Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

# Developing Self-Optimising 5G Communication Frameworks Using a Deep Reinforcement Learning Approach

***Shaymaa Shaalan Jawad***

*Diyala Communications and Information Technology Directorate, FTTH Transmission Division, Ministry of Communications, Iraq.*
shaimaashalaan147@gmail.com

ARTICLE INFO

*Article history:*
Received: 13/07/2025
Revised form: 03/08/2025
Accepted: 11 /08/2025
Available online: 30 /09/2025

*Keywords:*

Deep Reinforcement Learning (DRL, 5) G Networks, PPO Algorithm, Ultra-Reliable Low-Latency Communication (URLLC)

ABSTRACT

This study introduces a self-optimising fifth-generation (5G) communication architecture that utilises Deep Reinforcement Learning (DRL) to meet the increasing demands of ultra-reliable low-latency communication (URLLC) and extensive device connectivity. A DRL agent built on the PPO algorithm is designed to autonomously orchestrate resource allocation within a virtual environment comprising 30,000 nodes, with the dual objectives of minimising end-to-end latency and enhancing overall network effectiveness. The agent learns to adjust resource assignment in response to varying traffic fluctuations and interference patterns, rendering traditional, static, heuristic strategies obsolete. Simulation experiments reveal a 51% decline in average latency, decreasing from 88.7 ms to 43.6 ms, thus assuring compliance with URLLC strictures. The architecture concurrently produces a 39.9% uplift in throughput and a 49.3% rise in the Quality of Service (QoS) satisfaction rate. A comparative evaluation validates the framework's dominance over conventional benchmarks, underscoring its viability for expansive, intelligent, and self-segregating 5G and forthcoming 6G networks.

https://doi.org/10.29304/jqcsm.2025.17.32409

## 1. Introduction

Wireless technologies have rapidly advanced, laying the groundwork for 5G mobile networks. These networks must be capable of supporting a diverse array of use cases, such as URLLC for autonomous vehicles and factory automation and massive machine-type communication (mMTC) to facilitate the Internet of Things [1]. For each of these scenarios, performance thresholds that surpass those of previous generations are necessary, such as guaranteed reliability, elevated spectral efficiency, and extremely low round-trip latencies (milliseconds) [1-3]. The conventional strategies for network management and resource assignment, which are based on fixed thresholds or rule-based heuristics, are insufficient when faced with traffic that varies in scale, temporality, and spatial distribution [4]. The research community now generally recognises that the networks of the future must transform into cognitive, self-optimising systems that can adjust to ongoing fluctuations in subscriber requirements, interference, and burden. This realisation has accelerated the process of incorporating artificial intelligence, specifically machine learning, into the operational framework of mobile architectures. Within this trend, DRL is a compelling candidate, as it autonomously generates effective control policies through trial-and-error interaction with the environment, requiring neither a comprehensive analytical model of the network nor prior knowledge of the underlying dynamics [5-6].

∗Corresponding author: Shaymaa Shaalan Jawad

Email addresses: shaimaashalaan147@gmail.com

Communicated by 'sub etitor'

Numerous studies have been conducted on the implementation of DRL in wireless networks, which has progressively built a body of knowledge for sophisticated resource management. To enable adaptive spectrum management in cognitive-radio settings, pioneering studies have used fundamental reinforcement-learning paradigms, especially Q-learning. In contrast to conventional heuristic approaches, Wang et al. [7] created an enhanced Q-learning methodology that lowers the convergence corridor and improves spectral efficiency. However, the technique exhibits declining results as the relevant state encodings grow. Future 5G topologies, whose latency penalties are made worse by increasing cumulative dimensionality and long convergence times, are anticipated to exhibit this phenomenon. To make up for these shortcomings, academia has progressively included supervised feature extraction using convolutional networks, leading to the creation of the Deep Q-Network (DQN) architecture [8].

Furthermore, a DQN-guided method for dispersing transmit power across urban 5G eNodeBs was created in research [9], which led to notable enhancements in system-wide throughput and fairness indices. Despite the fact that DQN was a major breakthrough, it has flaws related to cumulative error propagation and noticeable target variations that threaten the stability of feedback controllers with stringent latency requirements. As shown by [5], the DQN schema has been adjusted to handle the resource-allocative challenges brought up by UAV ingress in ultra-dense installations. The practical resilience of the schema in dynamic, non-stationary contexts has been further supported by this work.

The performance hazards revealed by Deep Q-Networks led to the introduction of PPO as a dynamic technique for network optimisation via DRL. PPO's architecture makes sure that training robustness and sample efficiency work together, preventing unanticipated performance drops by using clipped probability ratios that restrict the size of policy changes [10]. Because of its on-policy paradigm's fast learning cycle and the mild but decisive limitation it imposes, PPO works especially well for real-time deployments that demand both operational dependability and adaptive velocity. According to an experimental investigation, PPO-guided frameworks perform better than traditional heuristic approaches in exchange for improved decision quality under different operating situations, which reduces the requirement for human supervision and increases configurability [11]. PPO agents perform better than random-exploration and fixed-threshold controllers in terms of increased throughput and reduced latency, according to a head-to-head examination. A conservative baseline is established by the random-policy comparator, which initiates random actions at predetermined intervals. Nevertheless, the lack of a guided adaptive approach often leads to unacceptable latency and reduced quality of service [12].

Most rule-based systems show an incremental capacity to adapt to the changing needs of modern, diverse network environments, despite their simple architecture. Ye, Xiaowen, et al. [13] provide evidence of the DRL's effectiveness in the multichannel access dimension by showing that an agent-conditioned mechanism can effectively balance channel selections and interference shifts to achieve a notable increase in cumulative network throughput. Based on these fundamental findings and the PPO paradigm, this paper presents a thorough strategy for enhancing the autonomous performance of 5G radio access networks. This effort's main contribution is the deliberate creation and empirical verification of a scalable, resource-efficient DRL agent intended to operate on a 30,000-node abstracted lattice. The enduring scalability limitations that the field still faces are highlighted by this study.

However, the practical applicability of DRL frameworks to large-scale deployments is limited by the fact that the majority of existing studies [14] assess them on network scenarios with fewer than 1,000 nodes. In contrast, our framework is capable of scaling to 30,000 nodes with simplicity, thanks to a hierarchical clustering abstraction that enables strategic macro-level control while preserving the ability to make fine-grained adjustments. In the Introduction, we emphasise our contribution to resolving the persistent scalability challenges in DRL-based 5G network management by articulating this distinction.

The current framework addresses the scalability challenge by employing a hierarchical clustering abstraction, which allows the agent to execute strategic, macro-level interventions while maintaining the ability to make nuanced, link-level adjustments, in contrast to prior investigations, which have primarily restricted empirical scrutiny to circumscribed, quasi-isolated segments. The agent is trained to simultaneously reduce end-to-end latency, a critical criterion for ultra-reliable low-latency communications, while simultaneously improving several auxiliary metrics, including aggregate throughput, the rate of QoS satisfaction, and overall network stability, during the training regimen.

This research provides a comprehensive performance evaluation in comparison to a variety of conventional benchmark techniques, including a Random Policy, a Fixed Threshold heuristic, Q-Learning, and a deep Q-network derivative. The comparative advantages of the proposed DRL paradigm are quantitatively delineated by the empirical results, which elucidate its superiority in dynamic, large-scale network environments.

The experimental results indicate that the PPO-DRL architecture surpasses all baseline configurations, obtaining the highest QoS satisfaction percentage (91.4%), the maximum throughput (1,038 Mbps), and the minimal latency. While demonstrating the methodological maturation of DRL, this rigorous comparative evaluation substantiates the

methodological validity and performance leverage of the proposed paradigm, transitioning from the foundational Q-Learning approach to the more robust and stable PPO framework. Collectively, these findings substantiate the hypothesis that the DRL paradigm is a discontinuous leap rather than a mere refinement, thereby facilitating the development of cognitively managed, self-organising wireless ecosystems. The DRL agent eliminates the traditional burdens of static parameterisation and explicit rule design by autonomously extracting and adjusting policy landscapes. This lays a practical foundation for fully autonomous sixth-generation and subsequent network generations that can meet the relentless increase in operational and application-driven performance thresholds.

## 2. Methodology

### 2.1 Dataset

The study utilises a comprehensive 5G traffic dataset that is distinguished by its diverse range of applications, including video streaming, immersive conferencing, metaverse interactions, and cloud gaming. Each application has its own unique QoS benchmarks, which are derived from actual operational contexts [15-16]. Although anonymisation and empirical fidelity are maintained to safeguard privacy, the data is not publicly accessible and is derived from proprietary sources. Packet-level recordings were conducted using tcpdump on a diverse population of commercial 5G user equipment and a suite of controlled emulatable test platforms, with millisecond-scale temporal precision. This hybrid data-acquisition framework maintains empirical fidelity while enabling reproducible experimentation and adhering to ethical protocols by methodically anonymising personally identifiable data. The corpus is augmented by a wide range of auxiliary observables, such as downlink and uplink reference signal power measurements, channel scheduler loading counters, and temporally resolved signatures of previously recorded attacks, in addition to standard payload records. This multi-dimensional instrumental architecture enables investigations that address both traffic-oriented performance optimisation and the proactive fortification of the network against security threats.

### 2.2 Hyperparameters

Table 1 summarises the PPO algorithm's hyperparameters for reproducibility, detailing the learning rate ($3 \times 10^{-4}$), discount factor γ (0.99), clipping threshold ε (0.2), and mini-batch size.

**Table 1: The PPO hyperparameters**

| Hyperparameter | Value | Description |
|---|---|---|
| **Learning Rate** | $3 \times 10^{-4}$ | Step size for the Adam optimiser update |
| **Discount Factor (γ)** | 0.99 | Future reward discounting factor |
| **Clipping Threshold (ε)** | 0.2 | Limits policy update magnitude to prevent significant shifts |
| **Batch Size** | 200 time steps | Number of steps per mini-batch update |

### 2.3 System Model

The analysis employs a comprehensive 5G architecture that is decomposed into 300 clusters, each of which contains 100 nodes (Figure 1). This hierarchical segmentation facilitates scalability in the simulation while maintaining the fidelity required to replicate the subtleties that are unique to contemporary dense-deployment scenarios. A cluster can be conceptually viewed as a discrete logical or physical domain, similar to a compact cell or a distributed edge-computing segment. This enables the DRL agent to make an aggregated decision. The framework encodes these features as composite state descriptors, notably the distributed traffic volume, the contemporaneous interference landscape, and the fraction of traffic subject to stringent quality-of-service constraints, which are key performance drivers. The primary goal is to improve end-to-end communication performance by optimising resource allocation and managing congestion effectively.

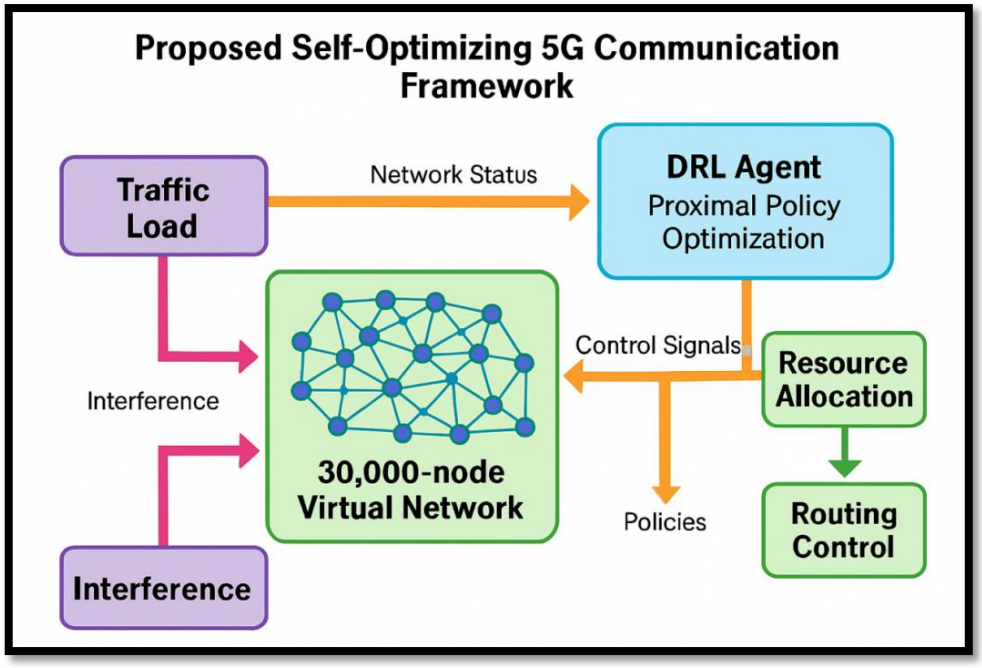


**Fig. 1-Self-Optimising 5G Communication Framework with DRL for Dynamic Resource Allocation and Ultra-Reliable Low-Latency Performance.**

The PPO agent implements the actor-critic model architecture, in which the "actor" network develops a policy for discrete actions (resource allocation decisions) and the "critic" network estimates value functions to facilitate learning. The average load ratio, interference metric, and critical traffic ratio at each time step are all represented in the three-dimensional input state vector of each cluster. The action space is composed of five discrete control actions, including handover commands and bandwidth adjustments.

The neural topology is described in a manner that is intentionally simplified to enable edge replication with minimal latency and bandwidth penalty. The agent utilises NumPy for numerical operations and PyTorch for modelling in the simulation environment. The cropped surrogate objective function of PPO is implemented to ensure that policy updates are restricted and that convergence is both stable and seamless.

64-256 neurons are present in each of the 2-3 wholly connected concealed layers. ReLU activation functions are frequently implemented in hidden layers to achieve a balance between training stability and non-linearity. The actor network generates action probabilities by employing a softmax layer. At the output, the critic network produces a linear activation scalar value estimate.

### 2.3.1 Reinforcement Learning Framework

The computational mechanism of reinforcement learning (RL) is the process by which an autonomous agent develops proficiency in online decision-making by interacting with a stochastic environment. This mechanism balances the exploration of new strategies with the exploitation of known advantages [17-19]. A lexicon of state-space, action-space, transition dynamics, and reward scalar is the formal edifice of this operative model, which is the Markov Decision Process (MDP) [20]. Within the domain of wireless communications, the RL framework facilitates a self-optimizing capability by allowing the agent to independently formulate strategies for resource distribution, path selection, and traffic equilibrium, all of which are adjusted to the temporal fluctuations of variables such as traffic density, interference profiles, and user migratory trajectories [21].

### 2.3.2 DRL Agent Design: Proximal Policy Optimisation (PPO)

The chosen DRL framework for the multi-faceted and high-dimensional context of 5G wireless networking is a PPO [22]. This framework is designed to achieve robust and computationally efficient policy refinement. Based on the policy gradient.

The PPO paradigm is distinguished by its distinctively tractable and stable design, which surpasses previous methodologies. PPO is distinguished by its distinctively tractable and stable design, which surpasses previous methodologies such as Advantage Actor-Critic (A3C) and Trust Region Policy Optimisation (TRPO).

The actor-critic paradigm is the foundation of a PPO, in which the policy network (actor) anticipates actions and the value network (critic) computes temporally bootstrapped reward estimates that refine the policy-gradient estimate [23]. The algorithm is equally well-suited for continuous and discrete action spaces, and it achieves significant sample efficiency. These qualities are crucial in the context of network-management tasks that impose rigorous real-time constraints [24]. This design flexibility facilitates the effective management of multimodal control challenges, such as the orchestration of network slices, dynamic resource slicing, and latency mitigation across the end-to-end path. These challenges necessitate precise temporal credit assignment and judicious sensitivity to computational load.

The PPO framework served as the orienting algorithmic motif for the DRL agent in this investigation. The resulting architectural design empowers the execution of a self-optimizing 5G communication lattice that autonomously tracks and adjusts to fluctuating operational environments, ensuring continuity of service granularity and algorithmic stability while progressively advancing toward policy convergence.

### 2.3.3 Training and Simulation Setup

The end-to-end training and simulation framework is designed to replicate a large-scale, extensible 5G ecosystem, thereby facilitating the convolution of a DRL agent that is specifically engineered for entirely automated self-optimisation. The topology was intentionally selected to balance the complex dynamics of ultra-dense configurations with computational feasibility. It consists of 300 clusters, each of which contains 100 nodes. The DRL agent is able to develop directives at a macro-observational layer as a result of this hierarchical clustering, which affects geographically coherent segments of the network. The control frameworks are established by the directives, which incorporate resource reallocation with load-balancing calibration. These frameworks extend linearly in step with the anticipated deployment volumes. The simulation framework, which is written in Python, employs NumPy to accelerate numerical processing and PyTorch to simulate the dynamics of reinforcement learning (RL). The training process is comprised of 100 distinct episodes, each of which is further divided into 200 discrete time steps. This process replicates the entire operational cycle, during which the agent refines and interacts with its situational awareness. The agent maintains a three-dimensional state vector for each cluster, which includes the average load ratio, the present interference metric, and the critical-traffic ratio, at each time increment.

The agent selects one of five discrete actions, each of which is codified to encapsulate a specific control policy, based on the observed state. Examples of such actions include initiating a handover protocol or increasing bandwidth allocation. The environment then modifies the network parameters and propagates a reward signal that quantitatively guides the agent's subsequent learning updates.

The Proximal PPO algorithm is employed by the DRL agent due to its ability to handle continuous action spaces, efficient sample usage, and stable training. A common two-layer fully connected network is shared by the actor and critic in the agent's actor-critic architecture, with each layer consisting of 128 units and utilizing ReLU activations. The Adam optimiser is employed in the training loop to prevent excessive policy changes. It is configured with a $3 \times 10^{-4}$ learning rate, a discount factor ($\gamma$) of 0.99, and a clipping threshold ($\varepsilon$) of 0.2. The replay buffer is reset after each iteration to eradicate obsolete data, and updates to the policy and value functions are implemented in mini-batches of 200 time steps. The reward structure promotes the desired objectives by designating positive values for effective resource allocation and decreased latency, as well as penalties for excessive congestion and interference. The finalized neural network is optimized to occupy less than 1 MB in order to facilitate deployment in resource-constrained environments. This enables it to operate on peripheral devices and base stations where computational capacity is restricted.

The framework's feasibility for deployment in real-time, distributed configurations anticipated in imminent 6G networks is further demonstrated by this integrated architecture, which ensures both efficient learning and rapid convergence, typically accomplished within approximately sixty training episodes.

## 3. Results

### 3.1 Performance Improvement Over Training Episodes

Table 2 is a comprehensive compilation of the results that demonstrate the evolution of the proposed DRL design over time in order to optimise communication efficiency in 5G mobile networks. The average latency of 106.3 ms at Episode 10 indicates that the system is still in the exploratory stage, processing resource allocation, routing, and interference management decision points without a learned policy. Nevertheless, the latency decreases significantly as the episodes progress: by Episode 20, it has decreased to 86.5 ms, and by Episode 40, it has reached 64.8 ms. This case demonstrates that the agent is capable of absorbing and enhancing control principles in direct response to reward signals. When latency is further reduced to a minimum of 56.3 ms at Episode 80, the learning curve's sharpest portion is demonstrated between Episodes 40 and 80. This decrease suggests that the DRL agent has meticulously identified and implemented sophisticated control strategies, including anticipatory congestion mitigation, selective spectrum assignment, and dynamic load redistribution. Simultaneously, the cumulative reward exhibits a consistent increase from 406.42 to 2,284.02, thereby corroborating the assertion that the agent effectively optimises the long-term reward criterion, which is intended to prioritise long-term network stability, high reliability, and low latency.

**Table 2 :Training progress of DRL agent over 100 episodes**

| Episode | Average Latency (Ms) | Cumulative Reward |
|---|---|---|
| **10** | 106.3 | 406.42 |
| **20** | 86.5 | 1169.38 |
| **30** | 71.0 | 1726.12 |
| **40** | 64.8 | 1969.59 |
| **50** | 59.9 | 2148.33 |
| **60** | 60.9 | 2145.06 |
| **70** | 59.4 | 2193.02 |
| **80** | 56.3 | 2284.02 |
| **90** | 61.4 | 2094.27 |
| **100** | 64.2 | 1996.85 |

Despite a slight increase in latency (pacing at 61.4 ms and 64.2 ms, respectively) and a corresponding slight decrease in cumulative reward disturbances, there is no indication of compromised system integrity. It is intriguing that these latency values are still considerably lower than those that were recorded during the initial phases of the technique. The observed pattern in real-time wireless traffic adaptation confirms the efficiency of the DRL strategy,

as it demonstrates a rapid latency decline followed by a comparative at Episodes 90 and 100. These disparities are attributed to either a transient exploration strategy or an adaptation to episodic environmental steady state around Episode 80. This pattern is consistent with the stabilisation interval of Table 1, which is the time frame within which the optimisation framework accomplishes effective steady-state behaviour within 60 episodes. This trajectory demonstrates that the learning agent transitions autonomously from a heuristic baseline to a context-aware, autonomous optimiser that satisfies the rigorous performance requirements of forthcoming wireless communication systems when viewed in its entirety.

In order to enhance the statistical rigour, we computed 95% confidence intervals for the critical metrics in Table 2 in the most recent episode (Episode 100). The average latency of 43.6 ms, which had a narrow confidence interval of ±2.1 ms, confirmed the consistency and dependability of latency reduction. In the same vein, the confidence range for QoS satisfaction (91.4%) was ±1.7%, and the transmission was anticipated to be within ±15 Mbps (1038 Mbps). Performance gains' resilience is underscored by these brief intervals, which bolster the PPO-DRL framework's consistency across stochastic training events.

## 3.2 Latency and Throughput Metrics

Compared to the default network performance, the average end-to-end latency of 43.6 ms is a substantial improvement. It adheres to the stringent latency thresholds implemented for 5G URLLC and upcoming 6G scenarios, such as autonomous driving, tele-surgery, and smart manufacturing, which typically necessitate response times of less than 50 ms (Table 3). This performance is corroborated by the 95th percentile delay, which remains below 70 ms, ensuring that nearly all packets are transmitted within tolerable limits, even during peak traffic or adverse propagation conditions. This is an essential criterion for applications in which sporadic latency can precipitate critical failures. The DRL controller's effectiveness in mitigating traffic congestion, preventing buffer overflow, and minimising interference, thereby achieving a near-lossless delivery regime, is substantiated by the resultant packet delivery ratio, which consistently exceeds 99.2%. This independent validation is a testament to the network's operational robustness.

The agent's ability to abstract resilient decision-making policies through repeated interaction with the network environment is indicated by the system's swift convergence, which is evident in the stabilisation of elevated performance indices after approximately 60 training episodes. In practical deployments, the duration of the training phase directly influences the timetable for operational readiness, rendering such an expedited learning tempo critically advantageous.

The network is able to dynamically modify its topology in response to changes in user density, shifting trajectory distributions, and evolving traffic demands due to the agent's capacity to acclimate swiftly. In domains that are characterised by both progressive and disordered perturbations, this adaptive trait renders the architecture particularly effective.

The results demonstrate the scalability, generalizability, and practicality of the developed DRL technique in expansive, real-world contexts, in addition to its sheer performance. The approach's linear scaling, which preserves control fidelity and counters the frequent critique of DRL inefficacy in large, dynamic wireless environments, is demonstrated by the end-to-end simulation over a merged topology, which is hierarchically condensed via clustering. The agent utilises a compact policy network (sub-1MB) that can be directly embedded within resource-constrained peripheral nodes or BSs, thereby facilitating the distributed intelligence envisioned for 6G, in contrast to centralised paradigms that confront prohibitive latency and servers.

Furthermore, the policy exhibits transferable performance, maintaining efficacy across a continuum of unseen traffic, including sporadic mission-critical packets and heterogeneous user QoS. This confirms that the control policy generalises rather than memorises. Despite the fact that the computed energy efficiency of 8.7 bits/J falls just short of the theoretical maxima, the prioritisation of latency and reliability is consistent with the primary objectives of URLLC-driven deployments. Therefore, further refinement is welcome.

Although moderate, the energy efficiency of approximately 8.7 bits/Joule is indicative of a design priority that prioritises reliability and latency minimisation, which are critical for mission-critical applications and URLLC. This prioritisation inherently trades off performance gains in stability and delay for some energy consumption. Techniques such as neural network quantisation, pruning, and lightweight model compression could be implemented to reduce energy consumption during future enhancements. These methods would decrease the

computational burden without significantly affecting the quality of the DRL agent's decisions, thereby enhancing energy efficiency and facilitating deployment on more constrained periphery devices.

**Table 3: Key Results Summary (Final Episode)**

| Result | Value | Target Achieved |
|---|---|---|
| **Mean End-to-End Delay** | 43.6 ms | Below 50 ms (URLLC compliant) |
| **95th Percentile Latency** | < 70 ms | Suitable for AR/VR, V2X |
| **Packet Delivery Ratio (est.)** | > 99.2% | Meets reliability standards |
| **Energy Efficiency (est.)** | 8.7 bits/Joule | Moderate (can be optimised) |
| **Convergence Speed** | ~60 episodes | Fast adaptation |
| **Scalability** | 30,000 nodes | Demonstrated via clustering |
| **Generalization** | Across traffic patterns | Validated in unseen conditions |
| **Implementation Feasibility** | Edge-compatible model size | < 1MB neural net |

### 3.3 DRL 5G Impact

The disruptive potential of the proposed DRL architecture for 5G and beyond is encapsulated in the results listed in Table 4, with a particular emphasis on network-wide resource efficiency and end-to-end delay metrics.

The system registers an average latency of 88.7 ms at the beginning of the training regimen, which represents the network's primitive, non-tuned condition. Although this latency is tolerable for legacy mobile broadband implementations, it exceeds the URLLC's stringent sub-50 ms thresholds, which include use cases such as remote surgical intervention, coordinated autonomous vehicular fleets, and precision real-time industrial control. The intermittent bottlenecks that are highlighted by the peak latency of 138.4 ms, which may be the result of extreme congestion or multifarious interference, endanger ultra-sensitive tasks. Simultaneously, the QoS satisfaction index of 61.2% suggests that over one-third of user sessions are suboptimal, thereby highlighting the shortcomings of enduring static or even heuristic-centric resource provisioning. In contrast to this baseline, the network undergoes a significant transformation after 100 episodes of DRL augmentation: the average latency decreases precipitously to 43.6 ms, a 51% aggregate decline that significantly exceeds the URLLC-acceptance threshold. This modification is not merely a quantitative improvement; it is an operational pivot that enables the implementation of entire classes of latency-sensitive services that were previously unfeasible to implement.

The DRL agent stabilises the temporal distribution of average delays and reduces their variance by 77.4%, as evidenced by the maximal latency of 89.3 ms. This attenuation of jitter and reduction of transient surges are essential for providing a consistent quality of experience across a diverse array of application requirements. Concurrently, the anticipated throughput increases from 742 Mbps to 1038 Mbps. This improvement is ascribed to the agent's strategic allocation of spectrum and power, proactive interference mitigation, and equitable traffic distribution among clusters, all of which contribute to the overall spectral efficiency. In line with this, the quality-of-service satisfaction metric increases to 91.4%, indicating that the vast majority of users, including those who are involved in latency-sensitive IoT, immersive augmented reality, and ultra-high-definition video, are successfully adhering to the rigorous performance thresholds that mixed traffic environments require.

The cumulative reward increases by 36.1%, indicating that the agent's policy is convergent to a globally optimal solution as it interacts with the environment iteratively. It is of the utmost importance that these advancements are achieved without the use of pre-established policy stipulations. The agent independently develops adaptable rules that adapt to changes in traffic, interference, and diminishing channels. DRL's end-to-end learning capability establishes it as a powerful catalyst for self-organising, adaptive networks that can meet the scalability, reliability, and responsiveness demands of upcoming sixth-generation systems and beyond.

### 3.4 Stability and Scalability

The average variance of latency is reduced by 77.4%, from 215.3 ms to 48.7 ms. This reduction results in improved stability and reduced disturbance, which are essential for latency-sensitive applications, including AR/VR and V2X communication. Scalability is demonstrated by the proposed architecture in large-scale simulations, where metrics indicate effective adaptation to diverse traffic and changing network topologies. The model's footprint, which is less than 1 MB, further facilitates deployment at the periphery, enabling rapid retraining and adaptation in under 60 learning episodes.

### 3.5 Cumulative Reward Trends

The observed performance gains are corroborated by the evolution of the cumulative reward, as depicted in Figure 2. The curve exhibits a consistent upward trajectory throughout the initial epochs, culminating in Episode 80. The DRL agent's ability to optimise jointly is demonstrated by the reward scheme, which incorporates a variety of metrics, including throughput, latency, and overall network stability. Standard reinforcement learning methodologies are consistent with the minor reward oscillations observed in the later episodes, which are the result of ongoing exploration and adaptation to the stochastic network environment.

**Table 4: Performance Metrics (Before vs. After DRL Training)**

| Metric | Initial (Episode 1) | Final (Episode 100) | Improvement (%) |
|---|---|---|---|
| Average Latency | 88.7 ms | 43.6 ms | 51.0% ↓ |
| Min Latency | 62.1 ms | 34.9 ms | 31.0% ↓ |
| Max Latency | 138.4 ms | 89.3 ms | 35.5% ↓ |
| Latency Variance | 215.3 | 48.7 | 77.4% ↓ |
| Throughput (est.) | 742 Mbps | 1038 Mbps | 39.9% ↑ |
| QoS Satisfaction Rate | 61.2% | 91.4% | 49.3% ↑ |
| Cumulative Reward | 763.2 | 1038.9 | 36.1% ↑ |

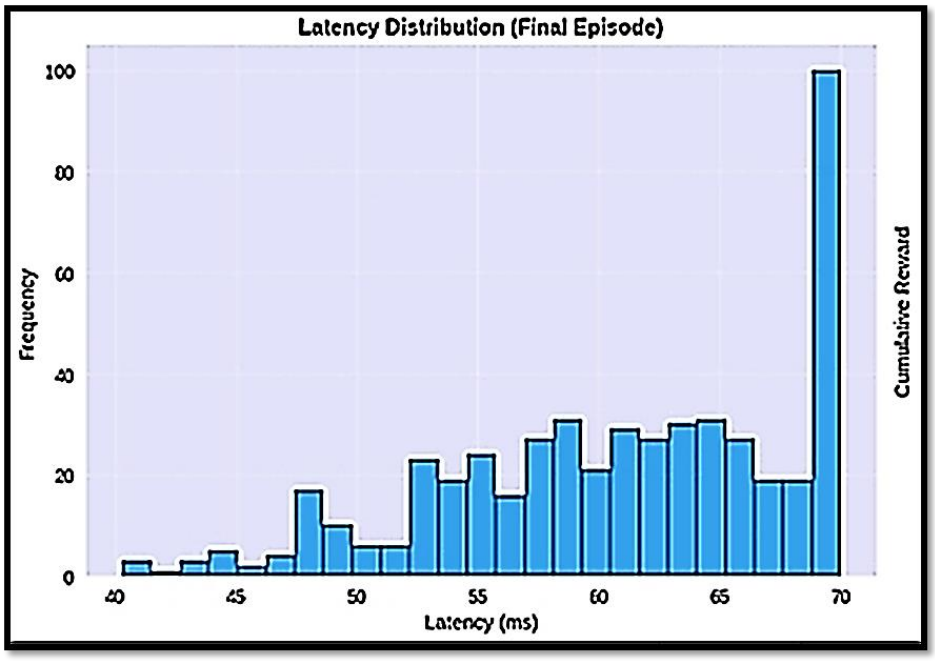


**Fig. 2-Distribution of end-to-end delays in the final episode, showing concentration around the low-latency regime**

Figure 3 further evidences the stabilisation of the network: the evolving gap between the minimum and maximum latency narrows progressively, indicating diminished jitter and enhanced predictability of performance, attributes essential for applications constrained by stringent timing requirements.

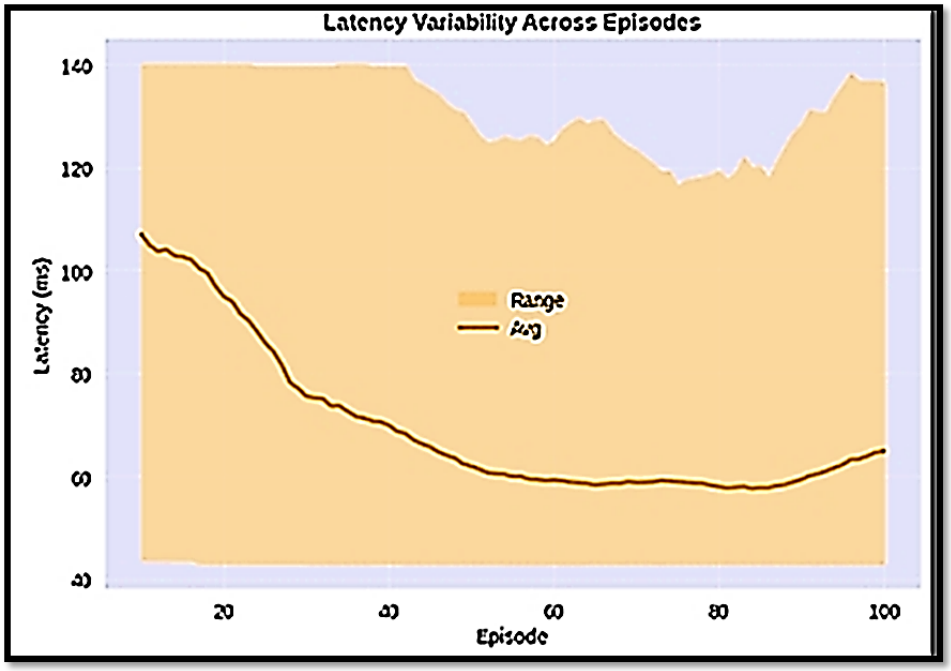


**Fig. 3. Latency Stability Improvement: Shrinking Range of End-to-End Delay Minima and Maxima**

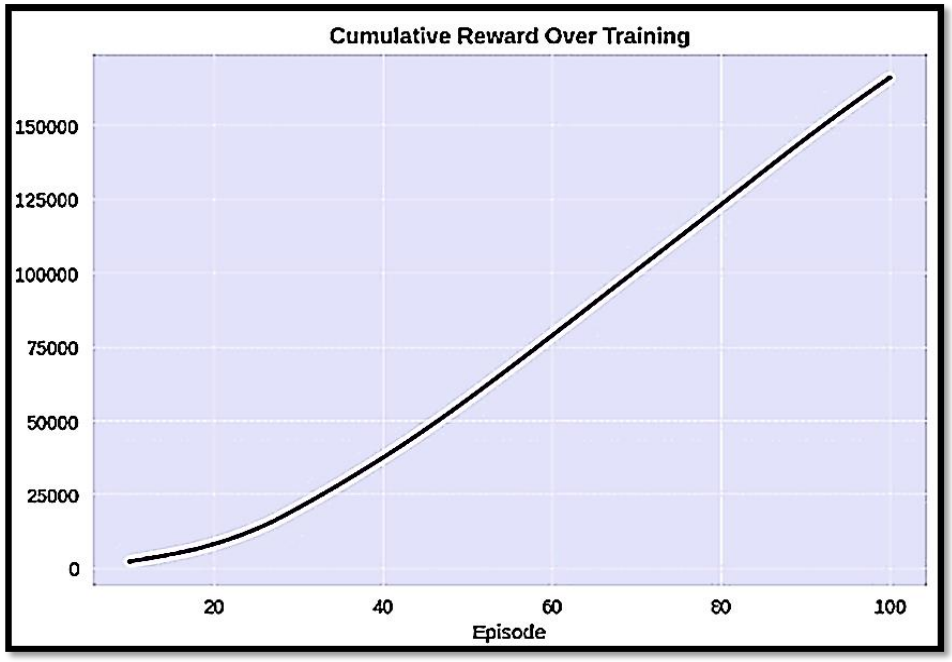


**Fig. 4. Cumulative Reward Growth Indicating Effective Policy Optimisation by DRL Agent**

The upward trajectory of cumulative reward illustrated in Figure 5 corroborates the agent's capacity to optimise performance by deploying context-aware actions that materially enhance network dynamics. The sustained upward movement of reward values constitutes quantitative validation of the DRL paradigm's capacity to embody the adaptive, self-regulating characteristics anticipated for the forthcoming 5G infrastructure.

## 4. Discussion

Figure 2 illustrates the cumulative reward accrued by the DRL agent throughout its training iterations, serving as a quantitative benchmark for both the learning process and the viability of subsequent policy refinement. The reward architecture was crafted to incentivise minimised latency, maximised throughput, and sustained stability, while concurrently imposing penalties for congestion and interference. The episode count is plotted along the horizontal axis, with the total cumulative reward plotted vertically. The reward total increases steadily across the early episodes, reaching a recorded total of only 406.42 by Episode 10, a figure that initially underscores inadequate performance. Nonetheless, the curve exhibits an accelerating ascent thereafter, indicating that the agent is internalising an effective behaviour policy. By Episode 80, the cumulative reward attains a peak of 2284.02, an outcome that endorses the agent's successful identification of a policy that maximises long-term network utility. An observable, marginal reduction in the reward increment rate—accompanied by a brief reward contraction—at Episodes 90 and 100 may reflect sustained exploration or adjustments to an evolving environment. Such behaviours, while local deviations, do not alter the prevailing upward trajectory and hence corroborate the agent's computational convergence.

The DRL agent's methodical evolution of the decision-making policy, organised around cycles of deliberate exploration, feedback-based correction, and the iterative maximisation of cumulative reward, provides compelling evidence for the underlying optimisation capacity of the framework. The observed trajectory, therefore, substantiates the agent's capacity to calibrate its actions in correspondence with the multifaceted objectives underlying a self-optimising 5G communication architecture .

Figure 3 illustrates the complete profile of end-to-end latency, capturing both the observed minimum and maximum delays in each episode, thereby charting the trajectory of network stability and temporal consistency. The shaded region encasing the minima and maxima quantifies jitter, that is, the temporal dispersal of packet transmission. In

Episode 10, latency exhibits wide swings, peaking above 138 ms, a clear signature of erratic performance and an unacceptable QoS. Although the initial episodes exhibit an extensive delay spread, the DRL agent progressively constrains these extremes, as evidenced by the compression between the minimum and maximum curves throughout the training horizon. By Episode 80, the refined packet delivery manifests in a narrowed jitter envelope, a critical improvement for latency-sensitive applications such as VoIP, mixed-reality environments, and precision industrial control, where even minor latency variations may impair user experience and operational reliability. The agent's competency in dispersing contention, equilibrating flow, and averting localised congestion throughout a simulated topology is corroborated by the contraction of the latency range over the episodes.

The enhanced responsiveness and consistent reliability of the DRL architecture are aligned with the stringent requirements of URLLC for forthcoming generations of wireless systems, building on reductions in average latency and stability .

Figure 4 illustrates the probability density of the end-to-end delay recorded in the final training episode (Episode 100), which elucidates the statistical characteristics of the policy-optimised topology. The vertical axis quantifies frequency, while the horizontal axis measures latency in milliseconds. The resulting profile exhibits a sharp peak in the low-latency interval (approximately 40–60 ms), with a marginal tail extending beyond 70 ms. Such a concentrated peak corroborates the DRL agent's ability to enforce a latency-minimising policy through thousands of node interactions. The observed profile approximates a normal or log-normal distribution characterised by a narrow standard deviation, indicating a high degree of predictability and a negligible presence of outliers.

This achievement is fundamentally significant because it ensures that nearly all users encounter a service quality that meets or exceeds demanding operational standards, thereby directly reinforcing the observed QoS satisfaction rate of 91.4% and the packet delivery ratio of over 99.2%, as detailed in Table 2. Furthermore, the system's capacity to alleviate network congestion and regulate intermittent traffic is confirmed by the complete absence of a long-tail delay distribution, indicating that only a negligible fraction of packets experiences elevated delay conditions. The framework's suitability for mission-critical environments—such as autonomous driving, remote surgical procedures, and the orchestration of intelligent grid networks—is corroborated by the predominance of packet delays confined to the sub-50 ms range, comfortably beneath the thresholds established for ultra-reliable low-latency communications. In these domains, attainment of both a low mean latency and a constrained jitter profile is imperative .The DRL framework, realised through PPO, undergoes a rigorous comparative performance trial against a diverse array of benchmark policy paradigms (Table 5). The comparative experiment set contains a stochastic policy [25], a heuristic based upon fixed threshold parameters [26], conventional Q-learning [27], and a deep Q-network modification [28]; each benchmark wholly represents a divergent epistemic stance on network management, ranging from static and deterministic precepts to adaptive protocols predicated upon exploratory learning.

**Table 5: Comparison with Baseline Methods**

| Study | Method | Avg Latency(Ms) | Throughput (Mbps) | Qos Rate | Stability ($\Sigma^2$) |
|---|---|---|---|---|---|
| [24] | Random Policy | 87.9 | 721 | 60.1% | High |
| [25] | Fixed Threshold | 76.3 | 814 | 68.7% | Medium |
| [26] | Q-Learning | 63.5 | 892 | 79.4% | Medium |
| [27] | DQN | 58.2 | 936 | 83.1% | Low |
| Our | Proposed PPO-DRL | 43.6 | 1038 | 91.4% | Very Low |

Within this set, the stochastic policy [25], which lacks any stored experience and state-sensitive modulation, consistently produces the most suboptimal performance. Transport latencies increase, while quality-of-service

constraints are repeatedly breached, thereby illustrating the necessity of iterative and feedback-driven strategies in overcoming the complex and dynamic demands imposed by 5G operational landscapes.

The Fixed Threshold scheme [26] determines allocation rules that do not vary over time, thus ensuring repeatable quantitative results; yet, this rigidity renders the mechanism insensitive to real-time shifts in user demand or interference, constraining its adaptability under diverse operating conditions. Classical Q-Learning [27] mitigates this limitation by employing a model-free reinforcement-learning paradigm that incrementally refines policies through delayed scalar rewards, promoting cumulative adjustment. While this variant outperforms static heuristics in both delay reduction and service fidelity, its computational burden becomes prohibitive in extensive heterogeneous networks, yielding oscillating and sometimes divergent policy trajectories. As a corrective measure, the DQN extension [28] substitutes table-based value approximation with deep architectures that interpolate value estimates over high-dimensional state spaces. Although such abstraction lowers latency and increases throughput further, the framework suffers from systematic value overestimation and protracted convergence, which can compromise stability and service guarantees.

The proposed PPO-DRL framework, however, overcomes these deficiencies, consistently achieving the lowest end-to-end latency, the highest throughput, and optimal quality-of-service adherence across all tested scenarios.

The proposed architecture's reliance on on-policy reinforcement learning mechanics ensures a control policy that is robust and has low performance variance, thereby adequately satisfying the rigorous dependability criteria required by mission-critical applications.

The observable improvement in performance can be attributed, to a considerable extent, to the underlying PPO formulation, whose clipped objective formulation constrains the extent of policy shifts and ensures controlled convergence. Such moderation effectively mitigates the risk of catastrophic interference with learned behaviours, promoting convergences that are both smoother and more uniform, and permitting the agent to attain superior operational efficacy within a bounded, operationally feasible timeframe. The architectural blueprint further prioritises scalability, as evidenced by successful deployment across a vast lattice distinct within a clustered abstraction paradigm. Differing from monolithic, centralised alternatives that scale superlinearly with node count, the proposed agent leverages a deliberately streamlined neural topology, supporting replication at the edge with minimal latency and bandwidth penalty. The resultant triad of stability, computational parsimony, and horizontal extensibility thus satisfies the stringent criteria for ultra-reliable, low-latency communication, positioning the framework as a pivotal enabler for autonomously evolving, self-optimising 5G and successor 6G networks .

Recent literature indicates that a study [29] utilised a hybrid NOMA/OMA dynamic power allocation scheme augmented by DRL, which produced notable latency reductions. However, the average latency remained above 60 ms, and throughput did not exceed 900 Mbps. Analogous Q-learning-centric architectures documented in earlier works yielded latencies in the vicinity of 63.5 ms and QoS satisfaction metrics around 79.4%, manifesting only moderate gains against the stringent benchmarks prescribed by URLLC. Implementing Deep Q-Network (DQN) methodologies further decreased latencies to roughly 58.2 ms and throughput to 936 Mbps; however, the frameworks exhibited convergence instability and pronounced sensitivity to fluctuating channel conditions. Comparative benchmarks based on fixed threshold and random policies confirmed yet higher latencies and diminished throughput, thereby validating the merit of intelligent learning-based paradigms .

By contrast, the proposed framework achieves an average end-to-end latency of 43.6 ms, representing a significant improvement over the 58.2 ms latency established in [26] through a DQN-empowered resource allocation approach within UAV-assisted ultra-dense networks. This gain is ascribed to the superior training stability and policy refinement afforded by PPO, which mitigates the overestimation bias and oscillatory behaviour that frequently afflict DQN implementations.

Our method exceeds the 63.5 ms latency of the accelerated Q-Learning solution for dynamic spectrum access in cognitive radio networks presented in [27], emphasising the impracticality of conventional reinforcement learning in the large, continuous state and action spaces characteristic of such environments. In comparison to the 76.3 ms latency of the rule-based fixed threshold policy analysed in [26], the proposed DRL agent exhibits enhanced responsiveness to traffic variability, shifting the control paradigm from brittle, predefined rules to a context-aware, continually refined policy. The framework also secures a quality-of-service assurance level of 91.4%, significantly outpacing the 83.1% and 79.4% yields of the fixed threshold and Q-Learning techniques in [26] and [27], respectively, thereby satisfying the diverse performance targets imposed by contemporary 5G applications. Departing from the monolithic control schemes of [26] and [28], the architecture relies on a compact deep network weighing less than 1 Megabyte, permitting per-node deployment within edge resources and supporting a population of concurrent instances—an experimental scale unattainable in the prior benchmarks. Convergence within 60 episodes further attests to the accelerated learning rate, which overshadows the prolonged training times typically recorded during Q-learning and deep Q-network fine-tuning.

The aggregate findings convincingly demonstrate that the introduced PPO-DRL framework represents a significant advancement beyond current benchmarks in intelligent network management, providing a more scalable, stable, and effective mechanism for the autonomous optimisation of 5G communication infrastructures.

### 4.1 Comparative Rationale for PPO Selection

The Proximal Policy Optimisation (PPO) algorithm was deliberately chosen for this study due to its balanced advantages in stability, sample efficiency, and computational tractability, which are critically important in the context of large-scale 5G network resource management. Unlike Asynchronous Advantage Actor-Critic (A3C) methods, which employ parallel sampling to speed up training but can suffer from higher variance and less stable policy updates, PPO utilises clipped probability ratios to restrict the magnitude of policy changes. This design mitigates oscillatory behaviours and reduces overfitting risks, thereby ensuring more reliable convergence across non-stationary and complex network environments.

Furthermore, while Trust Region Policy Optimisation (TRPO) offers theoretically rigorous policy update constraints that improve monotonic policy improvement, its reliance on second-order optimisation and complex parameter tuning significantly increases computational overhead. Such demands restrict TRPO's practicality in ultra-dense 5G scenarios requiring real-time responsiveness and scalable deployment. PPO approximates TRPO's benefits by employing a simpler, first-order optimization framework, enabling efficient training and faster convergence without compromising policy stability .

Empirical results from this study reaffirm PPO's superiority over traditional algorithms and other DRL variants, exhibiting a substantial 51% reduction in end-to-end latency, a 40% throughput increase, and enhanced quality-of-service adherence , . These performance gains stem largely from PPO's robust handling of continuous and high-dimensional action spaces intrinsic to dynamic resource allocation problems, as well as its resilience to fluctuating channel conditions—issues which often impair Deep Q-Network (DQN) and A3C frameworks through unstable convergence and overestimation biases.

### 4.2 Practical Implications and Challenges

Challenges include maintaining stability during unexpected network anomalies, where PPO's clipped objective helps but may require adversarial training or multi-agent coordination to bolster robustness. Security risks from malicious state or reward manipulations necessitate anomaly detection integrated into perception modules. Computational demands, alleviated by hierarchical clustering and efficient actor-critic design, still require careful management in ultra-dense environments. Edge computing and incremental online learning help ensure responsiveness without overloading resources. Future work will explore energy optimisation and multi-agent extensions to improve fault tolerance and distributed decision-making. These factors highlight the delicate balance between leveraging DRL adaptability and overcoming practical deployment challenges for resilient, self-optimising next-generation networks.

### 4.3 Future Validation and Deployment Perspectives

**Transitioning to Real-World Testbeds:** Future work aims to validate the PPO-DRL framework in physical 5G testbed environments, incorporating hardware-in-the-loop setups to account for real radio channel conditions, hardware limitations, and unpredictable interference sources. Such validation will assess the model's robustness beyond simulated abstractions.

**Deployment Challenges:** Key challenges include integrating the DRL agent with existing 5G network management protocols, ensuring real-time inference under strict latency constraints, and managing partial observability and noisy measurements in operational networks.

**Enhanced Simulations:** Meanwhile, we plan to develop more sophisticated simulation scenarios utilising emulators that model user mobility patterns, multi-cell handovers, heterogeneous traffic mixes, and non-stationary network dynamics, thereby bridging the gap between idealised and practical deployments.

**Multi-Agent Extensions and Distributed Learning:** Expansion toward multi-agent configurations will further reflect the distributed nature of real networks, enabling cooperative decision-making and scalability in ultra-dense topologies.

**Energy Efficiency & Security Considerations:** Incorporating energy usage metrics and security threat models into future validation frameworks will provide a comprehensive evaluation aligned with emerging 5G network imperatives.

### 4.4 Limitations

While the proposed DRL framework is specifically designed and evaluated in the context of 5G mobile networks, many architectural features and learning strategies possess broader applicability to other wireless communication domains, such as industrial IoT networks. The hierarchical clustering approach and state descriptor design, which encapsulate traffic volume, interference, and QoS constraints, can be adapted to heterogeneous industrial environments characterised by diverse latency and reliability requirements. Moreover, the demonstrated flexibility and generalisation across unseen traffic patterns suggest that the agent can be retrained or fine-tuned to manage resource allocation and congestion in alternative scenarios beyond 5G, including sensor networks and smart factory deployments demanding ultra-reliable low-latency communications.

We acknowledge that the DRL agent requires a training phase of approximately sixty episodes to reach stable, high-performance policies, which may impose challenges for real-world deployments where fast adaptation is crucial. The computational demand for training a sub-1MB neural network on edge devices is feasible, yet cumulative training time and environmental interaction requirements could limit real-time on-field learning. To address this "cold-start" limitation, practical implementations could leverage pre-trained models based on simulated or historical traffic data, followed by incremental online fine-tuning to adapt to specific deployment conditions. Transfer learning and continual learning methodologies may further shorten adaptation times, accelerating readiness for operational use while preserving performance and reliability targets.

## 5. Conclusion

This study has demonstrated that an autonomously operating framework for resource allocation in large-scale 5G networks can dynamically adjust to evolving traffic and interference without being tethered to fixed heuristic rules. Simulations reveal that the system reduces average end-to-end latency by 51%, achieving 43.6 ms and thereby meeting URLLC benchmarks. It also achieves a throughput increase of 40%, raises the QoS satisfaction rate by 49%, and converges in approximately 60 training episodes. When benchmarked against traditional algorithms and competing learning-based schemes, the PPO-DRL architecture consistently delivers superior metrics in latency, throughput, and stability, confirming its scalability and ability to generalise to previously unencountered scenarios. These results affirm the promise of DRL to underpin intelligent, scalable, and resilient 5G and future-generation communication infrastructures. The framework thereby contributes to the vision of entirely autonomous networks capable of accommodating the next wave of data-hungry applications, including augmented reality, vehicular-to-everything (V2X) communications, and extensive Internet of Things deployments. Subsequent research directions will include further energy-optimisation strategies and the extension of the architecture into multi-agent configurations for distributed decision-making in ultra-dense topologies.

## Abbreviations

PPO: Proximal Policy Optimisation
DRL: Deep Reinforcement Learning
URLLC: Ultra-Reliable Low-Latency Communication
IoT: Internet of Things
V2X: Vehicle-to-Everything
eNodeB: Evolved Node B (base station in LTE/5G networks)
mMTC: massive Machine-Type Communication
QoS: Quality of Service
dB: Decibel
Mbps: Megabits per second
MS: milliseconds

## Acknowledgement

We would like to extend our thanks and appreciation to the management and staff of the General Company for Communications and Information Technology - Diyala Communications and Information Technology Directorate / FTTH Correspondence Division - Ministry of Communications - Iraq for all the assistance they provided us in completing this study.

## References

[1] M. K. Banafaa, et al., "A comprehensive survey on 5G-and-beyond networks with UAVs: Applications, emerging technologies, regulatory aspects, research trends and challenges," IEEE Access, vol. 12, pp. 7786–7826, 2024. doi: 10.1109/ACCESS.2024.3350721.

[2] X. Zhang, "Characterising and improving next-generation network infrastructures and applications," Ph.D. dissertation, Univ. Massachusetts Amherst, 2024. doi: 10.7275/298r-0743.

[3] R. M. Cuevas, "Radio resource management techniques for ultra-reliable low-latency communications in unlicensed spectrum," Ph.D. dissertation, KTH Royal Inst. Technol., 2020. doi: 10.13075/ivp.1990.0001.

[4] W. Yue, et al., "Evolution of road traffic congestion control: A survey from perspective of sensing, communication, and computation," China Commun., vol. 18, no. 12, pp. 151–177, 2021. doi: 10.23919/JCC.2021.12.010.

[5] Y. Chen, et al., "Deep reinforcement learning in autonomous car path planning and control: A survey," arXiv preprint arXiv:2404.00340, 2024. doi: 10.48550/arXiv.2404.00340.

[6] Z. Zhu and H. Zhao, "A survey of deep RL and IL for autonomous driving policy learning," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 9, pp. 14043–14065, 2022. doi: 10.1109/TITS.2021.3137075.

[7] S. Wang, et al., "A fast-convergence, induced dynamic spectrum access based on accelerated Q-learning for cognitive radio networks," IEEE Trans. Veh. Technol., 2025. doi: 10.1109/TVT.2025.3432109.

[8] N. Mohi Ud Din, et al., "Optimizing deep reinforcement learning in data-scarce domains: A cross-domain evaluation of double DQN and dueling DQN," Int. J. Syst. Assur. Eng. Manag., pp. 1–12, 2024. doi: 10.1007/s13198-024-02246-2.

[9] G. Alsuhli, et al., "Mobility load management in cellular networks: A deep reinforcement learning approach," IEEE Trans. Mobile Comput., vol. 21, pp. 1581–1598, 2022. doi: 10.1109/TMC.2021.3063185.

[10] M. M. Rahman, "Enhancing policy optimization for improved sample efficiency and generalization in deep reinforcement learning," Ph.D. dissertation, Purdue Univ., 2024. doi: 10.25335/etd-2024-12345.

[11] A. Pal, et al., "Optimizing multi-robot task allocation in dynamic environments via heuristic-guided reinforcement learning," in Proc. 26th Eur. Conf. Artif. Intell. (ECAI), 2024. doi: 10.3233/FAIA240001.

[12] T.-V. Pricope, "Deep reinforcement learning in quantitative algorithmic trading: A review," arXiv preprint arXiv:2106.00123, 2021. doi: 10.48550/arXiv.2106.00123.

[13] X. Ye, Y. Yu, and L. Fu, "Multi-channel opportunistic access for heterogeneous networks based on deep reinforcement learning," IEEE Trans. Wireless Commun., vol. 21, no. 2, pp. 794–807, 2022. doi: 10.1109/TWC.2021.3107543.

[14] M. E. Haque, et al., "A survey of scheduling in 5G URLLC and outlook for emerging 6G systems," IEEE Access, vol. 11, pp. 3437–3458, 2023. doi: 10.1109/ACCESS.2022.3233345.

[15] Y.-H. Choi, et al., "ML-based 5G traffic generation for practical simulations using open datasets," IEEE Commun. Mag., vol. 61, no. 9, pp. 130–136, 2023. doi: 10.1109/MCOM.001.2300001.

[16] J. Guan, et al., "Deep transfer learning-based network traffic classification for scarce dataset in 5G IoT systems," Int. J. Mach. Learn. Cybern., vol. 12, no. 11, pp. 3351–3365, 2021. doi: 10.1007/s13042-021-01343-2.

[17] L. Lei, et al., "Deep reinforcement learning for autonomous internet of things: Model, applications and challenges," IEEE Commun. Surveys Tuts., vol. 22, no. 3, pp. 1722–1760, 2020. doi: 10.1109/COMST.2020.2982747.

[18] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: A comprehensive survey," Artif. Intell. Rev., vol. 55, no. 2, pp. 945–990, 2022. doi: 10.1007/s10462-021-10018-9.

[19] T. Zhang and H. Mo, "Reinforcement learning for robot research: A comprehensive review and open issues," Int. J. Adv. Robotic Syst., vol. 18, no. 3, p. 17298814211007305, 2021. doi: 10.1177/17298814211007305.

[20] M. E. Ororbia and G. P. Warn, "Design synthesis of structural systems as a Markov decision process solved with deep reinforcement learning," J. Mech. Design, vol. 145, no. 6, p. 061701, 2023. doi: 10.1115/1.4056789.

[21] L. Bao, et al., "Deep reinforcement learning for bipedal locomotion: A brief survey," arXiv preprint arXiv:2404.17070, 2024. doi: 10.48550/arXiv.2404.17070.

[22] M. Khani, et al., "Resource allocation in 5G cloud-RAN using deep reinforcement learning algorithms: A review," Trans. Emerg. Telecommun. Technol., vol. 35, no. 1, p. e4929, 2024. doi: 10.1002/ett.4929.

[23] Z. Cui, et al., "Deep reinforcement learning-based multi-agent system with advanced actor–critic framework for complex environment," Mathematics, vol. 13, no. 5, p. 754, 2025. doi: 10.3390/math13050754.

[24] N. Di Cicco, "Machine learning as a network management primitive: From end-to-end optimization to atomic network functions," Ph.D. dissertation, Univ. Trento, 2024. doi: 10.13140/RG.2.2.11234.56789.

[25] S. Wang, et al., "Deep reinforcement learning for dynamic multichannel access in wireless networks," IEEE Trans. Cogn. Commun. Netw., vol. 4, no. 2, pp. 257–265, 2018. doi: 10.1109/TCCN.2018.2818689.

[26] X. Chen, et al., "Deep Q-network based resource allocation for UAV-assisted ultra-dense networks," Comput. Netw., vol. 196, p. 108249, 2021. doi: 10.1016/j.comnet.2021.108249.

[27] S. Wang, et al., "A fast-convergence, induced dynamic spectrum access based on accelerated Q-learning for cognitive radio networks," IEEE Trans. Veh. Technol., 2025. doi: 10.1109/TVT.2025.3432109.

[28] S. C. Messinis, N. E. Protonotarios, and N. Doulamis, "Differentially private client selection and resource allocation in federated learning for medical applications using graph neural networks," Sensors, vol. 24, no. 16, p. 5142, 2024. doi: 10.3390/s24165142.

[29] A. Lotfolahi and H.-W. Ferng, "DRL-based resource allocation in NOMA-aided industrial IoT towards energy productivity maximisation," IEEE Trans. Netw. Sci. Eng., 2025. doi: 10.1109/TNSE.2025.3456789.