid expansion of NextG networks and new requirements for high-speed, low-latency communication, ensuring network stability has emerged as a vital problem. Classical methods based on predefined rules do not perform well in adapting to the fast-changing, intricate characteristics of today's wireless ecosystems. In this regard, reinforcement learning (RL) has emerged as a potent tool for facilitating intelligent, autonomous, data-driven decision-making in network management. This work introduces a new framework for NextG network management which is based on reinforcement learning and implements actor-critic architecture with Proximal Policy Optimization (PPO). Models are taught in a multi-cell scenario simulators and optimized for real-time network performance metrics to achieve goal-oriented handovers while maintaining system stability. The presented framework achieves optimum management of next-generation mobile networks through constant learning and adaptation across various scopes, ensuring versatility in complex systems.

The processes that involve the optimization of Handover (HO) in Beyond NextG (BNextG) and 6G networks is quite complicated due to the alternation in the technology and other requirements pertaining to the new concepts a network system entails. Small cells placed in greater numbers, along with new high frequency bands like mmWave and Terahertz (THz) communications, cause HOs to recur more often, resulting in increased signaling overhead and latency [2]. Moreover, the ability of older systems that rely on automatic handover (HO) to deliver seamless connectivity at enhanced levels of mobility is rather unbounded, which increases the chances of HO failures and reduced service quality (QoS) [10]. The problem is complicated by ultra-dense networks (UDNs) along with the internet of things (IoT) due to an increase in the number of devices making proficient HO management more crucial [5]. HO optimization using Machine Learning (ML) approaches, and solutions built with Deep Reinforcement Learning (DRL), have other types of problems like policy freezing, high computation costs, and adapting to changeable networks [6]. With Proximal Policy Optimization (PPO) and other techniques of reinforcement learning, positive outcomes are offered, yet the challenge of providing sufficient training

still exists when real-time demands and rapid adjustment to practical situations are necessary.

In BNextG and 6G networks, the optimization of handovers (HOs) encounters numerous challenges due to the changes in network requirements and technological advances. The deep deployment of small cells, as well as the addition of new high-frequency bands like mmWave and Terahertz (THz), result in frequent and fast handover (HO) events, increasing signaling overhead and latency for the network [2]. Furthermore, the traditional HO approaches also fail to provide seamless connectivity using high mobility scenarios which results in increased HO inadequacies and reduction in Quality of Service (QoS) [10]. The issue is complicated by ultra-dense networks (UDNs) and the Internet of Things (IoT) which increase the device density while simultaneously compressing the space, thus demanding HO management techniques. Intelligent HO management is made necessary [5]. Machine Learning (ML) and Deep Reinforcement Learning (DRL) are innovative approaches that have attempted to optimize HO decision making. However, unresolved issues such as policy convergence and fluctuation in resource load, especially in dynamic networks remain challenges [6]. The combination of Proximal Policy Optimization (PPO) alongside other reinforcement learning approaches has shown potential to resolve these issues. The requirement to efficiently train the system and adapt assumptions in real-time is still a challenge open for research [3].

In response to these challenges, a new approach is implemented through a PPO-based framework for the self-optimizing, intelligent management of NextG networks. This framework ensures the optimal network distribution by maintaining stability in the latency and signal strength equilibrium using dynamic control of network parameter. Through policy-based learning, PPO adapts and optimizes resource management and traffic flow in dynamic environments which reduces congestion and constitutes smoother traffic. In addition, it also controls the power (dBm) allocation needed for efficient network performance in terms of reduced latency and improved network performance. The proposed approach provides scalable and adaptable solutions to enable real-time traffic management control. This kind of approach forms a solid basis for next generation wireless communication systems.

## II. System Model and Problem Formulation

### A. Network Distribution

In NextG networks, the concern of network distribution is associated with how resources like bandwidth, power, and spectrum are managed across set small cells or base stations to optimize coverage, reliability, and interference mitigation.

This issue becomes further complicated due to the unpredictable nature of user mobility and everchanging data requirements, which calls for preemptive resource distribution. More flexible and adaptive approaches are needed in mobile networks as compared to traditional cellular networks where large fixed covered area base stations usually suffice.

### B. Network Traffic Optimization

NextG traffic optimization does involve controlling the flow of information data through the network which guarantees high signal strength, low latency, and optimal performance even when the network is stressed by heavy traffic. Beyond that, however, the primary focus tries to balance the distribution of traffic across different base stations to avoid congestion.

To accommodate the requirements of all the devices, it has to determine in real time how to optimally route and balance the loads – which is why traffic optimization in NextG turns into a problem of dynamic optimization.

### C. Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) is a sophisticated variant of reinforcement learning which, unlike Q-Learning that estimates values, enhances decisions directly on actions. Q-Learning is effective in cases with few fixed options, but PPO is more advantageous in cases of continuous actions such as modulation of power, bandwidth, or resource allocation in NextG networks. [3]

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[ \min \left( r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) \right] \tag{1}$$

TABLE I. Definitions of Symbols Used in the Algorithm

| Symbol | Description |
|---|---|
| $\theta$ | Policy parameter |
| $\hat{E}_t$ | Empirical expectation over timesteps |
| $r_t$ | Ratio of the probability under the new and old policies |
| $\hat{A}_t$ | Estimated advantage at time $t$ |
| $\varepsilon$ | Hyperparameter, usually 0.1 or 0.2 |

In Table 1, symbol and it's information are given.

## III. Learning Framework

In NextG networks, reinforcement learning (RL) is increasingly used to optimize dynamic tasks such as network resource allocation and traffic routing. Agents—such as base stations or user devices—interact with the environment by adjusting parameters like the number of connected users, transmission speed, and cell configurations. These interactions change the network state, with the environment providing feedback through key performance indicators like signal strength, latency, and throughput.

Proximal Policy Optimization (PPO) has emerged as a preferred RL technique in this domain due to its balance between exploration and policy stability. PPO is well-suited for NextG use cases because it supports continuous action spaces, which are essential for fine-grained tasks such as power control, load balancing, spectrum management, and bandwidth allocation. It offers the flexibility needed for real-time decision-making while avoiding drastic policy updates that can lead to performance instability.

## IV. Methodology

In a NextG network, decision making depends on numerous elements such as signal level, delay, number of UE cells and their respective velocities and the learning structure utilized. This study centers on assessing how Proximal Policy Optimization (PPO) copes with the RSRP of the base cell and how latency changes within the learning framework. Also, caring about the reward function evaluation is important since the learning process and its outcome are strongly determined by the reward signal designed.

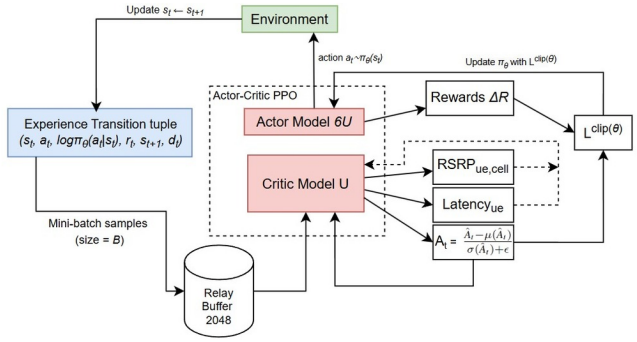


Fig. 1. System Model

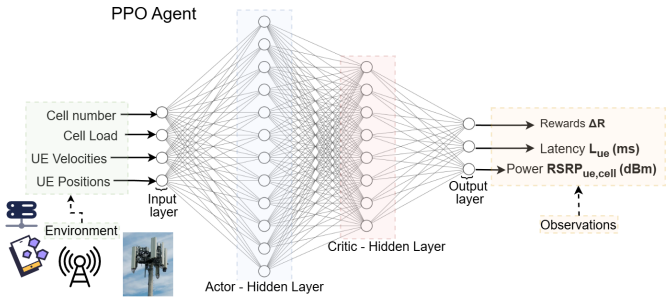


Fig. 2. Flow Graph

The flow graph for Our PPO-based handover management framework uses input parameters like cell count, individual cell load, and UE velocity and position to form a comprehensive state for dynamic decision-making in NextG networks. Outputs like reward, latency, and RSRP help assess and optimize network performance.

.

RSRP is a critical metric for assessing signal strength, calculated using the formula:

$$\text{RSRP}_{ue,cell} = 140 - 30 \cdot \log_{10}\left(\|\mathbf{p}_{ue} - \mathbf{p}_{cell}\|_2 + 10^{-5}\right) + \mathcal{N}(0,2) \tag{2}$$

where denotes the power of the reference signal in every resource element and represents the total number of reference signal resource elements. The network performance improves with higher RSRP values, indicating better signal quality, while low RSRP values may lead to poor connectivity.

Since network conditions fluctuate, analyzing latency variations is essential. The latency change can be modeled as:

**Algorithm 1** PPO for Post Handover Management

**Input:** Number of cells $C$, number of UEs $U$, grid size $G$, max steps $T$, learning rate $\alpha$, discount factor $\gamma$, clip parameter $\epsilon$, batch size $B$, epochs $E$.
**Output:** Trained PPO model for handover management.
Initialize NextG environment with $C$ cells, $U$ UEs, grid size $G$, and max steps $T$.
Initialize PPO actor-critic networks with state dimension $6U$ and action dimension $U$.
Initialize replay buffer $\mathcal{B}$ .
**for** episode $= 1$ to $N$ **do**
    Reset environment and observe initial state $s_0$.
    Initialize total reward $R \leftarrow 0$.
    Initialize episode latencies $\mathcal{L} \leftarrow \emptyset$.
    Initialize episode RSRP $\mathcal{R} \leftarrow \emptyset$.
    **for** step $t = 1$ to $T$ **do**
        Define state $s_t$ as: RSRP, latency, cell load, and UE velocities.
        Select action $a_t$ using PPO actor network: $a_t \sim \pi_\theta(s_t)$.
        Perform handover if $a_t = 1$ and target cell $\neq$ current cell.
        Execute action $a_t$, observe next state $s_{t+1}$, reward $r_t$, and done flag $d_t$.
        Calculate reward $r_t$ based on RSRP improvement, cell load, and handover penalties.
        Store transition $(s_t, a_t, \log \pi_\theta(a_t|s_t), r_t, s_{t+1}, d_t)$ in buffer $\mathcal{B}$.
        Update state: $s_t \leftarrow s_{t+1}$.
        Update total reward: $R \leftarrow R + r_t$.
        Record latency and RSRP metrics: $\mathcal{L} \leftarrow \mathcal{L} \cup \{\text{latency}\}$, $\mathcal{R} \leftarrow \mathcal{R} \cup \{\text{RSRP}\}$.
    **end for**
    **if** buffer size $> B$ **then**
        Sample mini-batch $\mathcal{M}$ of size $B$ from $\mathcal{B}$.
        Compute advantages $\hat{A}_t$ using critic network.
        Normalize advantages: $\hat{A}_t = \frac{\hat{A}_t - \mu(\hat{A}_t)}{\sigma(\hat{A}_t)+\epsilon}$.
        **for** epoch $= 1$ to $E$ **do**
            Compute actor loss $\mathcal{L}_{\text{actor}}$ using PPO clipped objective.
            Compute critic loss $\mathcal{L}_{\text{critic}}$ using MSE.
            Update actor and critic networks using gradient descent.
        **end for**
    **end if**
**end for**

$$\text{Latency}_{ue} = C_1 + C_2 \cdot \text{CellLoad}_{\text{current\_cell}} \tag{3}$$

RSRP and other network metrics such as congestion and inter- ference also impact latency variation. For instance, real-time video streaming and online gaming will require sub 50 ms latency, and any value exceeding this is damaging to the users' experience. Here, latency can be modeled as a multi-factor scope problem through the use of parameters such as cell load, using constant parameters like $C_1$for base latency and $C_2$ for cell load impact, to reflect pragmatic network dynamics.

To optimize network performance, PPO relies on a reward function, which determines whether an action improves or worsens network quality. The reward function is defined as:

$$\begin{aligned}\Delta R = 0.3(\text{RSRP}_{new} - \text{RSRP}_{old}) \\ - 0.02(\text{Load}_{new} - \text{Load}_{old}) - 0.2\end{aligned} \tag{4}$$

## V. RESULTS AND OBSERVATIONS

Our findings indicate that Proximal Policy Optimization (PPO) is more appropriate for NextG networks compared to Q-Learning, especially in scenarios that require dynamic and continuous control. The policy gradient method of PPO allows for real-time adjustments in bandwidth distribution, power regulation, and interference oversight. It provides stability through the use of clipped objective functions and strikes a strong balance between exploration and exploitation due to entropy regularization.

### Baseline Model

The Baseline model (Blue) indicates the default RL approach with hyperparameters tuned to what is assumed optimal default to not interfere with stable latency and power. The PPO model in this instance has a learning rate $\alpha$, discount factor $\gamma$, clipping threshold $\epsilon$, and batch size $N$.

### High Learning Rate ($\alpha_{\text{Low}}$) Model

The $\alpha_{\text{Low}}$ model (Orange) goes with a higher learning rate, which allows rewards to be harvested quickly, but too rapidly increases latency as well leading to a perception of instability over the long haul. The PPO setup includes $\alpha$, $\gamma$, $\epsilon$, batch size $N$.

### Low Discount Factor ($\gamma_{\text{Low}}$) Model

The $\gamma_{\text{Low}}$ (Green) which adjusts the environment so that latency and power stays as stable as possible. In this instance, the model is set with a learning rate of $\alpha$, discount factor $\gamma$, and clipping threshold of $\epsilon$. In addition, the model utilizes a batch size of ( N ).

### Deeper Neural Network ($\text{Net}_{\text{Deeper}}$) Model

The $\text{Net}_{\text{Deeper}}$ model (Red) incorporates a more complex neural network architecture, allowing it to capture deeper relationships within the data. This model achieves some of the highest rewards, demonstrating its ability to learn optimal policies effectively. In this variation, the PPO model is configured with a learning rate $\alpha$, a discount factor $\gamma$, a clipping threshold $\epsilon$, and a batch size of $N$. Additionally, we further improve the architecture by adding multi-layered structures for the actor and critic networks individually with aims at enabling greater learning efficiency and enhancing policy robustness. This leads to improving the model which in turn improves the generalization ability of the model for various scenarios with respect to networks and increases handover performance in dynamic NextG environments.

The following notations are used throughout the figures: $\gamma_{\text{Low}}$ refers to the Low_Gamma configuration which has a lower gamma value in the relevant context. In the same manner, $\alpha_{\text{Low}}$ relates to High_LR as a higher learning rate setting. $\text{Net}_{\text{Deeper}}$ is also used for the Deeper_Net architecture which is a neural network with greater depth or more layers.

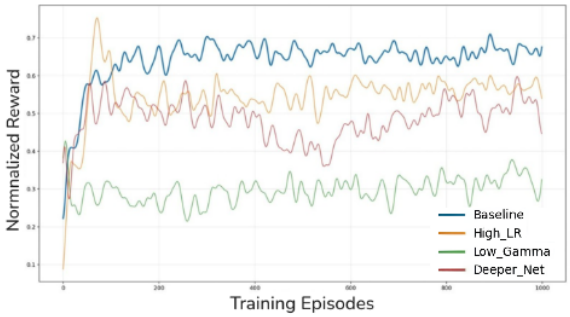


Fig. 3. Reward Progression in PPO-Based for NextG Environments

As shown in Fig. 3, the normalized reward performance of four RL configurations is compared over 1000 training episodes. The Baseline (blue) is the most stable, plateauing around 0.65, reflecting a balanced trade-off between learning and stability. $\alpha_{\text{Low}}$ (orange) reaches the highest reward (0.72) but with large fluctuations, indicating fast yet unstable convergence. $\text{Net}_{\text{Deeper}}$ (red) peaks at 0.55, showing moderate but steady performance. $\gamma_{\text{Low}}$ (green) performs worst, with rewards around 0.35 due to poor long-term reward estimation. The results highlight the trade-off between convergence speed and training stability, with Baseline offering the best balance.

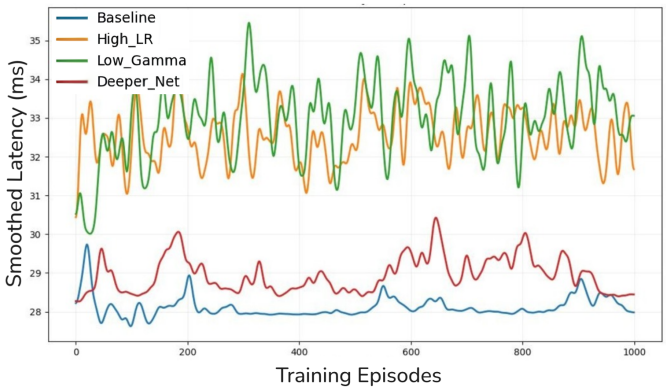


Fig. 4. Latency observation in PPO-Based for NextG Environments

As shown in Fig. 4, the effect of different training configurations on latency performance in the NextG environment is evaluated. The Baseline (blue) shows the most stable latency around 28 ms. $\text{Net}_{\text{Deeper}}$ (red) incurs a slight overhead (28.5–29.5 ms), indicating that deeper networks trade a small

latency increase for improved stability. $\alpha_{\text{Low}}$ (orange) and $\gamma_{\text{Low}}$ (green) display severe latency spikes above 35 ms, highlighting the negative impact of aggressive learning rates and low discount factors. These results confirm that deeper architectures enhance stability, while poorly tuned hyperparameters degrade latency performance.

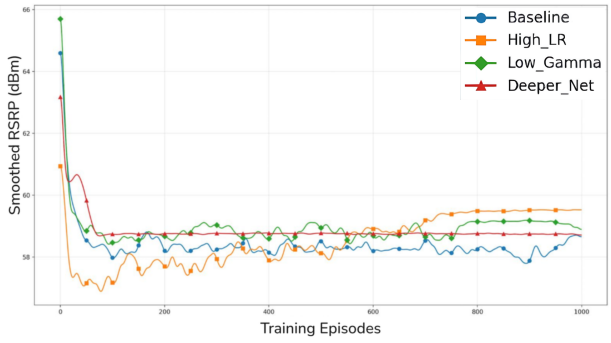


Fig. 5. Signal strength in PPO-Based for NextG Environments

As shown in Fig. 5, RSRP values over 1000 episodes show initial instability due to baseline values ¿64 dBm, which stabilize within 50 episodes. The Baseline model maintains steady strength at 59.5–60 dBm, $\text{Net}_{\text{Deeper}}$ hovers near 59 dBm, $\alpha_{\text{Low}}$ varies between 57.5–58 dBm, and $\gamma_{\text{Low}}$ is most unstable at 56.5–57.5 dBm. Results suggest deeper architectures yield higher rewards and stable signals, while low $\gamma$ and high learning rates increase variability, impacting reliability. PPO proves effective for optimizing NextG networks with minimal latency trade-offs.

The comparative analysis of PPO in NextG post-handover scenarios shows that networks with dynamic, continuous action spaces perform more effectively. PPO addresses key challenges like bandwidth allocation, power control, and interference management. Its clipped objective and entropy regularization ensure stable training and balanced exploration. PPO outperforms other models in total rewards across training episodes.

Latency analysis shows the Baseline model performs best ( 28 ms), while $\text{Net}_{\text{Deeper}}$ incurs a slight increase (28.5–29.5 ms). $\alpha_{\text{Low}}$ and $\gamma_{\text{Low}}$ exceed 35 ms and show instability. Deeper models improve reliability with minor latency costs, but high learning rates and low gamma values lead to divergence.

RSRP trends over 1000 episodes reveal initial instability (¿64 dBm), stabilizing by episode 50. Baseline holds at 59.5–60 dBm, $\text{Net}_{\text{Deeper}}$ at 59 dBm, $\alpha_{\text{Low}}$ between 57.5–58 dBm, and $\gamma_{\text{Low}}$ at 56.5–57.5 dBm. PPO proves effective, with deeper architectures yielding higher rewards and stable signals, with minimal latency trade-offs.

## VI. Conclusion

This research demonstrates the proficient use of Proximal Policy Optimization (PPO) within artificial intelligent-driven, stable, and adaptive control for NextG network management, showing significant strides in performance outcomes. Within the configurations tested, the Baseline architecture was the overall best in total reward, high signal strength, but minor latency issues. The $\alpha_{\text{Low}}$ setting converged quickly, but showed large performance swings while $\gamma_{\text{Low}}$ exhibited slower and unstable convergence. $\text{Net}_{\text{Deeper}}$ provided insights on stability but was deemed resource-heavy. For latency and signal strength, baseline remained the top trackable performer with a variance of 1.2 ms which indicates low but stable fluctuation further validating their superiority. Total reward is persistently sustained indicating maintained clarity through policy updates, which enables synthesis of new policies. RSRP values proved effective post-handover by remaining within 57-59 dBm. Strategically, PPO-based approaches showcased dynamic control over critical network parameters such as latency and power, exhibiting exceptional responsiveness in contrast to other forms, demonstrating significant proficiency.

## References

[1] S. Arya, J. Yang, P. T. Grogan, and Y. Wang, "Real-Time UAV Collaborative Beam Reforming for Coexistent Satellite-Terrestrial Communications," *IEEE Aerospace Conference*, vol. 2024, pp. 1–10, 2024, doi: 10.1109/AERO58975.2024.10521379.

[2] S. Alraih, R. Nordin, A. Abu-Samah, I. Shayea, and N. F. Abdullah, "A Survey on Handover Optimization in Beyond NextG Mobile Networks: Challenges and Solutions," *IEEE Access*, vol. 11, no. 7, pp. 1787–1799, 2023, doi: 10.1109/ACCESS.2023.3284905.

[3] Y. Gu, Y. Cheng, C. L. P. Chen, and X. Wang, "Proximal Policy Optimization With Policy Feedback," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, pp. 4600–4610, 2021, doi: 10.1109/TSMC.2021.3098451.

[4] J. Kaur, M. A. Khan, M. Iftikhar, M. Imran, and Q. E. U. Haq, "Machine Learning Techniques for NextG and Beyond," *IEEE Access*, vol. 9, pp. 23472–23488, 2021, doi: 10.1109/ACCESS.2021.305155.

[5] M. S. Mollel, A. I. Abubakar, M. Ozturk, S. F. Kaijage, M. Kisangiri, and S. Hussain, "A Survey of Machine Learning Applications to Handover Management in NextG and Beyond," *IEEE Access*, vol. 29, pp. 45770–45802, 2021, doi: 10.1109/ACCESS.2021.3067503.

[6] C. Lee, J. Jung, and J. M. Chung, "Intelligent Dual Active Protocol Stack Handover Based on Double DQN Deep Reinforcement Learning for NextG mmWave Networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 469–484, 2022, doi: 10.1109/TVT.2022.3170420.

[7] T. Li, X. Zhu, and X. Liu, "An End-to-End Network Slicing Algorithm Based on Deep Q-Learning for NextG Network," *IEEE Access*, vol. 8, no. 1, pp. 122229–122240, 2020, doi: 10.1109/ACCESS.2020.3006502.

[8] M. Raeisi and A. B. Sesay, "Power Control of NextG-Connected Vehicular Network Using PPO-Based Deep Reinforcement Learning Algorithm," *IEEE Access*, vol. 12, no. 12, pp. 96387–96403, 2024.

[9] A. del Rio, D. Jimenez, and J. Serrano, "Comparative Analysis of A3C and PPO Algorithms in Reinforcement Learning: A Survey on General Environments," *IEEE Access*, vol. 12, no. 10, pp. 146795–146806, 2024, doi: 10.1109/ACCESS.2024.3472473.

[10] N. Zohar, "Beyond NextG: Reducing the handover rate for high mobility communications," *Journal of Communications and Networks*, vol. 24, pp. 154–165, 2022, doi: 10.23919/JCN.2022.000001.

[11] A. Jain, E. Lopez-Aguilera, and I. Demirkol, "Evolutionary 4G/NextG Network Architecture Assisted Efficient Handover Signaling," *IEEE Access*, vol. 7, pp. 256–283, 2018, doi: 10.1109/ACCESS.2018.2885344.

[12] S. Schwarzmann, C. C. Marquezan, R. Trivisonno, and S. Nakajima, "ML-Based QoE Estimation in NextG Networks Using Different Regression Techniques," *IEEE Transactions on Network and Service Management*, vol. 6, no. 2, pp. 3516–3532, 2022, doi: 10.1109/TNSM.2022.3179924.

[13] M. B. M. Kamel, I. A. Najm, and A. K. Hamoud, "Congestion Control Prediction Model for NextG Environment Based on Supervised and Unsupervised Machine Learning Approach," *IEEE Access*, vol. 12, pp. 91127–91139, 2024, doi: 10.1109/ACCESS.2024.3416863.

[14] L. A. Garrido, A. Dalgkitsis, K. Ramantas, and A. Ksentini, "Resource Demand Prediction for Network Slices in NextG Using ML Enhanced With Network Models," *IEEE Access*, vol. 73, pp. 11848–11861, 2024, doi: 10.1109/TVT.2024.3373490.

[15] M. U. Iqbal, E. A. Ansari, and S. Akhtar, "Improving the QoS in NextG HetNets Through Cooperative Q-Learning," *IEEE Access*, vol. 10, pp. 19654–19676, 2022, doi: 10.1109/ACCESS.2022.3151090.

[16] J. Mu, X. Jing, Y. Zhang, Y. Gong, and R. Zhang, "Machine Learning-Based NextG RAN Slicing for Broadcasting Services," *IEEE Transactions on Broadcasting*, vol. 68, pp. 295–304, 2021, doi: 10.1109/TBC.2021.3122353.

[17] J. Li and X. Zhang, "Deep Reinforcement Learning-Based Joint Scheduling of eMBB and URLLC in NextG Networks," *IEEE Wireless Communications Letters*, vol. 9, pp. 1543–1546, 2020, doi: 10.1109/LWC.2020.2997036.

[18] . U. Iqbal, E. A. Ansari, and S. Akhtar, "Improving the QoS in 5G HetNets Through Cooperative Q-Learning," *IEEE Access*, vol. 10, pp. 19654–19676, 2022, doi: 10.1109/ACCESS.2022.3151090.

[19] S. Troia, A. F. R. Vanegas, and L. M. M. Zorello, "Admission Control and Virtual Network Embedding in NextG Networks: A Deep Reinforcement-Learning Approach," *IEEE Access*, vol. 10, pp. 15860–15875, 2022, doi: 10.1109/ACCESS.2022.3148703.

[20] L. A. Garrido, A. Dalgkitsis, K. Ramantas, and A. Ksentini, "Resource Demand Prediction for Network Slices in 5G Using ML Enhanced With Network Models" *IEEE Access*, vol. 73, pp. 11848–11861, 2024, doi: 10.1109/TVT.2024.3373490.