



Deep Reinforcement Learning for Autonomous Network Optimization

Shruthi Pokuri¹, Sameer Nair², Sumith Raina³, Vijay Ronith Reddy Duggempudi⁴

Department of Computer Science, College of Engineering, Trivandrum, India¹⁻⁴

Abstract: The growing complexity of next-generation communication networks — encompassing heterogeneous 5G and beyond-5G (B5G) infrastructures, software-defined networking (SDN), network function virtualization (NFV), and ultra-dense small cell deployments — has rendered manual and rule-based network management approaches fundamentally inadequate. Autonomous network optimization, in which self-configuring, self-optimizing, and self-healing control loops adapt network parameters in real time without human intervention, has emerged as the essential operational paradigm for future intelligent networks. This paper proposes a Deep Reinforcement Learning framework for Autonomous Network Optimization (DRL-ANO) that employs a multi-agent Soft Actor-Critic (MASAC) architecture to simultaneously optimize radio resource management, traffic engineering, and network slice resource allocation across a heterogeneous multi-cell environment. The framework introduces a Graph Attention Network (GAT)-based state encoder that captures the dynamic topology and interference relationships of the network as a relational observation, enabling the MASAC agents to make topology-aware optimization decisions at millisecond timescales. A safe exploration mechanism enforcing hard Quality of Service (QoS) constraints via Lagrangian relaxation is incorporated to prevent the DRL agents from violating service level agreements during online learning and deployment. Security provisions include adversarial observation detection, encrypted policy distribution, and anomaly-triggered safe fallback modes. Experimental evaluation on a simulated heterogeneous 5G network with 19 base stations, 6 active network slices, and dynamic traffic generated from real-world LTE traces demonstrates a 39% improvement in network-wide spectral efficiency, a 44% reduction in end-to-end slice latency, a 31% improvement in resource utilization, and a 97.2% QoS constraint satisfaction rate, significantly outperforming rule-based, single-agent DRL, and model predictive control baselines.

Keywords: Deep Reinforcement Learning; Autonomous Network Optimization; Soft Actor-Critic; Multi-Agent Systems; Graph Attention Networks; Network Slicing; 5G; Radio Resource Management.

I. INTRODUCTION

The deployment of fifth generation (5G) New Radio networks and the ongoing standardization of beyond-5G (B5G) and 6G architectures are fundamentally redefining the scope and complexity of network management. Where previous generations of mobile networks were characterized by relatively homogeneous radio access technologies and well-understood traffic patterns, modern networks combine macro cells, small cells, millimeter-wave links, device-to-device communication, and non-terrestrial network segments into a heterogeneous multi-layer architecture serving diverse traffic classes with radically different latency, throughput, reliability, and mobility requirements [1]. Network slicing — the virtualization of a shared physical infrastructure into multiple logically isolated end-to-end network instances, each tailored to a specific service vertical — adds a further dimension of management complexity by requiring simultaneous optimization of slice-level QoS guarantees alongside network-wide resource efficiency [2].

Traditional network optimization relies on a combination of static configuration tables, model-based controllers such as proportional-integral-derivative (PID) regulators and model predictive control (MPC), and rule-based self-organizing network (SON) functions standardized by 3GPP. These approaches offer well-understood stability properties but are inherently reactive: they optimize around a fixed operating point and cannot generalize to novel traffic conditions, topology changes, or failure scenarios outside their design envelope [3]. The scale and dynamism of next-generation networks — where a single urban deployment may encompass thousands of cells and tens of millions of simultaneously active devices — make the manual parameterization of rule-based controllers both operationally infeasible and economically unsustainable [4].

Deep reinforcement learning offers a compelling alternative: a DRL agent learns a general-purpose optimization policy through interaction with the network environment, adapting to changing conditions without explicit system models or manual re-parameterization [5]. However, direct application of single-agent DRL to heterogeneous network optimization faces three fundamental challenges. First, the joint action space of a multi-cell network — encompassing transmit power, beamforming weights, carrier aggregation configurations, and slice bandwidth allocations across dozens of base stations — is combinatorially intractable for a single agent. Second, DRL exploration in live networks risks violating QoS



constraints that service level agreements legally and technically mandate. Third, standard DRL state representations based on tabular network statistics fail to capture the spatial and interference topology relationships that are critical for coordinated resource management decisions [6].

This paper addresses all three challenges through the DRL-ANO framework, making the following principal contributions:

- A formal multi-agent MDP formulation of the heterogeneous network optimization problem, decomposing the joint action space across geographically local agent scopes while preserving global coordination through shared critic training and message-passing between neighboring agents.
- A multi-agent Soft Actor-Critic (MASAC) architecture with entropy regularization that promotes diverse exploration while maintaining policy stability in the non-stationary multi-agent environment of a live heterogeneous network.
- A Graph Attention Network (GAT)-based state encoder that represents the network as a dynamic interference graph, producing topology-aware node embeddings that equip each agent with relational awareness of its neighbors' states and interference contributions.
- A safe exploration mechanism based on Lagrangian relaxation that enforces hard per-slice QoS constraints as penalty terms in the actor optimization objective, provably bounding constraint violation probability during both training and deployment.
- Comprehensive experimental evaluation on a simulated 5G heterogeneous network with realistic LTE-derived traffic traces, demonstrating significant improvements over rule-based, single-agent DRL, and MPC baselines across spectral efficiency, latency, resource utilization, and QoS satisfaction metrics.

The rest of this paper is organized as follows. Section II presents the formal system model. Section III details the DRL-ANO scheme design and analysis. Section IV provides performance evaluation. Section V concludes the paper.

II. SYSTEM MODEL FOR DEEP REINFORCEMENT LEARNING FOR AUTONOMOUS NETWORK OPTIMIZATION (DRL-ANO)

Network Architecture, Formal Definitions, and Security Assumptions

The DRL-ANO system models a heterogeneous 5G network comprising a set of B base stations $\{b_1, b_2, \dots, b_B\}$ — including both macro eNodeBs and small cell gNBs — serving a dynamic population of user equipment (UE) across K active network slices $\{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}$. Each slice Σ_k is characterized by a QoS requirement tuple $Q_k = (\lambda_k, \delta_k, \rho_k)$ specifying the minimum guaranteed throughput λ_k (Mbps), maximum tolerated end-to-end latency δ_k (ms), and minimum packet delivery reliability ρ_k . The physical layer is modeled with realistic path loss, log-normal shadowing, and multi-path fading following 3GPP TR 38.901 channel models for Urban Macro and Urban Micro scenarios.

A network-wide SDN controller with northbound API interfaces coordinates the DRL-ANO agents and enforces global policy consistency. Each base station b_i hosts a local DRL-ANO agent A_i responsible for optimizing its local radio resource management (RRM) and slice bandwidth allocation decisions. The formal algorithmic components of DRL-ANO are as follows:

- $\text{SystemInit}(B, K, k_{\text{sec}}) \rightarrow (\Gamma_{\text{net}}, \{\Theta_i^0\}, G_0)$: Initializes global network configuration Γ_{net} (slice QoS profiles, channel model parameters, agent communication graph topology), the initial MASAC actor-critic parameters $\{\Theta_i^0\}$ for each agent A_i , and the initial network interference graph G_0 , given B base stations, K slices, and security parameter k_{sec} .
- $\text{EncodeState}(G_t, x_i^t) \rightarrow h_i^t$: Agent A_i 's GAT-based state encoder processes the network interference graph G_t and local observation vector x_i^t (comprising per-slice buffer occupancy, channel quality indicators, current resource allocations, and QoS violation counters) through L graph attention layers to produce the topology-aware hidden state embedding h_i^t .
- $\text{SelectAction}(h_i^t, \Theta_i^{\text{actor}}, \lambda_i) \rightarrow a_i^t$: The MASAC actor network samples action a_i^t from the maximum-entropy stochastic policy $\pi_{\{\Theta_i\}}(a_i^t | h_i^t)$, subject to Lagrangian QoS penalty terms parameterized by constraint multipliers λ_i , returning the joint RRM and slice allocation decision vector.
- $\text{ComputeReward}(s^t, a^t) \rightarrow r^t$: The reward function aggregates network-wide spectral efficiency, slice latency, and resource utilization into a scalar reward r^t , with additive Lagrangian penalty terms penalizing per-slice QoS violations in proportion to their magnitude and duration.
- $\text{UpdateCritic}(\{\tau_i\}, \Theta^Q) \rightarrow \Theta^Q_{\text{new}}$: The centralized critic network is updated using a mini-batch of joint experience tuples $\{\tau_i\} = \{(h_i^t, a_i^t, r^t, h_i^{t+1})\}$ from the shared replay buffer, minimizing the soft Bellman residual with entropy regularization coefficient α_{entropy} .



- UpdateActor(Θ_i^{actor} , Θ^Q , λ_i) $\rightarrow$ ($\Theta_i^{\text{actor_new}}$, λ_{i_new}): Each agent's actor network is updated by maximizing the constrained soft policy gradient objective; the Lagrange multipliers λ_i are simultaneously updated by gradient ascent on the constraint violation magnitude, implementing dual ascent for safe policy optimization.

Table I — Notation and Descriptions

Symbol	Description
B	Total number of base stations (macro eNodeBs and small cell gNBs)
K	Number of active network slices
b_i	i-th base station in the heterogeneous network
A_i	DRL-ANO agent co-located with base station b_i
Σ_k	k-th network slice with QoS requirement tuple Q_k
$Q_k = (\lambda_k, \delta_k, \rho_k)$	QoS tuple: min throughput, max latency, min reliability for slice k
G_t	Dynamic network interference graph at time step t
x_i^t	Local observation vector of agent A_i at time step t
h_i^t	GAT-encoded topology-aware state embedding of agent A_i at time t
a_i^t	Joint RRM and slice allocation action vector of agent A_i at time t
r^t	Scalar reward signal at time step t
Θ_i^{actor}	Actor network parameters of agent A_i
Θ^Q	Centralized critic network parameters (shared across all agents)
λ_i	Lagrange constraint multiplier vector of agent A_i for QoS enforcement
α_{entropy}	Entropy regularization temperature coefficient of SAC
τ_i	Experience tuple of agent A_i : ($h_i^t, a_i^t, r^t, h_i^{t+1}$)
Γ_{net}	Global network configuration: slice profiles, channel model, graph topology
k_{sec}	Security parameter for cryptographic and attestation functions

Reward Function and QoS Constraint Model

The reward function r^t is a weighted composite of three normalized network performance indicators: $r^t = w_{\text{SE}} * r_{\text{SE}}^t + w_{\text{lat}} * r_{\text{lat}}^t + w_{\text{util}} * r_{\text{util}}^t - \sum_{k=1}^K \lambda_k^t * C_k^t$, where r_{SE}^t is the network-wide spectral efficiency normalized by the Shannon capacity bound, r_{lat}^t is the negative mean per-slice end-to-end latency normalized by the tightest slice latency SLA, r_{util}^t is the physical resource block (PRB) utilization efficiency, and $C_k^t = \max(0, \delta_k^t - \delta_k) + \max(0, \lambda_k - \lambda_k^t) + \max(0, \rho_k - \rho_k^t)$ is the QoS constraint violation magnitude for slice k at time t. The Lagrange multiplier λ_k^t grows whenever $C_k^t > 0$, progressively increasing the penalty on QoS violations until the constraint is satisfied, implementing a self-tuning safe optimization mechanism that requires no manual penalty weight tuning.

Security and Adversarial Robustness Model

The DRL-ANO deployment faces a threat model in which adversarial entities — including compromised UEs, rogue small cells, or man-in-the-middle attackers on the SDN southbound interface — may attempt to manipulate the agents' observations to induce suboptimal or unsafe resource allocation decisions. Three defense mechanisms are integrated. First, an observation integrity verifier using a one-class neural network classifier detects out-of-distribution state observations that deviate significantly from the manifold of legitimate network measurements, triggering a safe fallback



mode that reverts to the rule-based SON controller until observation integrity is restored. Second, all MASAC policy network weights distributed from the SDN controller to base station agents are transmitted over TLS 1.3 encrypted channels and signed with ECDSA P-256, preventing unauthorized policy injection. Third, a model integrity monitor verifies actor network weight checksums at each policy update, detecting unauthorized in-situ model modifications on the base station hardware.

III. DRL-ANO SCHEME DESIGN AND ANALYSIS

Graph Attention Network State Encoder

The network interference graph $G_t = (V, E_t, W_t)$ models base stations as nodes $V = \{b_1, \dots, b_B\}$, with directed edges E_t representing interfering links between base stations whose coverage areas overlap at time t , and edge weights W_t^{ij} encoding the measured inter-cell interference-to-noise ratio (INR) on each link. The GAT encoder processes G_t through $L = 4$ graph attention layers, each computing node embeddings $h_i^{(l+1)} = \sigma(\sum_{j \in N(i)} \alpha_{ij}^{(l)} * W^{(l)} * h_j^{(l)})$, where $N(i)$ denotes the neighbors of node i in G_t , $W^{(l)}$ is a learnable linear transformation, and $\alpha_{ij}^{(l)}$ is the attention coefficient computed as $\alpha_{ij} = \text{softmax}(\text{LeakyReLU}(a^T [W_{hi} \parallel W_{hj} \parallel e_{ij}]))$, with e_{ij} being the edge feature encoding the INR between cells i and j .

The GAT encoder provides three critical advantages over tabular state representations for network optimization. First, it produces permutation-equivariant embeddings, meaning that the learned policy generalizes across networks with different numbers of base stations and varying cell configurations without retraining. Second, multi-head attention (8 heads in the DRL-ANO implementation) allows each agent to simultaneously attend to different aspects of its neighbors' states — such as load imbalance, interference severity, and QoS violation status — producing richer relational context than single-channel aggregation. Third, because G_t is updated at each control epoch to reflect the current interference topology, the encoder naturally tracks time-varying interference patterns arising from user mobility and traffic fluctuation.

Multi-Agent Soft Actor-Critic with Centralized Training

DRL-ANO adopts the Centralized Training with Decentralized Execution (CTDE) paradigm. During training, a centralized critic $Q_{\phi}(h^t, a^t)$ receives the joint state-action tuple (h^t, a^t) from all agents, enabling it to learn value estimates that account for the global consequences of joint agent decisions, including inter-cell interference externalities that are invisible to individual agents' local observations. During deployment, each agent's actor $\pi_{\theta_i}(a_i | h_i^t)$ executes based solely on its local GAT-encoded state, requiring no inter-agent communication at inference time and enabling sub-millisecond distributed decision-making at each base station.

The SAC objective adds a maximum-entropy regularization term to the standard RL objective: $J(\pi) = E[\sum_t \gamma^t (r^t + \alpha_{\text{entropy}} * H(\pi(h^t)))]$, where $H(\pi) = -E[\log \pi(a|h)]$ is the policy entropy. Entropy maximization encourages the agents to maintain diverse, exploratory policies that avoid premature convergence to locally optimal but globally suboptimal solutions — a particularly important property in the non-stationary multi-agent setting where the environment experienced by each agent changes as other agents' policies evolve during training. The temperature parameter α_{entropy} is adaptively tuned by minimizing a cross-entropy objective targeting a desired minimum entropy H_{target} , eliminating manual temperature scheduling.

Safe Exploration via Lagrangian Relaxation

The Lagrangian safe exploration mechanism transforms the constrained policy optimization problem: $\max_{\pi} J(\pi)$ s.t. $E[C_k^t] \leq \xi_k$ for all k , into an unconstrained saddle-point problem by introducing dual variables (Lagrange multipliers) $\lambda_k \geq 0$ for each constraint. The Lagrangian $L(\pi, \lambda) = J(\pi) - \sum_k \lambda_k * (E[C_k^t] - \xi_k)$ is optimized by alternating gradient ascent on λ_k (when constraint k is violated) and gradient descent on π (maximizing L for fixed λ). This dual ascent mechanism is provably convergent to a constrained optimum under mild conditions on the policy parameterization and reward structure [10]. In practice, the Lagrange multipliers act as adaptive penalty weights that grow in response to persistent QoS violations, providing a self-correcting enforcement mechanism that tightens the optimization constraint automatically when the agents' exploratory actions risk service degradation.

Inter-Agent Coordination via Message Passing

To enable coordination between geographically neighboring agents without requiring synchronous communication at inference time, DRL-ANO incorporates a lightweight message-passing module inspired by the Graph Network framework. During training, each agent A_i broadcasts a compact coordination message $m_i^t = \text{MLP}_{\text{msg}}(h_i^t)$ of dimension 32 to its graph neighbors. Received messages are aggregated as $m_i^{in,t} = \text{mean}_{j \in N(i)} m_j^t$ and



concatenated with the local embedding h_i^t before being passed to the actor network. The message-passing module learns to communicate latent coordination signals — such as anticipated load shedding, proactive interference reduction, or slice handover intentions — that improve joint decision quality. At inference time, the module can optionally be disabled for fully decentralized execution or retained for coordinated multi-cell scenarios where inter-agent communication links are available.

IV. PERFORMANCE ANALYSIS

Experimental Setup

DRL-ANO was evaluated on a simulated heterogeneous 5G network implemented in the ns-O-RAN simulation framework extended with custom DRL-ANO modules. The network comprises 19 base stations in a hexagonal layout — 7 macro cells (inter-site distance 500m) and 12 small cells (inter-site distance 150m) — deployed in an urban macro scenario following 3GPP TR 38.901 channel models at 3.5 GHz (macro) and 28 GHz (small cell mmWave). Six network slices are active concurrently: enhanced Mobile Broadband (eMBB, 2 slices), Ultra-Reliable Low-Latency Communications (URLLC, 2 slices), and massive Machine-Type Communications (mMTC, 2 slices) with distinct QoS profiles. Dynamic traffic was generated from real-world LTE base station traffic traces from a metropolitan network operator, resampled to 5G traffic volumes, with a peak network load of 94% PRB utilization during rush-hour periods. The following baseline methods were evaluated against DRL-ANO:

- Rule-Based SON (RB-SON): 3GPP-standard self-organizing network functions including Automatic Neighbor Relation (ANR), Mobility Load Balancing (MLB), and Coverage and Capacity Optimization (CCO) with vendor-default parameterization, representing the current operational standard.
- MPC-Net: A model predictive control approach using a linear time-varying network model identified online from traffic measurements, optimizing a receding-horizon QoS objective over a 10-step horizon. Represents the strongest model-based baseline.
- Single-Agent DRL (SA-DRL): A single centralized Soft Actor-Critic agent managing all base stations jointly with a fully connected state encoder (no GAT), representing the DRL baseline without multi-agent decomposition or topology awareness.
- DRL-ANO (Ours): The full proposed framework with MASAC, GAT state encoder, Lagrangian safe exploration, and inter-agent message passing.

Computational Performance

GAT state encoding for the 19-node interference graph requires 2.1ms per control epoch on the SDN controller's Intel Xeon Silver 4214 CPU. Individual MASAC actor inference at each base station requires 0.8ms on an embedded ARM Cortex-A72 CPU, enabling a 10ms distributed control cycle compatible with 5G NR slot-level scheduling granularity. The centralized critic update during online training requires 34ms per mini-batch (batch size 512) on an NVIDIA V100 GPU. Policy distribution from the SDN controller to all 19 base station agents via the O-RAN A1 interface requires 4.3ms total, completing well within the 100ms E2 interface control loop budget specified by O-RAN Alliance standards.

Table II — Performance Comparison Across Network Optimization Methods

Method	Spectral Efficiency Gain (%)	Slice Latency Reduction (%)	PRB Utilization Gain (%)	QoS Satisfaction Rate (%)	Control Latency (ms)
RB-SON	Baseline	Baseline	Baseline	88.4	N/A
MPC-Net	18.7	22.3	15.4	92.1	280
SA-DRL	27.4	31.6	21.8	93.7	24
DRL-ANO (Ours)	39.2	44.1	31.3	97.2	10

Fig 1. Per-Slice QoS Satisfaction Rate Over a 24-Hour Traffic Cycle for DRL-ANO vs. Baselines (eMBB, URLLC, and mMTC Slices)

Scheme Comparison and Discussion

Table II confirms that DRL-ANO achieves the highest performance across all evaluated metrics by substantial margins. Compared to the strongest model-based baseline MPC-Net, DRL-ANO improves spectral efficiency by 20.5 percentage points, reduces slice latency by 21.8 percentage points, and improves QoS satisfaction rate from 92.1% to 97.2%. The



spectral efficiency advantage reflects the MASAC agents' ability to learn coordinated interference mitigation strategies — such as asymmetric transmit power patterns that create interference-free corridors for URLLC traffic — that model-based MPC cannot discover due to the linearity constraint of its prediction model.

Comparing DRL-ANO against SA-DRL isolates the contribution of multi-agent decomposition with GAT state encoding: DRL-ANO improves spectral efficiency by a further 11.8 percentage points, slice latency by 12.5 percentage points, and QoS satisfaction by 3.5 percentage points over the single centralized agent. This improvement demonstrates that graph-structured multi-agent decomposition — which provides each agent with topology-aware relational context about its neighbors — produces qualitatively superior coordination strategies compared to a single agent processing a flattened global state representation. The 2.4x control latency reduction (from 24ms to 10ms) further confirms that distributed execution enables significantly faster closed-loop control than centralized decision-making.

The key advantages of DRL-ANO over existing approaches are summarized as follows:

- **Topology-aware coordination:** The GAT state encoder equips each agent with dynamic relational context about its interference neighborhood, enabling coordinated beam management and load balancing decisions that exploit the spatial structure of the network interference graph.
- **Guaranteed QoS compliance:** The Lagrangian safe exploration mechanism achieves a 97.2% QoS satisfaction rate — a 4.5 percentage point improvement over the unconstrained SA-DRL baseline — providing the reliability guarantees required for commercial slice SLA contractual commitments.
- **Scalable distributed execution:** Decentralized actor inference requiring only local state observations enables sub-10ms control decisions at each base station, compatible with 5G NR scheduling granularity and O-RAN near-real-time RIC (nRT-RIC) control loop specifications.
- **Generalization across traffic conditions:** Entropy-regularized SAC training produces robust policies that maintain high performance across the full 24-hour traffic cycle — including off-peak periods where the traffic distribution shifts significantly from the training data distribution — without requiring separate policy training per traffic regime.

Fig 2. Network-Wide Spectral Efficiency and PRB Utilization Heat Maps for DRL-ANO vs. RB-SON During Peak Traffic Hour

Simulation Experiment Details

The simulation was implemented using Python 3.10 with PyTorch 2.0 for neural network components, the Deep Graph Library (DGL) 1.1 for GAT implementation, and the ns-O-RAN 2.0 network simulator for 5G PHY/MAC layer emulation. The MASAC training used a shared replay buffer of capacity 500,000 transitions, mini-batch size 512, actor and critic learning rates of $3e-4$ and $1e-3$ respectively, discount factor $\gamma = 0.99$, and soft target network update coefficient $\tau = 0.005$. GAT encoder: $L = 4$ layers, 8 attention heads per layer, hidden dimension 128, ELU activation, dropout 0.1. Lagrange multiplier update rate: $1e-3$. The desired minimum entropy target H_{target} was set to $-\dim(A_i) = -12$, following the SAC recommendation of negative action dimensionality. Training was conducted over 2,000 episodes of 1,440 control steps each (representing 24 hours at 60-second control intervals) on 4 NVIDIA V100 GPUs in data-parallel mode, completing in approximately 18 hours. Online deployment uses 10ms control intervals via the O-RAN nRT-RIC A1 and E2 interfaces.

Ablation studies demonstrated that removing the GAT encoder (replacing with a fully connected state encoder of equal parameter count) reduced spectral efficiency gain by 10.8 percentage points, confirming that the topology-aware relational encoding — rather than raw parameter count — is the primary driver of DRL-ANO's performance advantage over SA-DRL. Removing the Lagrangian safe exploration mechanism while retaining all other components reduced QoS satisfaction rate from 97.2% to 93.9%, with particularly severe degradation in URLLC slice latency compliance during peak traffic hours, validating the essential role of constraint enforcement in commercially deployable network optimization agents.

V. CONCLUSION

This paper presented DRL-ANO, a deep reinforcement learning framework for autonomous network optimization that combines multi-agent Soft Actor-Critic with Graph Attention Network state encoding, Lagrangian safe exploration, and centralized training with decentralized execution to deliver topology-aware, QoS-guaranteed, real-time optimization of heterogeneous 5G networks. By decomposing the intractable joint optimization problem across geographically local agents equipped with relational neighborhood context, DRL-ANO achieves a 39% spectral efficiency improvement, 44% slice latency reduction, 31% resource utilization improvement, and 97.2% QoS satisfaction rate — representing the strongest reported performance on the evaluated heterogeneous 5G network configuration. The sub-10ms distributed control latency is compatible with O-RAN nRT-RIC specifications and 5G NR scheduling granularity, confirming the



framework's readiness for integration into operational network management platforms. Future research directions include extension to B5G and 6G architectures incorporating terahertz links and reconfigurable intelligent surfaces, transfer learning for rapid policy adaptation to new deployment configurations without full retraining, and formal verification of safety constraint satisfaction guarantees for safety-critical network slices supporting emergency services and industrial automation.

REFERENCES

- [1] J. G. Andrews et al., "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065-1082, Jun. 2014.
- [2] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94-100, May 2017.
- [3] S. Ayyamgari et al., "Quantum computing: Challenges and future directions," Jul. 2023, *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*. doi: 10.48175/IJARSCT-14000V.
- [4] S. Hamalainen, H. Sanneck, and C. Sartori, Eds., *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*. Chichester, UK: Wiley, 2012.
- [5] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Commun. Surv. Tuts.*, vol. 25, no. 2, pp. 1376-1411, 2023.
- [6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [7] A. S. Sribhashyam et al., "Federated AI agents for collaborative decision analytics across retail, banking, and network infrastructures," 2023, *World Journal of Advanced Research and Reviews (WJARR)*, vol. 17, no. 2, pp. 988-998. doi: 10.30574/wjarr.2023.17.2.0249.
- [8] F. Wilhelmi, B. Bellalta, C. Cano, and A. Jonsson, "Collaborative spatial reuse in wireless networks via selfish multi-armed bandits," *Ad Hoc Netw.*, vol. 88, pp. 129-141, May 2019.
- [9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learning (ICML)*, Stockholm, 2018, pp. 1861-1870.
- [10] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vancouver, 2018.
- [11] C. Thumma et al., "AgenticXAI: An explainable multi-agent architecture for autonomous enterprise intelligence," Mar. 2024, *Global Journal of Engineering and Technology Advances (GJETA)*, vol. 18, no. 3.
- [12] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, 2017, pp. 6379-6390.
- [13] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. 34th Int. Conf. Mach. Learning (ICML)*, Sydney, 2017, pp. 22-31.
- [14] N. Suresh et al., "Graph-aware hierarchical federated intelligence with adaptive differential privacy for robust non-IID spectrum sensing in next-generation wireless networks," Sep. 2024, *GSC Advanced Research and Reviews (GSCARR)*, vol. 20, no. 3.
- [15] O-RAN Alliance, "O-RAN Architecture Description v06.00," O-RAN Alliance Technical Specification, Feb. 2022.
- [16] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP Technical Report TR 38.901, v17.0.0, Mar. 2022.
- [17] M. Wang, B. Yu, Y. Zheng, H. Gan, L. Xu, and Y. Li, "Deep graph library: A graph-centric, highly performant package for graph neural networks," *arXiv preprint arXiv:1909.01315*, 2019.
- [18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [19] V. K. Kasula et al., "Agentic AI-driven CBDC: A privacy-preserving and regulatory-compliant digital payment framework," Dec. 2024, *International Conference on Applied Machine Learning and Data Analytics*, Springer Nature Switzerland, pp. 74-88. doi: 10.1007/978-3-032-18293-7_6.
- [20] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surv. Tuts.*, vol. 21, no. 4, pp. 3397-3431, 2019.
- [21] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and learning in O-RAN for data-driven NextG cellular networks," *IEEE Commun. Mag.*, vol. 59, no. 10, pp. 21-27, Oct. 2021.
- [22] R. Azmeera et al., "Enhancing blockchain communication with named data networking: A novel node model and information transmission mechanism," Mar. 2022, *Journal of Recent Trends in Computer Science and Engineering (JRTCSE)*. doi: 10.70589/JRTCSE.2022.1.4.