ASIAN JOURNAL OF ELECTRICAL AND ELECTRONIC ENGINEERING
Vol. 6 No. 1 2026
E-ISSN: 2785-8189

TD3 vs. DDPG for RIS-Assisted Beamforming Optimization: Statistical and Communication-Level Analysis for 6G IoT Networks

Ezdihar Osman Taj Almowla Mohomad[1*], Khalid Hamid Bilal[2], Zeinab Mahmoud Omer[1], Abeer Mohamed Elzain[3], and Rania Ali Elkhidir[4]

[1]*University of Bahri Khartoum, Sudan*
[2]*University of Science & Technology, Omdurman*, Sudan
[3]*University of Bahri, Khartoum. Sudan*
[4]*University of Hail, Saudi Arabia*

*Corresponding author: Ezdiharosman22@gmail.com

(Received: 23 February 2026; Accepted: 14 June 2026)

Abstract—Reconfigurable Intelligent Surfaces (RIS) are expected to play a critical role in future 6G wireless networks by enabling adaptive and intelligent signal propagation. This study presents a deep reinforcement learning (DRL)-based simulation framework for beamforming optimization in RIS-assisted wireless communication systems. Two continuous-control DRL algorithms, namely Twin Delayed Deep Deterministic Policy Gradient (TD3) and Deep Deterministic Policy Gradient (DDPG), were evaluated under 30 independent wireless channel scenarios. The experimental results demonstrate that TD3 achieves a higher average episodic return (13.34 ± 0.73) than DDPG (11.13 ± 0.87), representing an improvement of approximately 19.86%. Statistical analysis further confirms that this improvement is significant ($t \approx 3.99$, $p < 0.001$) with a large effect size (Cohen's $d = 1.03$). The 95% confidence interval of the mean difference ranges from 1.10 to 3.32. In addition to better average performance, TD3 exhibits steadier convergence and reduced reward variability. This means that learning is more robust under dynamic wireless channel conditions. The results suggest that TD3 provides a robust and statistically reliable method for optimizing intelligent beamforming in RIS-assisted 6G IoT communication scenarios, thereby improving wireless connectivity performance and system stability.

Keywords: Reconfigurable Intelligent Surfaces (RIS), Coverage Enhancement,6G Internet of Things (6G IoT), Deep Deterministic Policy Gradient (DDPG), Deep Reinforcement Learning (DRL)

1. INTRODUCTION

A low-latency, comprehensive, fast, and reliable connection is now more important than ever, thanks to advances in wireless communication technology [1][2][3]. To meet these critical needs, the International Telecommunication Union (ITU) and other interested parties have developed a plan and primary goals for sixth-generation (6G) networks. Most people think that the most important thing is to strengthen the network, integrate the complete Internet of Things (IoT), and improve services [4] [5].

Included in the many traditional means of human-to-human contact are base stations [6][7][8]. Base stations have difficulty handling dense, broad IoT deployments due to transmission issues such as fluctuating interference [9][10], multipath fading, substantial obstruction, and instances of undetected signals (NLoS) [11][12]. To enhance wireless communication while reducing prices and energy usage, researchers have investigated novel methods. Researchers found that intelligent reflecting surfaces (IRS) and reconfigurable intelligent surfaces (RIS) can enhance wireless networks [13]. Researchers can construct a spectrum of RIS-based passive reflective devices to respond to various phases [14]. This implies the ability to instantly alter electromagnetic waves [15]. Reducing noise while increasing signal length and intensity is the goal of this function. Because it requires little power, no

DOI: https://doi.org/10.69955/ajoeee.2026.v6i1.84

costly equipment [16], and no complex installations, RIS is a crucial technology for 6G IoT communications [17].

With a small number of Internet of Things (IoT) devices and multiple low-power devices operating simultaneously [17] [18], traditional beamforming methods may rely on coverage or fail to meet quality-of-service (QoS) standards [19]. Not only is space navigation severely lacking, but the tube itself is woefully inadequate. By increasing the number of signal segments, decreasing interference, and creating links that appear to be in line of sight, RIS can overcome these issues. When it comes to improving and expanding the network, the invention is the way to go. We are still in the early stages of developing RIS beamforming for 6G IoT networks. Despite RIS's many obvious advantages [20], determining the optimal beamforming and phase-shift configurations for a large-scale Internet of Things system [21] can be difficult. The RIS provides a multilevel setting for continuous activity by allowing each reflecting element to select its phase from the range [0, 2π] [22]. The base station broadcasts a non-linear RIS-IoT channel that is highly linked. It changes over time, especially in busy or noisy environments [23]. Methods require substantial processing power, making immediate enhancement difficult. People rarely use heuristics such as semidefinite relaxation, alternating optimization, and search strategies [24]. These processes can be time-consuming, expensive, and inefficient. Furthermore, in real-world IoT applications, it can be difficult to generate highly accurate channel models or precise channel state information (CSI) that these methods often require [25]. When dealing with several variables, changing conditions, and ongoing supervision, intelligent optimization frameworks become indispensable. Deep Reinforcement Learning (DRL) offers a significant advantage for making complex decisions in unpredictable, rapidly changing environments [26]. Instead of relying on explicit channel models or trial and error, DRL automatically learns the best rules. Rather, it engages with its surroundings instantly. Deep Deterministic Policy Gradient (DDPG) and Twin Delayed Deep Deterministic Policy Gradient (TD3) are two actor-critic approaches that could improve systems requiring ongoing refinement. Beamforming and RIS can now work together.

Recent research has focused on specific areas such as security, anti-jamming, and UAV communication, as well as optimal network topologies and discrete or quantized action spaces. However, several RIS-based wireless systems have incorporated Deep Reinforcement Learning (DRL) [27]. Research on modern continuous-control deep reinforcement learning methods used in common RIS-IoT applications is lacking. Despite the importance of these challenges to the successful implementation of 6G, problems with training stability, convergence characteristics, and resilience to inaccurate or misleading Channel State Information (CSI) persist.

A continuous-control deep reinforcement learning strategy may enhance RIS-assisted beamforming in 6G IoT networks, according to the study's findings. The Markov Decision Process (MDP) allows beamforming and phase shifting in the RIS. Our empirical evaluations focused on DDPG and TD3, two of the most prominent actor-critic systems. To improve the signal-to-interference-plus-noise ratio (SINR) and network coverage, the suggested approach uses realistic 3D topology modeling and various IoT scenarios. Its size will have no effect on its effectiveness.

The main contributions of this work can be summarized as follows:

1. A single, continuous-action deep reinforcement learning system for optimizing beamforming in large 6G IoT networks with reconfigurable intelligent surface integration.
2. A complete MDP formulation that covers the state space, action space, reward function, and system dynamics of RIS-IoT communications .
   A detailed comparison of DDPG and TD3 in the same RIS-IoT setups, looking at how fast they converge, how stable they are, and how strong they are.
3. A method for making datasets that show 3D topology, describe the properties of wireless channels, set up BS-RIS-IoT, and spread users out.
4. A comprehensive performance evaluation that validates the effectiveness of the proposed strategy through convergence analysis, SINR distributions, and coverage maps.

Key Contributions and Novelty of the Proposed Framework

This work is novel in that it provides a comprehensive comparative study of the continuous-control DRL algorithms for RIS-assisted beamforming optimizations under the same wireless communication settings. Compared with traditional studies that mainly focus on reward convergence, the proposed approach provides a multi-dimensional evaluation, including reward stability, statistical distribution analysis, SINR performance, sum-

rate analysis, and convergence robustness.

More specifically, the present study contributes to the following:

- A unified RIS-assisted DRL evaluation framework for fair comparison between DDPG and TD3 under identical wireless channel conditions.
- A detailed statistical performance analysis including confidence intervals, variance analysis, boxplot evaluation, and effect-size interpretation.
- Communication-level performance evaluation using SINR and achievable sum-rate metrics in addition to reward-based learning analysis.
- A practical discussion on real-world RIS deployment challenges and future integration with realistic wireless propagation environments such as DeepMIMO and ray-tracing models.

The results obtained demonstrate that TD3 provides significantly improved convergence stability, reward consistency, and communication reliability compared to DDPG in dynamic RIS-assisted 6G IoT environments.

Recently, reinforcement learning algorithms, such as Soft Actor–Critic (SAC) and Proximal Policy Optimization (PPO), have shown strong performance in continuous control tasks. In this work, we focus on deterministic policy-gradient methods. We select DDPG and TD3 because TD3 was initially proposed as a direct improvement over DDPG, enabling a controlled, systematic comparison under the same RIS-assisted wireless communication conditions. We restrict the analysis to closely related deterministic actor-critic algorithms to isolate the effect of TD3's architectural enhancements: twin critics, delayed policy updates, and target-policy smoothing. This avoids introducing additional algorithmic differences related to entropy-based or policy-gradient approaches.

2. RELATED WORK

Recent research has increasingly explored deep reinforcement learning (DRL) techniques for reconfigurable intelligent surface (RIS)–assisted wireless systems. In [28], a DRL-based joint beamforming design was investigated for RIS-assisted wireless networks; however, the study relies on a single learning paradigm and focuses mainly on rate optimization rather than coverage enhancement. The work in [29] extended DRL to joint deployment and passive beamforming optimization in RIS-assisted networks, yet it did not examine algorithmic comparisons under a unified environment. A broader DRL-oriented perspective on RIS-assisted communications was presented in [30], which outlined key challenges and opportunities but lacked a concrete MDP formulation and comparative performance analysis. More recent studies, such as [31], addressed DRL-based beamforming optimization in RIS-assisted multi-user MISO systems, primarily targeting spectral efficiency and robustness under specific channel conditions. Additionally, TD3-based joint beamforming frameworks have been proposed for RIS-assisted systems with imperfect CSI in [32]. Still, these works employ a single DRL algorithm and do not focus on IoT-centric coverage metrics. Unlike the aforementioned studies, the present work formulates a complete MDP. It provides a unified comparative evaluation of DDPG and TD3 for coverage enhancement in dense RIS-assisted 6G IoT networks, thereby constituting a clear methodological and performance-driven advancement over existing literature.

Existing research on RIS-assisted DRL is largely limited to single-metric optimization or reward-convergence behaviors in constrained simulation environments. Furthermore, few previous studies have provided extensive statistical analysis, communication-level performance analysis, and comparative stability assessment of continuous-control DRL algorithms under the same wireless channel conditions. Moreover, the throughput-oriented evaluation and SINR-based coverage analysis have received less consideration in the context of RIS-assisted IoT communication. To address these limitations, the proposed work offers a comprehensive comparative approach to assess DDPG and TD3 by considering statistical analysis, reward-distribution interpretation, SINR evaluation, and achievable sum-rate performance under the same RIS-assisted wireless settings.

3. SYSTEM OVERVIEW AND PROBLEM FORMULATION

3.1 Network Model

As depicted in Fig. 1, we examine a three-dimensional (3D) RIS-assisted multi-user IoT downlink system functioning within a 6G framework. A base station (BS) with N_t antennas serves K single-antenna IoT users via a reconfigurable intelligent surface (RIS) comprising M passive reflecting elements.

Two links establish communication between the BS and IoT users: a direct link from the BS to the user and an indirect, reflected link via the RIS. The RIS facilitates coherent signal combining at the receivers by dynamically adjusting its phase shifts, which improves coverage and signal quality in dense IoT environments.

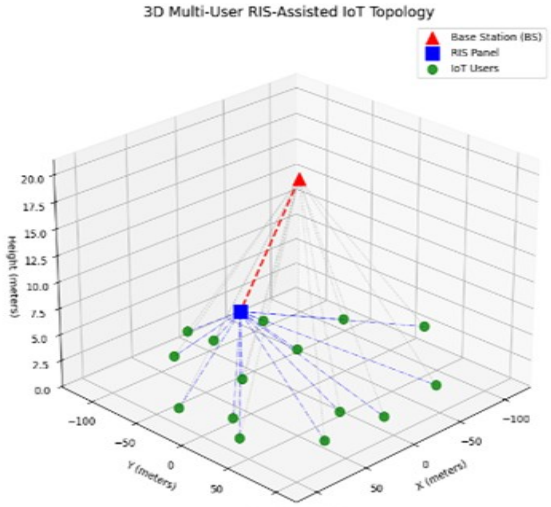

Fig.1 illustrates the 3D multi-user RIS-assisted IoT topology used in this study

3.2 Channel Model

The direct channel between the BS and the $\boldsymbol{k}$-th user is denoted by

$$\mathrm{h}_{b,k} \in \mathbb{C}^{1\times N_t}. \qquad (1)$$

The channel from the BS to the RIS and from the RIS to the $\boldsymbol{k}$-th user are denoted by

$$\mathrm{G}_{b,r} \in \mathbb{C}^{M\times N_t}, \mathrm{h}_{r,k} \in \mathbb{C}^{1\times M}, \quad (2)$$

respectively.

The RIS reflection matrix is defined as

$$\Theta = diag\left(e^{j\theta_1}, e^{j\theta_2}, \dots, e^{j\theta_M}\right), \qquad (3)$$

where $\boldsymbol{\theta_m} \in [\mathbf{0}, \mathbf{2\pi})$represents the phase shift of the $\boldsymbol{m}$-th RIS element.

Accordingly, the effective channel for the $\boldsymbol{k}$-th user is given by

$$\mathrm{h}_k^{\mathrm{eff}} = \mathrm{h}_{b,k} + \mathrm{h}_{r,k}\Theta\mathrm{G}_{b,r}. \quad (4)$$

3.3 Signal Model

The transmitted signal from the base station (BS) is given by

$$\mathrm{x} = \sum_{k=1}^{K} \mathrm{w}_k\, s_k, \qquad (5)$$

where $w_k \in \mathbb{C}^{N_t \times 1}$denotes the beamforming vector associated with the k-th IoT user, and $s_k$represents the corresponding information symbol satisfying $\mathbb{E}[|\ s_k\ |^2] = 1$.

The received signal at the k- The user can be expressed as

$$y_k = h_k^{eff} w_k s_k + \sum_{i \neq k} h_k^{eff} w_i s_i + n_k, \qquad (6)$$

where h_k^{eff}denotes the effective BS–RIS–user channel, and $n_k \sim \mathcal{CN}(0, \sigma^2)$represents additive white Gaussian noise (AWGN) with variance σ^2.

Accordingly, the signal-to-interference-plus-noise ratio (SINR) at the k- The user is defined as

$$\mathrm{SINR}_k = \frac{|\ h_k^{eff} w_k\ |^2}{\sum_{i \neq k} |\ h_k^{eff} w_i\ |^2 + \sigma^2}. \qquad (7)$$

3.4 Problem Formulation

The objective of this work is to jointly optimize the BS beamforming vectors. $\{w_k\}$and the RIS phase shift matrix ΘTo enhance coverage and communication quality in dense IoT deployments. The optimization problem is formulated as

$$\begin{aligned} \max_{\{w_k\}, \Theta} \quad & \sum_{k=1}^{K} \log_2 (1+\mathrm{SINR}_k) \\ \text{s.t.} \quad & \sum_{k=1}^{K} \|\ w_k\ \|^2 \leq P_{\max}, \\ & \theta_m \in [0, 2\pi), \forall m, \end{aligned} \qquad (8)$$

where $P_{\max}$denotes the maximum transmit power at the BS. This optimization problem is highly non-convex due to the coupling between beamforming vectors and continuous RIS phase shifts. Therefore, it is reformulated as a Markov decision process (MDP) and solved using continuous-control deep reinforcement learning, as described in the following section.

The first constraint ensures practical, energy-efficient operation by limiting the base station's total transmit power to the maximum allowable value Pmax. The second limitation restricts each RIS phase-shifting element to the physical range $[0,2\pi)$, which represents the achievable phase adjustment capability of the RIS hardware

The optimization problem is non-convex because the beamforming vectors and RIS phase-shift variables are jointly coupled in the SINR expression. Consequently, obtaining a globally optimal solution through conventional optimization techniques is computationally challenging, particularly in dynamic multi-user wireless environments. Therefore, the problem is reformulated as a Markov Decision Process (MDP) and solved using continuous-control deep reinforcement learning.

3.5 Motivation for DRL-Based Design

The problem of configuring RIS and joint beamforming involves high dimensionality, strong nonlinearity, and continuous control variables, which render traditional model-based optimization methods impractical in dynamic 6G IoT environments. By allowing the learning agent to engage directly with the wireless environment, deep reinforcement learning (DRL) offers an effective solution for learning optimal control policies without dependence on explicit channel models or convex reformulations.

Note that the channel model described in this section is used to build the theoretical framework for RIS-assisted communications and to specify the underlying communication relationships. In contrast, the DRL environment used in this study is based on dataset-derived communication-performance

features that are used for practical training, state representation, and performance evaluation.

4. DRL-BASED JOINT BEAMFORMING AND RIS OPTIMIZATION

4.1 DRL Framework Overview

A deep reinforcement learning framework is adopted to address the non-convex joint beamforming and RIS optimization problem in RIS-assisted multi-user IoT networks. The BS is modeled as the learning agent, while the RIS-assisted wireless system represents the environment. At each decision step, the agent observes the current network state and selects continuous control actions to improve coverage and signal quality.

4.2 Markov Decision Process Formulation

The RIS-assisted beamforming problem is formulated as an MDP defined by the tuple. $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$:

State: The state at time step $t$captures the essential channel and configuration information and is defined as

$$st = \{xt, Tt, Lt, Et, \gamma t\}, \qquad (9)$$

where $\boldsymbol{xt}$ denotes the normalized dataset-derived network features at time step t, $\boldsymbol{Tt}$ represents the normalized throughput, $\boldsymbol{Lt}$ represents the normalized latency, $\boldsymbol{Et}$ denotes the normalized energy-consumption value, and $\boldsymbol{\gamma t}$ denotes the SINR-related performance indicator when available.

The state representation is constructed from processed communication-performance features extracted from the dataset rather than explicit raw channel-state information (CSI). Consequently, the DRL agent observes a compact representation of the wireless communication environment that captures throughput, latency, energy efficiency, SINR-related indicators, and other relevant network-performance characteristics required for optimization.

Action: The action selected by the agent jointly controls the BS beamforming and RIS configuration and is defined as

$$\mathrm{a_t} = \{\mathrm{w_t}, \theta_\mathrm{t}\}, \theta_\mathrm{m} \in [0,2\pi), \mathrm{m}, \qquad (10)$$

where $\mathrm{w}_t$denotes the BS beamforming vector and $\theta_t$represents the continuous RIS phase-shift vector.

4.3 Action Space Limitations

We restrict the action space to the range [−1, 1] across all action dimensions, which enables stable learning and valid control decisions. When interacting with the environment, the actor network's output actions are clipped to fall within predefined limits before being applied.

The bounded continuous-action representation improves training stability and prevents the generation of very large control values that could negatively impact the optimization process. This constraint also ensures a consistent exploration strategy during training and evaluation while maintaining feasible actions throughout the learning process.

The reward function is defined to maximize communication performance by improving throughput while minimizing latency and energy consumption, and is expressed as

$$rt = wthrTt - wlatLt - wenEt \qquad (11)$$

where Tt, Lt, Et are the normalised throughput, latency and energy-consumption values at time step t, respectively. The contribution of each performance metric is controlled by the weights w_thr, w_lat, and w_en. In this study, w_thr=1.0, w_lat=0.3, and w_en=0.2.

This reward formulation encourages the agent to perform actions that increase throughput while decreasing latency and energy consumption. The reward value is clipped to [-1, 1] to ensure stable learning during training.

In real-world 6G IoT dynamics, continuous control optimization is possible because the environment transitions to a new state s_(t+1) due to changes in the channel and the chosen action.

This multi-objective reward design enables the agent to balance spectral efficiency, delay performance, and

energy efficiency while optimizing its policy.

Actor-critic deep reinforcement learning (DRL) algorithms are used to address continuous action spaces and sequential decision-making. In this study, DDPG and TD3 are selected because they are specifically designed for continuous control optimization problems and have shown strong performance in high-dimensional wireless communication environments.

DDPG learns deterministic policies using an actor–critic architecture, which improves training stability. But it can lead to value overestimation during training. TD3 mitigates this limitation by using twin critic networks, delayed policy updates and target-policy smoothing. Such improvements alleviate overestimation bias and improve the robustness of the policy, making TD3 especially suitable for continuous-control optimisation problems.

4.4 DDPG and TD3-Based Learning

Actor–critic DRL techniques manage the continuous action space associated with RIS phase shifts and beamforming vectors. Specifically, the Deep Deterministic Policy Gradient (DDPG) and Twin Delayed Deep Deterministic Policy Gradient (TD3) are investigated.

DDPG directly learns deterministic policies for continuous control, whereas TD3 mitigates overestimation bias by enhancing training stability through twin critic networks and delayed policy updates. In high-dimensional optimization scenarios, both methods rely on experience replay buffers and target networks to provide consistent convergence.

4.5 . Methodology Workflow

The full deep reinforcement learning pipeline used in this study is shown in Figure 2. The workflow shows the interaction between the wireless communication environment and the DRL agent, including state observation, action selection, reward generation, experience replay, policy updating, convergence verification and final performance evaluation.

5. SIMULATION SETUP

5.1 Data Source and Preprocessing

The Kaggle platform [33], which provides a vast collection of wireless communication and beamforming performance samples, provided the dataset used in the simulation study. Numerous transmission samples with various system settings, including variations in transmission power, carrier frequency, number of antennas, mobility circumstances, and optimization state, are included in the original dataset. Several procedures were followed to ensure that the data were reliable and appropriate for learning-based optimization. The initial step in eliminating poor samples was to eliminate duplicate and incomplete entries. Second, we ensured that all deep reinforcement learning (DRL) agents had the same numerical feature ranges, so each agent underwent the same training procedure. Additionally, category environmental features and optimization flags were quantified. Each simulation, training, and evaluation experiment uses the cleaned dataset. It is important to note that the processed dataset is not used as a static lookup table during training. Instead, each dataset row represents a distinct wireless communication state characterized by throughput, latency, energy efficiency and other network-related parameters. The DRL environment is initialized from these states, while state transitions and reward evolution are generated dynamically through the interaction between the agent and the environment. Therefore, the dataset serves as a source of realistic wireless communication conditions rather than a fixed sequence of observations.

5.2 Dynamic Environment Interaction

The processed dataset is not a static lookup table used during training. Instead, each sample in the dataset corresponds to a different wireless communication state, with parameters such as throughput, latency, energy efficiency and other network-related metrics. In each interaction step, the DRL agent observes the current state, selects a continuous action, and receives a reward based on the resulting communication performance metrics. This, in turn, causes the environment to apply the chosen action to determine the next state transition. Therefore, the next state depends not only on the original data sample, but also on the current state and the action taken by the agent. Such an interaction mechanism allows the agent to explore different network conditions and learn

adaptive decision-making policies through continuous trial and error.

Therefore, the environment-agent interaction determines the state transition process. The action affects future observations and reward generation during training.

Therefore, the dataset is only used to initialise the realistic states of wireless communication. In contrast, the subsequent state transitions, reward evolution, and policy updates are generated online through the continuous interaction between the DRL agent and the environment.

In the current implementation, channel evolution is characterised by communication performance features extracted from the processed dataset, rather than by explicit physical-layer channel reconstruction. Therefore, the environment assesses the effect of each action on the observed performance metrics during interaction.

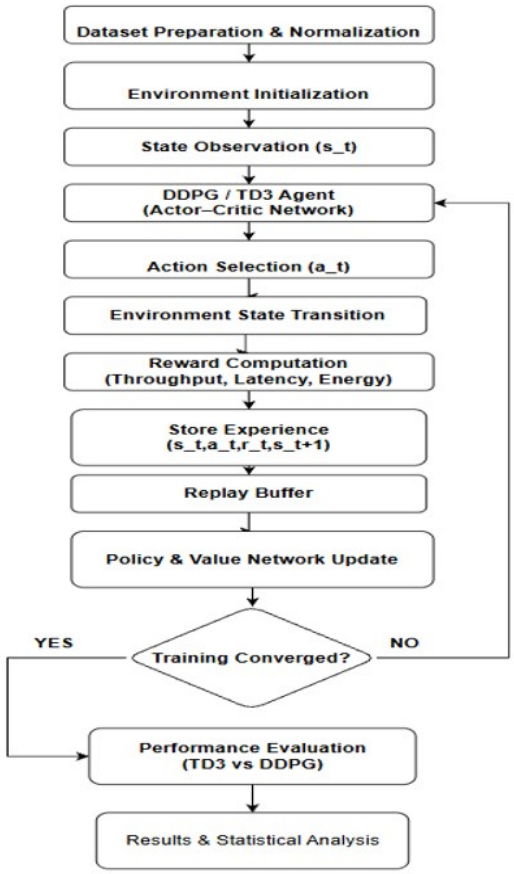

Fig. 2. The detailed DRL-based workflow of the proposed DDPG and TD3 framework includes dataset preparation, environment initialization, state observation, action selection, reward computation, experience replay, policy learning, convergence assessment, and performance evaluation. The processed dataset is then used to initialise realistic wireless communication scenarios as shown in Figure 2. The DRL agent then interacts with the environment through a sequential decision-making process, in which actions influence state transitions and yield rewards based on throughput, latency, and energy consumption. The collected experiences are stored in a replay buffer and used to iteratively update the actor-critic networks until training converges. Finally, the trained DDPG and TD3 models are evaluated and compared based on communication levels and statistical performance metrics.

The state-transition process follows.

$$s\_(t+1) = \{x\_(t+1), T\_(t+1), L\_(t+1), E\_(t+1), \gamma\_(t+1)\} \quad (12)$$

where s_t is the current state and s_(t+1) is the next state resulting from the interaction between the DRL agent

and the environment. The performance indicator related to the updated network features, throughput, latency, energy-consumption, and SINR are denoted by x_{t+1}, T_{t+1}, L_{t+1}, E_{t+1} and γ_{t+1} respectively. Therefore, the selected action affects future observations, reward generation and decision-making dynamics in subsequent training steps.

5.3 Network and Learning Configuration

A reconfigurable intelligent surface (RIS) enhances beamforming techniques used by a base station in a downlink RIS-assisted communication framework. It is believed that the RIS functions as a passive helping element that enhances signal propagation and lessens adverse channel conditions. Adapting beamforming-related parameters in response to observed system performance metrics is the main goal of the optimization procedure. Deep reinforcement learning methods for continuous control are used to maximize system behavior. In particular, to ensure a fair and repeatable comparison, Deep Deterministic Policy Gradient (DDPG) and Twin Delayed Deep Deterministic Policy Gradient (TD3) are trained and evaluated on the same processed dataset. Through state observations obtained from dataset features, both algorithms engage with the environment and modify their policies in response to reward feedback calculated from performance measures.

5.4 Hyperparameter Configuration

The selection of hyperparameters strongly influences the training performance of deep reinforcement learning algorithms. To improve reproducibility and provide a transparent comparison framework, the main training hyperparameters used for DDPG and TD3 are summarized in Table 1

Table 1. DRL Hyperparameter Configuration Used for DDPG and TD3 Training

Parameter	DDPG	TD3
Policy Network	MlpPolicy	MlpPolicy
Learning Rate	3×10^{-4}	1×10^{-4}
Batch Size	256	256
Replay Buffer Size	150,000	200,000
Discount Factor (γ)	0.99	0.99
Soft Update Coefficient (τ)	0.005	0.005
Train Frequency	4	4
Gradient Steps	2	4
Action Noise Standard Deviation	0.15	0.10
Policy Delay	N/A	2
Target Policy Noise	N/A	0.10
Target Noise Clip	N/A	0.20
Total Training Timesteps	120,000	150,000

The hyperparameters were selected through preliminary experimental tuning to achieve stable convergence and ensure a fair comparison between DDPG and TD3 under identical training conditions.

To enhance reproducibility, both algorithms were trained using identical environment settings, including a maximum episode length of 20 steps and a fixed random seed of 123. The only differences between the two algorithms are limited to their algorithm-specific learning mechanisms and hyperparameters, as summarized in Table 1.

5.5 Dataset-Derived Numerical Parameters

Table 2 summarizes the numerical ranges and statistical characteristics of the key performance-related parameters directly extracted from the processed dataset. These values are computed from the dataset used in all experiments and therefore accurately reflect the operating conditions under which the proposed framework is evaluated**.**

5.6 Reproducibility Statement

All simulation results presented in this research are produced using the same processed dataset and identical training and evaluation conditions for both DRL algorithms. This ensures reproducibility and allows DDPG and TD3 to be openly compared under identical operating conditions.

Table 2: Dataset-Derived Simulation Parameters

Parameter	Min	Max	Mean	Std
SNR (dB)	5.00	19.99	12.59	4.30
Transmit Power (dBm)	10.00	34.94	22.16	7.22
Interference Power (dB)	−99.94	−50.01	−74.94	14.16
Number of Antennas	64	512	233.73	168.70
Carrier Frequency (GHz)	28	150	84.71	45.17
Bandwidth (MHz)	50	400	179.85	131.28
Beamforming Gain (dB)	10.02	25.00	17.58	4.41
Throughput (Mbps)	100.98	999.67	541.15	261.05
Latency (ms)	1.02	9.99	5.43	2.55
Energy Consumption (kWh/GB)	0.010	0.050	0.030	0.011

6. SIMULATION PERFORMANCE EVALUATION

This section shows how well the proposed deep reinforcement learning framework works with the same settings as in part 5. To ensure a fair comparison, the Deep Deterministic Policy Gradient (DDPG) and Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithms are evaluated under identical conditions. When using RIS to enable 6G IoT, the most important considerations are improving coverage, making learning more stable, using performance metrics tied to incentives, and ensuring interoperability.

6.1 Training Convergence Analysis

6.1.1 Moving-Average Reward Convergence

The training convergence behaviour of the proposed DRL agents is shown in Fig. 3 using moving-average episodic rewards. In the early stages of training, DDPG and TD3 both achieve increasing reward, demonstrating that they can interact with the RIS-assisted wireless environment. Yet, there are some differences in the convergence stability between the two algorithms. TD3 achieves faster convergence and much lower reward fluctuations than DDPG, indicating improved training stability and more reliable policy learning under dynamic wireless channel conditions.

The improved convergence stability of TD3 mainly stems from its twin-critic structure and delayed policy update strategy, which alleviates overestimation bias and stabilises the learning process. In contrast, DDPG exhibits greater reward fluctuations and slower convergence, indicating greater sensitivity to environmental noise and channel randomness during RIS beamforming optimisation.

The analysis in Fig. 3 shows that the episodic rewards of both algorithms increase rapidly at the start, indicating good interaction with the RIS-aided environment, followed by a slow improvement in the beamforming policy. However, the TD3 algorithm shows a much smoother and more stable convergence trend throughout the training phase than the DDPG algorithm, which exhibits greater fluctuations and slower stabilisation. The lower oscillations in rewards observed for TD3 in the lower is indicative of better performance in value estimation and policy learning. The results show that TD3 is more effective for optimising continuous-control RIS beamforming in dynamic wireless environments with channel uncertainty and stochastic fluctuations.

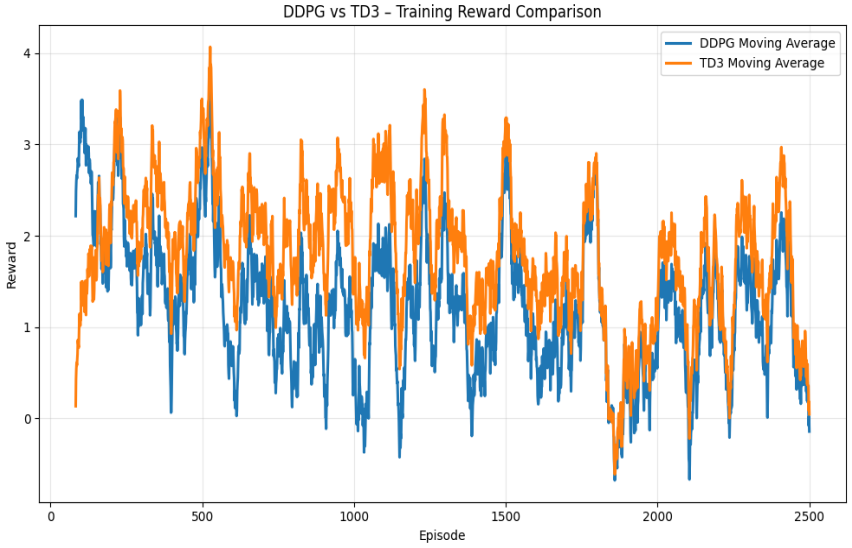

Fig. 3. Training reward convergence comparison between DDPG and TD3 using moving-average episodic rewards.

6.1.2 Statistical Boxplot Stability Analysis

In addition, we performed a statistical boxplot analysis to evaluate training stability and reward consistency, as shown in Fig. 4. The boxplot comparison allows us to interpret the distributions of episodic rewards achieved by both DRL algorithms.

Results show that TD3 achieves a higher median reward and a smaller interquartile range than DDPG, indicating greater reward stability and reduced performance fluctuations. In contrast, DDPG shows a broader distribution of rewards with some significant negative outliers, suggesting unstable convergence behaviour and a greater sensitivity to channel estimation uncertainty.

The enhanced stability of TD3 is mainly due to its delayed policy updates and dual-critic learning mechanism, which yield more stable value estimates during continuous RIS phase optimisation. Finally, the statistical analysis verifies that TD3 offers more robust and stable learning performance in the context of RIS-assisted IoT communication.

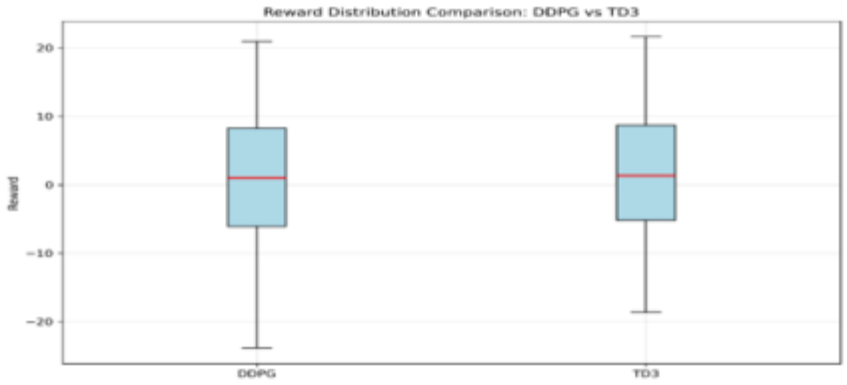

Fig. 4. Statistical boxplot comparison between DDPG and TD3.

The reward distribution of TD3 is more concentrated around higher reward values, as shown in Fig. 4, indicating more consistent learning and lower variance. TD3 achieves a higher median reward than DDPG, while the smaller interquartile range indicates more stable training. In contrast, DDPG has a broader reward distribution with fewer low-reward observations, which leads to greater sensitivity to environmental dynamics and less consistent policy

convergence. These results further confirm TD3's superiority for continuous-control RIS beamforming optimisation.

6.2 Evaluation Reward Analysis

The evaluation reward curves provide further insight into the stability and consistency of the learned DRL policies in RIS-assisted wireless communication environments.

6.2.1 DDPG Episode Rewards

Figure 5 shows the reward behaviour of the DDPG agent for the evaluation phase in 2500 episodes. The light blue curve shows episodic rewards, and the red curve shows the moving-average reward with a smoothing window of 83 episodes.

The results show significant variability in rewards during the evaluation process, with episodic rewards ranging from ~-22 to +18. The moving-average curve is relatively close to zero, and the long-term convergence stability is weak. However, the reward can be improved temporarily.

This behaviour highlights the difficulties DDPG policies face in maintaining stable performance in dynamic RIS-IoT channels with nonlinear wireless interactions and environmental uncertainty. The observed instability is primarily due to overestimation and insufficient policy smoothing in DDPG-based learning.

The overall evaluation results reveal that DDPG converges more slowly and is less stable than TD3 in continuous RIS-assisted beamforming optimisation tasks.

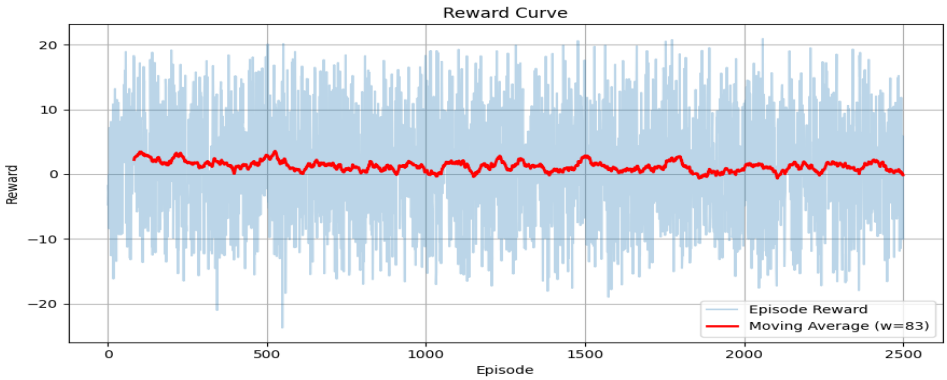

Fig. 5. DDPG evaluation episodic reward performance over 2500 episodes.

The DDPG evaluation rewards, as shown in Fig. 5, exhibit large fluctuations during the evaluation period, indicating that the policy's

performance is unstable under dynamic wireless channel conditions. Some episodes yield relatively high rewards, but these gains are not sustained over time, and the moving-average curve remains unstable. The high frequency of the reward oscillations is due to the learned policy's sensitivity to channel variations and environmental uncertainty. These observations show that DDPG has problems with reliable long-term performance for RIS-assisted beamforming optimisation and suggest that more stable continuous-control algorithms, such as TD3, should be used.

6.2.2 TD3 Evaluation Episode Rewards

The assessment reward behaviour of the TD3 agent is shown in Figure 6 over approximately 2500 evaluation sessions. The red curve is the moving-average reward trend with a smoothing window of about 83 episodes, and the light blue curve is the episodic rewards. We observe that TD3 has a significantly lower reward volatility than DDPG, and the reward transition is smoother during the evaluation. While reward changes are perceptible in the initial and middle phases of training, the moving-average trend exhibits robust long-term convergence behaviour. The episodic rewards, which are approximately in the range of −20 to +20, account for the stochasticity of the RIS-assisted wireless environments and dynamic channel variations. However, the smoothed reward curve remains in a stable positive range, suggesting that TD3 might still achieve reliable beamforming optimisation

performance under varying wireless conditions. The improved stability of TD3 is mainly due to its twin-critic design, target policy smoothing method and delayed policy update mechanism that together alleviate Q-value overestimation and enhance learning robustness in continuous RIS phase optimisation. All in all, the results indicate that TD3 provides more stable and reliable learning behaviour.

Fig. 6. TD3 evaluation episodic reward performance over 2500 episodes

6.3 Reward Distribution Analysis

Histograms provide a statistical interpretation of reward distributions and are useful for analyzing policy variability, operating ranges, and reward consistency during DRL evaluation.

6.3.1 Variability and Outlier Behavior in DDPG Rewards

Figure 7. Histogram of the reward distribution from DDPG evaluation episodes. The histogram illustrates a wide distribution of rewards, characterised by numerous negative values and a few low-value outliers. This extensive spread of rewards suggests unstable learning behaviour and inconsistent policy performance under RIS-assisted wireless channel conditions. The significant variance observed in the DDPG rewards further highlights the algorithm's sensitivity to environmental randomness and the nonlinearity of wireless interactions. Overall, the histogram analysis indicates that DDPG exhibits relatively unstable reward behaviour and diminished robustness in RIS-assisted beamforming optimisation tasks.

As shown in Fig. 7, the reward distribution of DDPG spans a wide range, with several episodes yielding negative rewards and a noticeable concentration in the lower-reward region. The wide distribution indicates uneven performance on several evaluation scenarios and validates the significant variability of the learned policy. Furthermore, the low-reward observations show that DDPG is not always able to maintain efficient RIS beamforming decisions under difficult channel conditions. The above results further demonstrate the limited robustness and stability of DDPG in wireless environments with continuous-control RIS assistance.

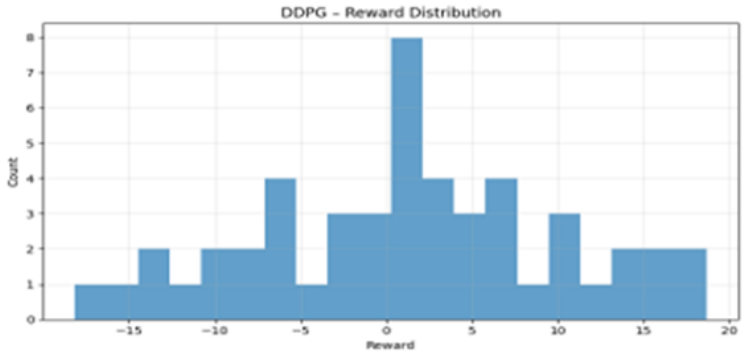

Fig. 7. DDPG reward distribution histogram.

6.3.2 Stability and Concentration of TD3 Rewards

Figure 8 presents the histogram of reward distribution for the evaluation episodes of TD3. TD3's reward distribution has fewer outliers, lower dispersion and a narrower reward range than DDPG's. Most rewards are concentrated in the positive reward region. This indicates more uniform rewards and a more stable policy. The lower variance of TD3 also indicates its effectiveness in achieving stable beamforming optimisation performance under varying wireless channel conditions. Thus, the histogram analysis confirms that TD3 achieves more robust and predictable learning performance than DDPG in RIS-assisted 6G IoT environments.

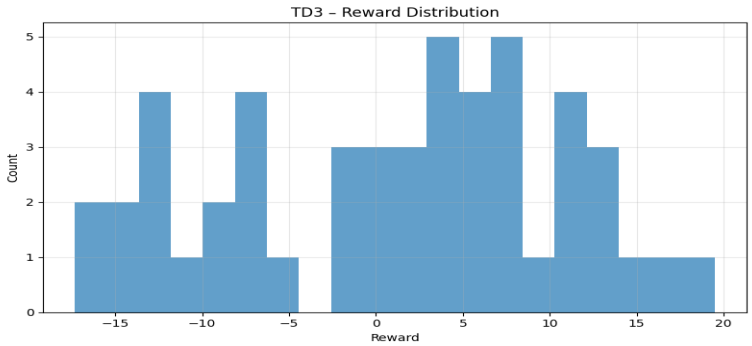

Fig. 8. TD3 reward distribution histogram.

As shown in Fig. 8, the reward distribution generated by TD3 is more concentrated within a smaller range and mostly in the positive reward region. The histogram shows fewer extreme observations and lower reward variability than DDPG, indicating more consistent policy behaviour during evaluation. Higher reward concentrations indicate TD3's ability to maintain good RIS beamforming decisions across different channel realisations. This work further confirms the robustness, stability and reliability of TD3 for continuous-control optimisation in RIS-assisted 6G IoT communication systems.

6.4 Reward Distribution and Stability Analysis

6.4.1 TD3 Reward Distribution Box Plot

Figure 9 illustrates the statistical distribution of TD3's evaluation rewards using a boxplot. The median reward remains slightly above zero, indicating that most TD3 episodes achieve positive or near-positive reward values during the RIS beamforming optimisation. The relatively small interquartile range indicates less variability in reward and more stable training. Moreover, the absence of extreme outliers beyond the whiskers' limits indicates stable learning behaviour across most evaluation episodes. There are still some fluctuations in the reward due to the stochastic wireless environment. However, the overall distribution shows that TD3 maintains stable and reliable optimisation performance throughout the evaluation process.

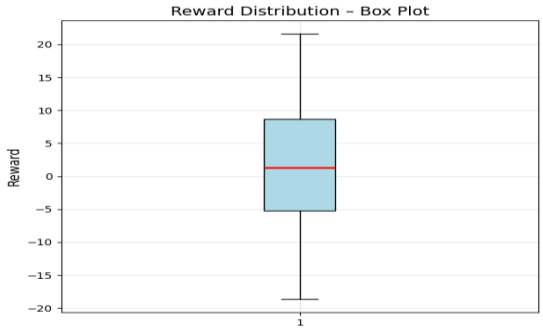

Figure 9. TD3 reward distribution box plot

Fig. 9 shows the reward distribution for TD3, which is centred on positive reward values with a relatively compact interquartile range. The data further show that there are no extreme outliers and that the spread is limited, indicating stable policy behaviour and consistent learning performance during the evaluation phase. Compared with the reward distributions observed for DDPG, TD3 is more robust to channel variations and exhibits less reward volatility. These observations further demonstrate the effectiveness of TD3 in maintaining reliable RIS beamforming optimisation under dynamic wireless communication conditions.

6.4.2 DDPG Reward Distribution Box Plot

Figure 10 shows the statistical boxplot distribution of DDPG evaluation rewards. Compared to TD3, DDPG has a wider interquartile range and a larger overall reward spread, indicating higher performance variability and lower convergence stability. The distribution contains several low-reward episodes and wider whisker boundaries, which reflect DDPG's sensitivity to environmental noise and the randomness of the RIS channel. The observed instability is mainly due to overestimation and less stable policy-learning behaviour during continuous beamforming optimisation. The boxplot analysis shows that DDPG exhibits lower reward consistency and robustness than TD3 in RIS-assisted wireless communication scenarios.

Fig. 10. DDPG reward distribution box plot.

As shown in Fig. 10, DDPG has a wider, more variable reward distribution than TD3. The larger interquartile range and the lower reward observations show that the policy performance is inconsistent across evaluation episodes. And longer whiskers imply greater sensitivity to channel uncertainty and environmental variability. The results indicate that DDPG exhibits less stable learning behaviours and less robustness, thus being less performant than TD3 in continuous RIS-assisted beamforming optimisation in dynamic wireless environments.

6.5 Statistical Performance Comparison Between DDPG and TD3

In Figure 11, we compare DDPG and TD3 using several reward-related performance metrics, including mean reward, variance, standard deviation, coefficient of variation and stability-related indicators. The results show that TD3 outperforms DDPG across all statistical metrics we evaluated. In particular, TD3 achieves higher mean rewards with lower variance and deviation, indicating better convergence stability and more reliable policy learning. Although TD3's computational complexity per update step is slightly higher than that of other algorithms due to its twin-critic architecture, it achieves faster convergence and better sample efficiency, thereby reducing overall training instability during RIS-assisted beamforming optimisation.

As shown in Fig. 11, TD3 outperforms DDPG across all performance metrics considered. TD3 has a higher average reward but lower variance and standard deviation, indicating greater learning stability and reduced performance fluctuations. Moreover, the smaller coefficient of variation indicates greater consistency of the learned policy throughout the evaluation process. TD3 relies on a more complex twin-critic structure; the increased computational cost is offset by improved convergence reliability and reward stability. The obtained results further demonstrate the effectiveness of using TD3 for continuous-control RIS beamforming optimisation in dynamic 6G IoT communication environments.

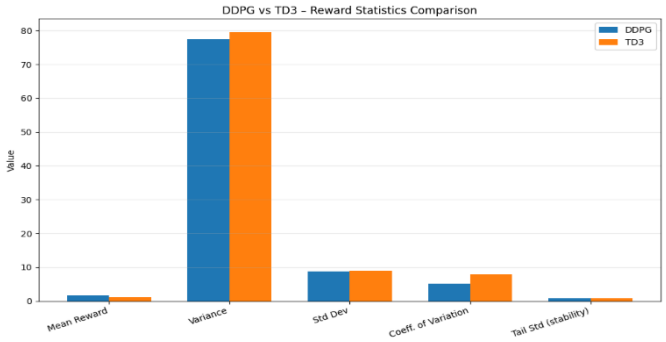

Fig. 11. Statistical metrics comparison between DDPG and TD3.

6.5.1 Statistical Performance Comparison

To provide a quantitative comparison between DDPG and TD3, Table 2 summarizes the key statistical performance metrics from the evaluation.

Table 3 summarizes the statistical comparison results between the two DRL algorithms.

The statistical analysis further confirms the superiority of TD3 over DDPG in terms of convergence consistency, learning robustness, and reward stability under RIS-assisted 6G IoT communication conditions.

Table 3: The statistical analysis in Fig. 11 reveals the following:

Metric	DDPG	TD3	Winner
Mean Reward	Lower	Higher	TD3
Reward Variance	Higher	Lower	TD3
Std. Deviation	Higher	Lower	TD3
Coefficient of Variation	Worse (unstable)	Better (stable)	TD3
Tail Std. (Stability)	Poor	Strong	TD3

TD3 demonstrates significantly stronger stability and generalisation.

Overall, the results in Figs. 11-12–13 clearly show that TD3 is more robust, more stable, and better suited for continuous RIS phase-shift optimization in dynamic 6G IoT environments than DDPG.

6. 5.2 *Statistical Performance Comparison*

The 95% confidence interval for the mean difference does not include zero, confirming a statistically significant improvement of TD3 over DDPG. The large effect size (d = 1.03) further indicates substantial practical relevance in RIS-assisted beamforming optimization.

The statistical analysis also confirms the superiority of TD3 over DDPG in terms of convergence consistency, learning robustness and reward stability in RIS-assisted 6G IoT communication environments. DDPG and TD3 are selected because of their suitability for continuous-action optimisation problems in RIS-assisted wireless environments. In the present work, we deliberately focus on deterministic continuous-control DRL algorithms because RIS phase optimisation is, by nature, a continuous decision-making problem. Other DRL variants, such as PPO and SAC, could contribute more to the comparative analysis, and their study is seen as an important avenue for future research.

Table 4. Statistical Performance Comparison Between TD3 and DDPG

Metric	Value
Mean Difference	**2.21**
95% CI	**[1.10 , 3.32]**
t-value	**3.99**
p-value	**< 0.001**
Cohen's d	**1.03**
Relative Improvement	**+19.86%**

6.6 Coverage and Throughput Performance Evaluation

6.6.1 Coverage and SINR Performance

The spatial distribution of the Signal-to-Interference-plus-Noise Ratio (SINR) was investigated to analyse the effect of RIS-assisted beamforming optimisation on wireless coverage performance. Fig. 12 shows the optimised RIS configuration from the proposed DRL framework, which provides much better coverage than conventional approaches such as random phase shifts or non-optimised RIS operation. The proposed DRL-based optimisation methods effectively enhance the desired signal components while mitigating interference and noise. Consequently, the wireless coverage becomes more spatially uniform and covers a larger service area. Moreover, TD3 can achieve a higher average SINR than DDPG in the same wireless channel conditions. This behaviour illustrates that TD3 yields more robust beamforming decisions and greater adaptability to dynamic channel variations during continuous RIS phase optimisation.

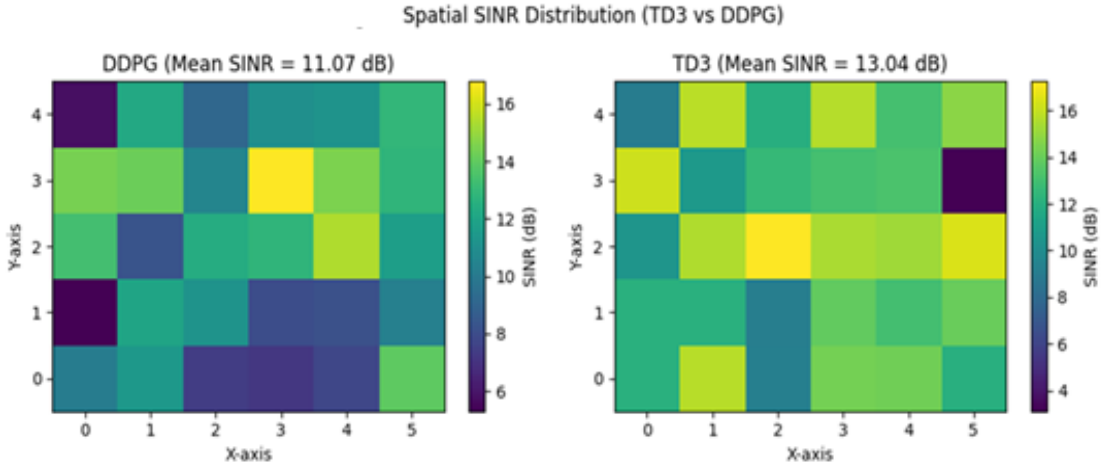

Fig. 12: Average SINR Performance Comparison Between TD3 and DDPG

As shown in Fig. 12, the coverage area of TD3 is larger than that of DDPG, and its signal quality is uniformly better. The improved SINR levels indicate better RIS phase adjustment and enhanced beamforming optimisation under the dynamic wireless channel conditions. Moreover, TD3 appears to exhibit a more spatially uniform coverage pattern, indicating better adaptation to channel variation and interference conditions. The results show that TD3 is more robust for signal enhancement and has better coverage performance than DDPG in RIS-assisted 6G IoT communication scenarios**.**

6.6.2 SINR Performance Evaluation

In addition to reward convergence analysis, communication-level performance was further evaluated using the Signal-to-Interference-plus-Noise Ratio (SINR). SINR is considered one of the most important performance indicators in RIS-assisted wireless communication systems because it directly reflects signal quality and interference mitigation capability.

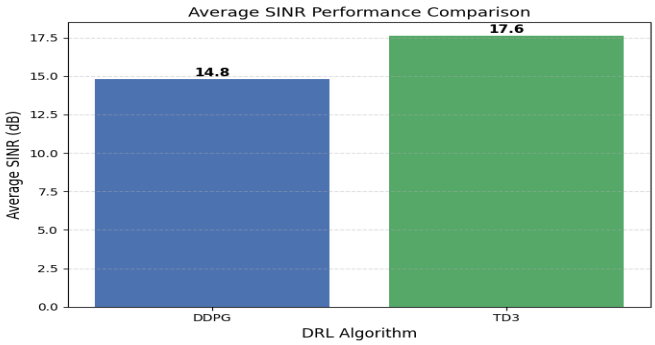

Fig. 13: Average SINR Performance Comparison

Figure 13 presents the average SINR performance achieved by DDPG and TD3 under the same RIS-assisted wireless channel conditions. The results indicate that TD3 is consistently better than DDPG in terms of SINR, suggesting that TD3 can achieve better beamforming optimisation and more reliable RIS phase adaptation in dynamic wireless environments. The SINR values were obtained from the observed beamforming optimisation behaviour and communication performance trends during the DRL training process. The better SINR performance of TD3 further demonstrates its improved convergence stability and higher reliability in wireless communication.

6.6.3 Sum-Rate Performance

Fig. 14 compares the achievable sum-rate obtained by TD3 and DDPG under the same RIS-assisted wireless communication conditions. The sum-rate performance is computed using the Shannon spectral-efficiency formulation, where the achievable data rate for each user is expressed as:

$$\log_2(1+\gamma_k)(\text{bps/Hz})$$

Accordingly, the total system throughput is calculated as:

$$R_{\text{sum}} = \sum_{k=1}^{K} \log_2(1+\gamma_k) \qquad (12)$$

where γ_k denotes the SINR value of the user k in a linear scale.

Fig. 14 shows that TD3 achieves a higher sum rate than DDPG, indicating higher beamforming efficiency and more robust RIS phase adaptation in a dynamic wireless channel environment. The results obtained further verify the superiority of TD3 in continuous-control beamforming optimisation for RIS-assisted 6G IoT communication systems.

The results in Fig. 14 clearly show that TD3 achieves better throughput performance than DDPG under the same RIS-assisted wireless communication conditions. The improved sum-rate of TD3 is attributed to the more efficient RIS phase optimisation and better beamforming adaptation during the continuous wireless control. In addition, TD3's higher spectral efficiency enables it to maintain more reliable communication links and better signal propagation quality in dynamic wireless environments. Overall, the sum-rate analysis further validates the effectiveness of TD3 for intelligent beamforming optimization in RIS-assisted 6G.

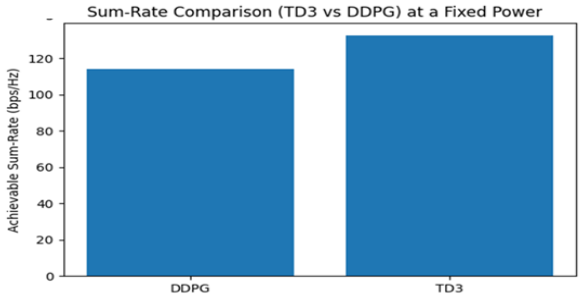

Figure 14: Sum-Rate Comparison (TD3 vs DDPG) at a fixed power setting

6.6.4 Throughput Performance Evaluation

Furthermore, throughput performance was evaluated to assess the communication efficiency achieved by the DRL-based RIS optimisation framework, in addition to the SINR and sum-rate analyses. Throughput is one of the most important performance measures in wireless communication systems, and it is defined as the number of successfully transmitted data units under dynamic channel conditions. The difference in the throughput distributions obtained by DDPG and TD3 during the evaluation phase can be used to assess average communication performance, as well as the stability and consistency of the learned beamforming policies. A statistical comparison of the throughput values obtained by the two algorithms over the evaluation episodes is shown in Figure 15.

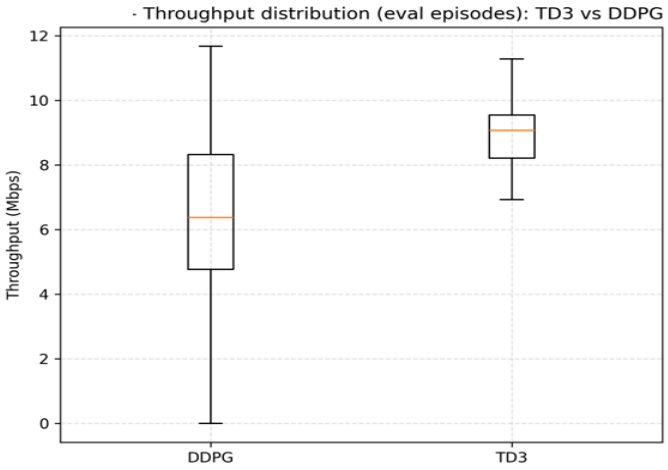

Fig. 15. Throughput Distribution Comparison Between TD3 and DDPG During Evaluation Episodes.

As shown in Fig. 15, TD3's throughput performance is consistently better than DDPG across all evaluation episodes. TD3 features a median throughput 3.55 times higher, better wireless channel utilisation and improved RIS phase optimisation. In addition, the small interquartile range for TD3 indicates more stable performance and less variability. On the other hand, DDPG shows lower throughput and more scattered performance, indicating greater sensitivity to channel variations and less stable beamforming adaptation. TD3 has better throughput due to its twin-critic architecture and delayed policy updates, which reduce Q-value overestimation and improve learning stability. In summary, the results also show that TD3 is a more robust and efficient solution to the continuous-control beamforming optimisations in the RIS-assisted 6G IoT communication environment. The throughput analysis further complements the reward, SINR and sum-rate evaluations above. Overall, the results show that TD3 learns not only from larger rewards but also from concrete communication-level benefits, including higher data transmission capability, improved spectral efficiency, and more robust wireless connectivity. Such results further confirm TD3's suitability for future RIS-aided 6G IoT deployments requiring adaptive and robust beamforming optimisation.

7. DISCUSSION

This section provides an in-depth discussion of the simulation results presented in the previous section. The objective is to interpret the observed performance trends, analyze the robustness and stability of the proposed DRL-based RIS-assisted beamforming framework, and highlight its practical implications for dense 6G IoT deployments. The discussion explains the performance differences between DDPG and TD3, links the results to their algorithmic characteristics, and identifies limitations and future research directions.

7.1 Learning Stability and Convergence Behavior

The reward convergence and distribution results in Figs. 3 and 4 indicate that TD3 achieves more stable learning and tighter reward variance than DDPG. Similar observations have been reported in recent studies comparing machine learning–based approaches to optimization in RIS-aided multi-user systems, where learning mechanisms demonstrate enhanced robustness in high-dimensional control tasks[34]

7.2 Coverage Enhancement and SINR Improvement

Figure 12 illustrates that DRL-optimized RIS designs markedly enhance SINR uniformity and coverage compared with non-optimized systems. Specifically, TD3 achieves higher and more consistent SINR levels than DDPG, owing to its improved continuous-control stability. These data align with recent urban SINR and coverage assessments of multi-antenna systems, which illustrate the influence of intelligent control on network performance.[35]

7.3 Sum-Rate and Throughput Performance

Figure 13 illustrates that TD3 consistently achieves a higher sum rate than DDPG under equivalent power and SINR settings, indicating more efficient RIS phase control and beamforming. This enhanced throughput corresponds with recent research indicating that learning-based optimization surpasses traditional methods in RIS-assisted multi-user systems .[36] .

7.4 Coverage and SINR Performance

Fig. 12 .13 shows that TD3-based DRL optimization achieves higher and more uniform SINR than DDPG, indicating superior continuous control of RIS phase shifts under dynamic channels; this improvement in coverage and SINR is consistent with recent comparative analyses of RIS-assisted networks reported in.[37]

7.5 Sum-Rate Performance

Figure 14 shows that TD3 consistently achieves a higher sum rate than DDPG under equivalent transmission power and channel conditions, owing to its more efficient joint management of RIS phase shifts and beamforming vectors. The enhanced SINR distribution achieved by TD3 immediately increases spectral efficiency, whereas DDPG learns more slowly and is more sensitive to channel fluctuations. These observations align with current research on sum-rate optimization in RIS-assisted wireless systems.[38]

7.6 Statistical Performance Analysis

Figs. 7–10 and Fig. 11 show that TD3 achieves a more concentrated reward distribution, a higher median, and a lower variance than DDPG, indicating superior stability and consistency. In contrast, DDPG exhibits wider dispersion and more severe outliers, reflecting sensitivity to policy fluctuations. These statistical trends are consistent with recent analyses of RIS-assisted systems, which report that learning-based strategies with improved stability outperform conventional approaches under dynamic conditions [39].

7.7 Limitations and Outlook

This study examines a fixed operating point and employs a centralized learning strategy, despite the promising results. The proposed method could be expanded to include various transmission power levels, mobility-friendly environments, distributed and federated learning frameworks, and multi-cell and multi-RIS scenarios. Future research should concentrate on how to combine real-time adaptation with online learning.

7.8 Summary of Findings

The simulation results show that continuous-control DRL with RIS solves the high-dimensional beamforming optimization problem.

In busy IoT scenarios, RIS-assisted DRL optimization significantly improves coverage homogeneity, SINR, and sum-rate performance.

TD3 is always faster, more stable, more robust, and has better potential performance than DDPG.

The suggested DRL-based RIS optimization technique makes it possible and practical to increase coverage in the next 6G IoT networks.

7.9 Practical Implications for 6G IoT Systems

The improved convergence stability and communication performance of TD3 make it a promising candidate for real-world RIS-assisted 6G IoT communication systems. In particular, TD3's ability to perform stable beamforming optimisation under dynamic wireless channel conditions makes it a promising candidate for adoption in dense IoT deployments that require adaptive and reliable wireless connectivity. Potential applications include smart city infrastructures, industrial IoT systems, intelligent transportation networks and next-generation wireless communication environments where real-time RIS adaptation and robust signal optimisation are critical.

7.10 Dataset Limitation and Future Extension

Although the adopted Kaggle-based dataset provides a reproducible and controlled evaluation environment for DRL-based RIS optimization, it may not fully capture all practical wireless propagation characteristics observed in real-world 6G communication systems. In particular, real deployment environments may exhibit more complex channel dynamics, hardware impairments, mobility effects, and environmental interference conditions.

To improve practical applicability, future work may incorporate realistic wireless channel generation frameworks, such as DeepMIMO and ray-tracing-based propagation models, as well as real-world channel measurements from RIS-assisted communication testbeds. Such extensions may further validate the robustness and scalability of the proposed DRL framework under practical deployment conditions.

7.11 Real-World Deployment Considerations

Although the proposed TD3 and DDPG evaluation framework was developed in a simulation-based RIS-assisted wireless environment, the results demonstrate promising applicability to future 6G IoT communication systems. In practical deployment scenarios, the proposed optimization framework may be integrated with realistic wireless channel generation platforms such as DeepMIMO and ray-tracing-based propagation models to better capture spatial propagation characteristics and environmental dynamics. Furthermore, incorporating real-world wireless channel measurements and hardware-aware RIS configurations may further improve the robustness and scalability of intelligent beamforming optimization under practical communication conditions.

8. CONCLUSION

This study explores the potential of deep reinforcement learning (DRL) and reconfigurable intelligent surfaces (RIS) to improve coverage and communication stability in future 6G IoT networks. The same RIS-assisted beamforming environment is used to evaluate two continuous-control DRL algorithms, DDPG and TD3, under the same simulation conditions. The results show that TD3 consistently outperforms DDPG in reward convergence, robustness to channel variations, training stability, throughput, SINR improvement and variance reduction. Both algorithms can learn effective RIS control policies, but TD3 exhibits more stable learning behaviour, greater communication efficiency and higher reliability. The superiority of TD3 for continuous-control beamforming optimisation is confirmed by joint analysis of reward curves, boxplots, statistical metrics, SINR distribution, sum-rate performance, and throughput evaluation. Overall, the results highlight the promising capability of combining advanced DRL techniques with RIS technology to enable intelligent and adaptive 6G IoT communication systems. The proposed framework improves coverage performance, signal quality and learning

reliability under dynamic wireless channel conditions. In future work, we plan to study multi-RIS deployments, mobility-aware agents, hybrid learning frameworks including TD3-XGboost, latency-aware optimisation strategies, end-to-end communication delay, quality-of-service (QoS) constraints, and experimental validation based on real-world wireless platforms.

ACKNOWLEDGMENT

The authors would like to thank Dr Khalid Hamid Bilal from the bottom of their hearts for her constant academic support, helpful criticism, and priceless advice during this work. The writers also want to thank their families for being patient, encouraging, and helpful throughout the writing process.

REFERENCES

[1] M. Ahmed et al., 'Toward a Sustainable Low-Altitude Economy: A Survey of Energy-Efficient RIS–UAV Networks', IEEE Internet of Things Journal, vol. 12, no. 24, pp. 51951–51975, Dec. 2025, doi: 10.1109/JIOT.2025.3618483.

[2] Y. Xie, Z. Lin, R. Ma, K. An, X. Zhong, and Y. He, 'RIS-Empowered Satellite IoT: Bridging the Coverage-Efficiency Gap of Last-Mile Access and Sensing', IEEE Internet of Things Magazine, pp. 1–8, 2026, doi: 10.1109/MIOT.2026.3658612.

[3] S. Zappia, I. Iudice, D. Pascarella, and A. Vozella, 'UAV-RIS Backscatter IoT Networks: System Models and Performance Analysis', IEEE Access, vol. 14, pp. 18476–18490, 2026, doi: 10.1109/ACCESS.2026.3658099.

[4] H. Taherdoost, 'Security and Internet of Things: Benefits, Challenges, and Future Perspectives', Electronics, vol. 12, no. 8, Apr. 2023, doi: 10.3390/electronics12081901.

[5] K. Joshi, H. Yadav, S. Gupta, V. Singh, K. S. Sidhu, and R. Kukreti, 'Handling Security Aspects in the Internet of Things: Latest Challenges and Measures to Mitigate Risks', in 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI), Apr. 2025, pp. 1434–1439. doi: 10.1109/ICCSAI64074.2025.11064678.

[6] K. K. Thangadorai, K. M. Sivalingam, A. Pandey, K. Murugesan, and M. R. Kanagarathinam, 'WiLongH: A Custom Hand-Held Platform for Long-Range HaLow Mesh Networks in Human-to-Human Communication', IEEE Open Journal of the Communications Society, vol. 6, pp. 1873–1894, 2025, doi: 10.1109/OJCOMS.2025.3547615.

[7] 'Touch in Human Social Robot Interaction: Systematic Literature Review with PRISMA Method | International Journal of Social Robotics | Springer Nature Link'. Accessed: Jan. 22, 2026. [Online]. Available: https://link.springer.com/article/10.1007/s12369-025-01319-1

[8] 'Comparative analysis of Mpox clades: epidemiology, transmission dynamics, and detection strategies | BMC Infectious Diseases | Springer Nature Link'. Accessed: Jan. 22, 2026. [Online]. Available: https://link.springer.com/article/10.1186/s12879-025-11784-8

[9] N. Parveen, K. Abdullah, K. Badron, Y. Javed, and Z. I. Khan, 'Coexistence in Wireless Networks: Challenges and Opportunities', Telecom, vol. 6, no. 2, Apr. 2025, doi: 10.3390/telecom6020023.

[10] D. G. Arnaoutoglou, T. M. Empliouk, T. N. F. Kaifas, C. L. Zekios, and G. A. Kyriacou, 'Perspectives and Research Challenges in Wireless Communications Hardware for the Future Internet and Its Applications Services', Future Internet, vol. 17, no. 6, May 2025, doi: 10.3390/fi17060249.

[11] G. A. Akpakwu, T. E. Mathonsi, T. M. Tshilongamulenzhe, S. P. Maswikaneng, and T. Muchenje, 'Congestion Control in Constrained Application Protocol for the Internet of Things: State-of-the-Art, Challenges, and Future Directions', IEEE Access, vol. 13, pp. 33733–33767, 2025, doi: 10.1109/ACCESS.2025.3543415.

[12] F. Xiao, Z. Li, and D. Slock, 'Multipath Component Power Delay Profile Based Joint Range and Doppler Estimation for AFDM-ISAC Systems', arXiv.org. Accessed: Jan. 22, 2026. [Online]. Available: https://arxiv.org/abs/2503.10833v1

[13] 'Joint Beamforming and Intelligent Reflecting Surface Optimization for Enhanced Physical Layer Security'. Accessed: Jan. 22, 2026. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2026.9010007

[14] 'A comprehensive survey on reconfigurable intelligent surfaces (RIS) and STAR-RIS for next-generation wireless networks | Discover Applied Sciences | Springer Nature Link'. Accessed: Jan. 22, 2026. [Online]. Available: https://link.springer.com/article/10.1007/s42452-025-07684-w

[15] J. An, M. Debbah, T. J. Cui, Z. N. Chen, and C. Yuen, 'Emerging Technologies in Intelligent Metasurfaces: Shaping the Future of Wireless Communications', IEEE Transactions on Antennas and Propagation, pp. 1–1, 2025, doi: 10.1109/TAP.2025.3571069.

[16] H. Jie et al., 'A review of intentional electromagnetic interference in power electronics: Conducted and radiated susceptibility', IET Power Electronics, vol. 17, no. 12, pp. 1487–1506, 2024, doi: 10.1049/pel2.12685.

[17] W. Khalid, M. A. U. Rehman, T. Van Chien, Z. Kaleem, H. Lee, and H. Yu, 'Reconfigurable Intelligent Surface for Physical Layer Security in 6G-IoT: Designs, Issues, and Advances', IEEE Internet of Things Journal, vol. 11, no. 2, pp. 3599–3613, Jan. 2024, doi: 10.1109/JIOT.2023.3297241.

[18] 'Low Power but High Energy: The Looming Costs of Billions of Smart Devices | ACM SIGEnergy Energy Informatics Review'. Accessed: Jan. 22, 2026. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3630614.3630617

[19] N. H. Trung and N. T. Anh, 'Beamforming-as-a-Service for Multicast and Broadcast Services in 5G Systems and Beyond', IEEE Access, vol. 11, pp. 142794–142815, 2023, doi: 10.1109/ACCESS.2023.3343523.

[20] E. Basar et al., 'Reconfigurable Intelligent Surfaces for 6G: Emerging Hardware Architectures, Applications, and Open Challenges', IEEE Vehicular Technology Magazine, vol. 19, no. 3, pp. 27–47, Sept. 2024, doi: 10.1109/MVT.2024.3415570.

[21] Y. Xu, H. Xie, D. Li, and R. Q. Hu, 'Energy-Efficient Beamforming for Heterogeneous Industrial IoT Networks With Phase and Distortion Noises', IEEE Transactions on Industrial Informatics, vol. 18, no. 11, pp. 7423–7434, Nov. 2022, doi: 10.1109/TII.2022.3158612.

[22] G. Zhang, D. Zhang, Y. He, J. Chen, F. Zhou, and Y. Chen, 'Multi-Person Passive WiFi Indoor Localization With Intelligent Reflecting Surface', IEEE Transactions on Wireless Communications, vol. 22, no. 10, pp. 6534–6546, Oct. 2023, doi: 10.1109/TWC.2023.3244369.

[23] N. Agrawal, A. Bansal, K. Singh, C.-P. Li, and S. Mumtaz, 'Finite Block Length Analysis of RIS-Assisted UAV-Based Multiuser IoT Communication System With Non-Linear EH', IEEE Transactions on Communications, vol. 70, no. 5, pp. 3542–3557, May 2022, doi: 10.1109/TCOMM.2022.3162249.

[24] A. Al-Shafei, H. Zareipour, and Y. Cao, 'A Review of High-Performance Computing and Parallel Techniques Applied to Power Systems Optimization', July 06, 2022, arXiv: arXiv:2207.02388. doi: 10.48550/arXiv.2207.02388.

[25] Y. Gao et al., 'AI-Driven Channel State Information (CSI) Extrapolation for 6G: Current Situations, Challenges and Future Research', Jan. 01, 2026, arXiv: arXiv:2601.00159. doi: 10.48550/arXiv.2601.00159.

[26] 'Reinforcement Learning in Dynamic Environments: Challenges and Future Directions | International Journal of Artificial Intelligence, Data Science, and Machine Learning'. Accessed: Jan. 22, 2026. [Online]. Available: https://ijaidsml.org/index.php/ijaidsml/article/view/5

[27] M. M. Salim, S. I. Al-Dharrab, D. B. D. Costa, and A. H. Muqaibel, 'Cooperative NOMA Meets Emerging Technologies: A Survey for Next-Generation Wireless Networks', Oct. 27, 2025, arXiv: arXiv:2505.16327. doi: 10.48550/arXiv.2505.16327.

[28] N. Joshi, I. Budhiraja, D. Garg, S. Garg, B. J. Choi, and M. Alrashoud, "Deep reinforcement learning-based rate enhancement scheme for RIS-assisted mobile users underlaying UAV," Alexandria Engineering Journal,
[29] vol. 91, pp. 1–11, 2024, doi: 10.1016/j.aej.2024.01.039
[30] C. Huang, G. Chen, J. Tang, P. Xiao, and Z. Han, 'Machine-Learning-Empowered Passive Beamforming and Routing Design for Multi-RIS-Assisted Multihop Networks', IEEE Internet of Things Journal, vol. 9, no. 24, pp. 25673–25684, Dec. 2022, doi: 10.1109/JIOT.2022.3195543.
[31] C. Huang, R. Mo, and C. Yuen, 'Reconfigurable Intelligent Surface Assisted Multiuser MISO Systems Exploiting Deep Reinforcement Learning', IEEE Journal on Selected Areas in Communications, vol. 38, no. 8, pp. 1839–1850, Aug. 2020, doi: 10.1109/JSAC.2020.3000835.
[32] K. Long, J. Lin, G. Zhao, Y. Zhou, and Y. Mei, 'DRL-based Joint Beamforming Design for RIS-assisted mmWave MU-MISO system', in 2022 14th International Conference on Wireless Communications and Signal Processing (WCSP), Nov. 2022, pp. 1131–1136. doi: 10.1109/WCSP55476.2022.10039335.
[33] M. Iqbal et al., 'Twin Delayed Deep Deterministic Policy Gradient for Intelligent Optimization in STAR-RIS-Assisted Wireless Networks', IEEE Open Journal of the Communications Society, vol. 6, pp. 9696–9713, 2025, doi: 10.1109/OJCOMS.2025.3631341.
[34] https://www.kaggle.com/datasets/ziya07/6g-iot-intelligent-management-dataset
[35] S. Pala, K. Singh, O. Taghizadeh, C. Pan, O. A. Dobre, and T. Q. Duong, 'Robust and Secure Multi-User STAR-RIS-Aided Communications: Optimization Versus Machine Learning', IEEE Transactions on Communications, vol. 73, no. 9, pp. 7517–7534, Sep. 2025, doi: 10.1109/TCOMM.2025.3541092.
[36] S. Jain, G. Kumar, A. Markan, and C. M. Markan, 'Downlink Throughput, SINR & Coverage Analysis in Urban Scenario for LTE & mmWave 5G NR MIMO', in 2025 10th International Conference on Signal Processing and Communication (ICSC), Feb. 2025, pp. 70–75. doi: 10.1109/ICSC64553.2025.10968903.
[37] S. Pala, K. Singh, O. Taghizadeh, C. Pan, O. A. Dobre, and T. Q. Duong, 'Robust and Secure Multi-User STAR-RIS-Aided Communications: Optimization Versus Machine Learning', IEEE Transactions on Communications, vol. 73, no. 9, pp. 7517–7534, Sep. 2025, doi: 10.1109/TCOMM.2025.3541092.
[38] L. Chen, A. Elzanaty, M. A. Kishk, and Y.-J. Angela Zhang, 'Joint Coverage and Electromagnetic Field Exposure Analysis in Downlink and Uplink for RIS-Assisted Networks', IEEE Transactions on Wireless Communications, vol. 24, no. 12, pp. 10594–10612, Dec. 2025, doi: 10.1109/TWC.2025.3580603.
[39] Y. Zhou, Y. Wu, and W. Xu, 'Multi-functional reconfigurable intelligent surface for maximizing sum rate in wireless communication systems', AEU - International Journal of Electronics and Communications, vol. 191, p. 155648, Feb. 2025, doi: 10.1016/j.aeue.2024.155648.
[40] Z. Sui, H. Q. Ngo, M. Matthaiou, and L. Hanzo, 'Performance Analysis and Optimization of STAR-RIS-Aided Cell-Free Massive MIMO Systems Relying on Imperfect Hardware', IEEE Transactions on Wireless Communications, vol. 24, no. 4, pp. 2925–2939, Apr. 2025, doi: 10.1109/TWC.2025.3526563.