*International Journal of Computational and Experimental Science and ENgineering (IJCESEN)*

*Vol. 11-No.3 (2025) pp. 4603-4614*
*http://www.ijcesen.com*

***ISSN: 2149-9144***



**Research Article**

# Distributed Reinforcement Learning Based Efficient Radio Resource Allocation for Heterogeneous IoT Wireless Network

**Abhishek Kumar Verma[1*], Vinay Kumar Singh[2], M. R. Khan[3]**

[1]Jhada Sirha Government Engineering College, Jagdalpur, Chhattisgarh, India
*** Corresponding Author Email:** akverma@gecjdp.ac.in – **ORCID:** 0000-0001-6919-9773

[2] Jhada Sirha Government Engineering College, Jagdalpur, Chhattisgarh, India
**Email:** vinay.rewa@gmail.com - **ORCID:** 0000-0001-8642-2885

[3] Government Engineering College, Raipur, Chhattisgarh, India.
**Email:** mrkhan@gecjdp.ac.in – **ORCID:** 0009-0000-0574-6053

**Article Info:**

**DOI:** 10.22399/ijcesen.2852
**Received :** 11 April 2025
**Accepted :** 19 June 2025

**Keywords**

Internet of Things (IoT)
Power Allocation
Heterogeneous Network
Reinforcement Learning
Quality of Service (QoS)

**Abstract:**

The heterogeneous cellular network has emerged as a critical infrastructure in supporting diverse Internet of Things (IoT) based services. As next-generation technologies continue to evolve, they will provide a unified framework capable of seamlessly connecting huge number of IoT devices. This integration will support the complex requirements of modern IoT-driven business processes. To achieve maximum capacity for IoT applications there is a need to integrate multiple wireless network technologies which makes the environment dense. However, this network densification also raises interference levels from neighbouring devices which can negatively impact the Quality of Service (QoS). In a dense IoT environment with limited radio resources there is need of efficient radio resource allocation strategy to maintain QoS across various connected devices. This work presents a distributed reinforcement learning based power allocation algorithm for heterogeneous IoT networks. We also propose a reward function that accounts for the QoS needs of multiple IoT users, promotes fairness, and ensures reliable connectivity. We carry out complexity analysis and convergence analysis of our proposed algorithm and also we explore different learning frameworks to evaluate the performance of the algorithm. Results demonstrated that the proposed method is effective in improving network capacity and other performance measures in dense heterogeneous environment

## 1. Introduction

The rapid advancement of the Internet of Things (IoT) has driven a substantial increase in devices utilizing wireless networks to support diverse applications across sectors like smart cities, healthcare, energy, and industry. Forecasts indicate that the global number of IoT devices could reach to billions by 2030 [2]. To meet the varying needs of the increasing number of IoT devices it is necessary to integrate multiple wireless networks. This integration creates dense heterogeneous network architecture with numerous base stations, access points, and devices within the coverage region with an objective to ensure seamless connectivity and efficient service delivery across a wide range of applications. However, such densification can cause higher interference due to overlapping frequencies thereby impacting network reliability. To address the growing demands of reliable and low latency services from IoT devices, one of the possible solutions is to deploy heterogeneous networks where small cells are overlaid within the coverage area of macro cells. In dense heterogeneous networks interference is a limiting factor which potentially affects the performance. Hence, an effective power allocation strategy is essential to enhance network capacity while meeting QoS requirements, ensuring fairness, and keeping SINR above the threshold for reliable connectivity across all IoT devices. This process involves dynamically adjusting transmission power to reduce interference and improve overall performance.

Various power allocation methods for heterogeneous wireless networks have been studied in the literature and are generally classified into optimization-based and learning-based techniques. Optimization-based techniques require full channel

state information (CSI) for resource management and also it is difficult to acquire CSI in dynamic environment [3]–[10]. On the other hand, learning-based approaches provide a flexible alternative by using adaptive techniques to optimize power allocation without the need for complete channel information. These methods adjust themselves according to evolving network conditions and can operate in distributed or decentralized frameworks which make them well-suited for dynamic and large scale IoT networks. In [3], Peng et al. addressed an energy-efficient power allocation in heterogeneous access networks using a convex optimization framework. Similarly, Zhang et al. [4] tackled the power allocation challenge in dense heterogeneous wireless networks by formulating it as a convex problem and through this method, they obtained analytical solutions to effectively manage resources and optimize user associations. Farooq et al. [5] presented a method to reduce interference in two-layer heterogeneous IoT networks by jointly optimizing uplink user association and power allocation along with maintaining the QoS requirement. In [6], Song et al. framed the issue as a multi-objective optimization problem to balance between energy efficiency and QoS in IIoT. Wang *et al.* [7] formulated the power control optimization problem for UAV networks. In [8], Bakht et al. proposed a cognitive radio-based approach for jointly managing power control and user assignment in 5G heterogeneous networks. Ha et al. [9] addressed the uplink channel and power allocation using a distributed and low-complexity algorithm that outperformed traditional exhaustive methods. Likewise, Kai et al. [10] developed a framework for channel allocation and power control in D2D communication to enhance network capacity.

For learning-based techniques, such as RL-based power allocation approaches, they are particularly effective in dynamic environments because of their ability to adapt and respond to changing conditions. These algorithms continuously interact with environment and update their strategies through feedback which make them well-suited for changing environment. In [11], Saad et al. presented a distributed architecture based cooperative Q-learning algorithm for power allocation, focusing on ensuring the QoS of macrocell users and does not take into account the QoS needs of femtocell users. A Q-learning-based power allocation approach is developed to improve overall network capacity while meeting QoS demands and their method also focuses on balanced resource distribution between macrocell and femtocell users [12, 13]. Shahid et al. [14] presented a docitive Q-learning approach for power allocation in heterogeneous architecture and performance is evaluated against independent Q-learning method. Zhang et al. [15] developed a learning-based power control method for networks with randomly distributed base stations which aims to maximize system capacity and their approach outperformed the traditional water-filling power allocation technique. Meng et al. [16] introduced a data-driven power allocation method for heterogeneous cellular networks to enhance sum-rate performance and their results demonstrated that data-driven techniques outperform traditional model-based methods. Ding et al. [17] presented a deep RL-based approach for power control and their approach shows improved energy efficiency and faster convergence as compared to conventional methods. Giannopoulos et al. [18] presented a demand-aware power allocation strategy designed to maximize user throughput while addressing individual user requirements. Wang et al. [19] developed a power allocation scheme for underwater acoustic networks aimed at maximizing node capacity and they also analyze the model's convergence behavior. Jiang et al. [20] proposed a Q-learning-based spectrum and power control method for D2D underlay networks aimed at maximizing network capacity and results shows an improvement over random power allocation method. A deep RL-based power control approach is applied to dense networks with the goal of maximizing sum-rate and energy efficiency within a multi-agent framework and its performance is compared against traditional resource allocation algorithms [21, 22, 23]. Sun et al. [24] introduced a joint approach combining deep deterministic policy gradient and unsupervised learning for channel and power allocation in centralized architecture which aims to enhance the energy efficiency of modern communication systems. Based on the existing literature we can say that learning-based power allocation algorithms outperform optimization-based methods in dynamic heterogeneous wireless networks. However, all learning approaches focus on maximizing network resources under specific constraints and often overlook reliability in dense heterogeneous wireless environments. Hence, our work focuses on fulfilling the gaps which arises in dense networks and hence, it is as follows:

- Reliability becomes a key challenge in interference limited dense heterogeneous networks. To tackle this, a reward function is designed that ensures QoS for both macrocell and smallcell users alongwith also maintaining the required SINR level which is necessary to meet the performance demands of IoT devices.
- To deliver optimal performance in dynamic environments, we present a scalable and flexible distributed Q-learning based power allocation for

dense heterogeneous IoT networks. As compared to traditional Q-learning, our method reduces computational effort, minimizes memory requirements, and achieves faster convergence by narrowing the exploration to a smaller state space.

- We carry out convergence analysis and complexity analysis of our proposed Q-learning based power allocation method for heterogeneous IoT networks. We conduct comprehensive simulations to analyze the performance and examine the behavior under both static and dynamic learning rates. We compare our work with other learning strategies and findings highlight the algorithm's flexibility, reliability, and suitability for diverse IoT applications.

The organization of this paper is as follows: Section 2 introduces the system model and details the power allocation problem. Section 3 explains the key elements of the Q-learning algorithm. Section 4 presents the simulation results and analysis and finally section 5 concludes the paper with a summary and suggestions for future work.

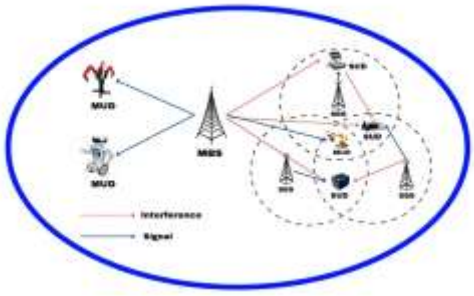


***Figure 1.*** *Heterogeneous IoT wireless network architecture*

***Table 1.*** *Abbreviations*

| Symbols | Descriptions |
|---|---|
| 5G | Fifth Generation |
| B5G | Beyond Fifth Generation |
| AWGN | Additive White Gaussian Noise |
| IoT | Internet of Things |
| RL | Reinforcement Learning |
| IIoT | Industrial Internet of Things |
| D2D | Device to Device |
| QoS | Quality of Service |
| MBS | Macro Base Station |
| SBS | Small Base Station |
| MUD | Macrocell User Device |
| SUD | Smallcell User Device |
| IL | Independent Learning |
| CL | Cooperative Learning |
| DL | Docitive Learning |
| $S$ | Set of all States |
| $A$ | Set of all Actions |
| $S_t$ | State at time $t$ |
| $A_t$ | Action at time $t$ |
| $r$ | Reward Function |
| $\lambda$ | Discount Factor |
| $\alpha$ | Learning rate |

## 2. System Model and Problem Formulation

We consider a heterogeneous IoT wireless network architecture consisting of a single macro base station (MBS) and N small base stations (SBSs). The MBS supports M macrocell user devices (MUDs), while each SBS connects to different types of IoT-enabled smallcell user devices (SUDs), as illustrated in Fig. 1. To minimize interference between users served by the same base station, resource multiplexing is utilized. Thus, for simplicity and without affecting generality, it is assumed that each of the N SUDs is connected to a unique SBS.

We focus on a downlink communication scenario where MUDs experience interference from multiple SBSs, while SUDs may face interference from both the MBS and other SBSs. Let $p_t^M$ be the power transmitted by MBS at time $t$ and $p_t^n, for\ n = 1,2 \ldots . N$ be the power transmitted by the $n^{th}$ SBS at time $t$ respectively. Let $g_{mbs,mud}$ represents the channel gain between MBS and the MUD and $g_{sbs_n,mud}$ represents the gain between $n^{th}$ SBS and MUD. Similarly, let $g_{sbs_n,sud_n}$ represents the gain between $n^{th}$ SBS and its serving SUD and $g_{mbs,sud_n}$ represents the gain between MBS and $n^{th}$ SUD. The noise variance is denoted by $\sigma^2$. Thus,

the capacity $C_t^m$ of the $m^{th}$ MUD at time $t$ is given by following expression

$$C_t^m = \log_2\left(1 + \frac{p_t^M g_{mbs,mud}}{\sum_{n=1}^{N} p_t^n g_{sbs_n,mud} + \sigma^2}\right) \quad (1)$$

Similarly, the capacity $C_t^n$ of $n^{th}$ SUD at time $t$ is given by following expression:

$$C_t^n = \log_2\left(+ \frac{p_t^n g_{sbs_n,sud_n}}{p_t^n g_{mbs,sud_n} + \sum_{i=1,i\neq n}^{N} p_t^i g_{sbs_i,sud_i} + \sigma^2}\right) \quad (2)$$

In this work, we assume that the MBS transmits with a constant power level. Let $p_{min}$ and $p_{max}$ represent the minimum and maximum allowable transmit power levels for all SBSs, respectively. Furthermore, the minimum QoS requirements for SUDs and MUDs are denoted by $\Gamma_{SUD}$ and $\Gamma_{MUD}$ along with the minimum received SINR thresholds needed to maintain a reliable connection are given by $SINR_{min}^{SUD}$ for SUDs and $SINR_{min}^{MUD}$ for MUDs. The main goal is to maximize the total capacity of all SBSs within a densely deployed network, while ensuring that the QoS demands of all IoT users are met. Based on this, we define the following optimization problem:

$$\max_{\{p_t^1, p_t^2 \ldots\ldots p_t^N\}} \sum_{n=1}^{N} C_t^n, \quad (3)$$

$$s.t.\ p_{min} \leq p_t^n \leq p_{max}, \forall n = 1,2 \ldots N, (3a)$$

$$C_t^m \geq \Gamma_{MUD}, \forall m = 1,2 \ldots\ldots M, \quad (3b)$$

$$C_t^n \geq \Gamma_{SUD}, \forall n = 1,2 \ldots\ldots N, \quad (3c)$$

$$SINR_t^m \geq SINR_{min}^{MUD}, \forall m = 1,2 \ldots\ldots M, \quad (3d)$$

$$SINR_t^n \geq SINR_{min}^{SUD}, \forall n = 1,2 \ldots\ldots N \quad (3e)$$

The constraint in (3a) ensures that the transmit power of each SBS remains within the defined minimum and maximum limits. Constraints (3b) and (3c) enforce the minimum capacity requirements for MUDs and SUDs, respectively, while (3d) and (3e) ensure that the received SINR for both MUDs and SUDs stays above the required thresholds. However, the optimization problem described in Eq. (3) is non-convex due to the SINR expression, which involves coupled variables in both numerator and denominator and it forms a ratio of linear and affine components. Since, the problem is non-convex which makes it difficult to solve using conventional optimization techniques. To address this, we introduce reinforcement learning approach capable of handling the dynamic conditions of heterogeneous networks efficiently.

## 3. Design of Proposed Algorithm

Section 3.1 outlines the fundamental components of the Q-learning algorithm and introduces a reward function designed to enhance power allocation efficiently. In Section 3.2, different learning rate configurations are analyzed to evaluate their effect on algorithm performance. Section 3.3 explores multiple modes of collaboration among agents and how they are incorporated into the proposed learning-based framework.

### 3.1. Elements of Q-Learning

**States:** Let $\boldsymbol{S} = [S_1, S_2, \ldots\ldots\ldots S_N]$ denote the collection of states for all $N$ SBSs where $S_n$, $for\ n = 1,2 \ldots\ldots N$ corresponds to the state of the $n^{th}$ SBS. The state of an SBS is defined by two key parameters: $S = [D_{mbs}, D_{mud}]$, where $D_{mbs}$ refers to the distance between SBS and MBS and $D_{mud}$ refers to the distance between SBS and MUD. In the following, let $d_{mbs,sbs_n}$ represent the distance between MBS and $n^{th}$ SBS and this distance is partitioned into $X$ equal segments and the value of $D_{mbs}$ is defined based on these segments

$$D_{mbs} \in \{1,2, \ldots\ldots X\} \quad (4)$$

The distance indicator $D_{mbs} = x$ if distance between $n^{th}$ SBS and MBS satisfies both the condition i.e.

$d_{mbs,sbs_n} \geq d_{mbs,sbs_{(x-1)}}$ and $d_{mbs,sbs_n} < d_{mbs,sbs_x}$.

Similarly, let $d_{mud_m,sbs_n}$ represents distance between $n^{th}$ SBS and $m^{th}$ MUD and we partitioned the distance into $Y$ equal segments. Hence, distance indicator $D_{mud}$ of each SBS consists of $M$ index values, i.e. one for each MUD and it is expressed as $D_{mud}^m = \{1,2, \ldots\ldots M\}$ reflecting the segmented distance levels to all MUDs:

$$D_{mud}^m \in \{1,2, \ldots\ldots Y\} \quad (5)$$

where $D_{mud}^m = y$ if distance between each SBS to MUD '$m$' satisfies both the condition

$d_{mud_m,sbs_n} \geq d_{mud,sbs_{(y-1)}}$ and $d_{mud_m,sbs_n} < d_{mud,sbs_y}$.

Thus, in the overall multi-agent framework, the complete set of states is denoted as $\boldsymbol{S} = [S_1, S_2, \ldots\ldots\ldots S_N]$, where each individual state $S_n$ for the $n^{th}$ SBS is defined as $S_n = [D^n_{mbs}, D^n_{sud}]$, with $n = 1,2 \ldots\ldots\ldots N$. Since the location of an SBS is fixed, its state remains constant over time. This property allows SBSs with the same state to share their Q-tables effectively. Instead of exchanging the full Q-table, they only need to share the specific row related to the common state. This approach minimizes complexity and improves the learning efficiency.

**Action:** Let $\boldsymbol{A} = [A_1, A_2, \ldots\ldots\ldots A_N]$, represent the collection of actions for all $N$ SBSs where $A_n$, corresponds to the action choices available to the $n^{th}$ SBS. The transmission power for each SBS is discretized by dividing the range from $p_{min}$ and $p_{max}$ into $Z$ uniform levels. As a result, the action set for each SBS is composed of $Z$ possible transmit power values within this defined range and is represented as:

$$A_n \in \{a_1, a_2, \ldots\ldots\ldots a_Z\} \quad (6)$$

Here, $a_n$ represents the chosen transmit power level for the $n^{th}$ SBS from its corresponding action set.

**Reward Function:** The reward function is pivotal in reinforcement learning algorithms as it assesses the desirability of the agent's actions within its environment. In the context of optimization problems, the reward function is often aligned with the objective function that needs to be maximized or minimized. The authors in [11, 12, 13] have defined reward functions which we termed as RwF1, RwF2 and RwF3 respectively and all reward functions is in accordance with the optimization problem of maximizing the total capacity of all SUDs but all the reward functions fails to maintain a minimum acceptable SINR for reliable connection which is important in dense heterogeneous IoT network where interference is a limiting factor. Motivated by the shortcomings of the reward functions as defined in [11, 12, 13], we have proposed our reward function which is in positive relation with the optimization problem and also consider the QoS of both SUDs and MUDs. Also our reward function maintains a minimum received SINR level for reliable connection. We can conclude that reliability is utmost important to meet the needs of various IoT based applications in various sectors, and hence we have defined our reward function and termed as PRwF and can be expressed as follows:

$$r^n_m(t) = \begin{cases} d_n C^n_t - \dfrac{1}{d_n}(C^m_t - \Gamma_{MUD})^2 - (C^n_t - \Gamma_{SUD})^2, if\ both\ Eq.(3d)\ and\ Eq.(3e) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad is\ satisfied \\ v\,, \qquad otherwise \end{cases} \quad (7)$$

Where $v$ is a large negative constant value. The large negative reward discourages the selection of infeasible solutions that violate the constraints. This effectively forces the optimization algorithm to explore only the feasible region of the solution space, guiding the search toward valid solutions. The parameter $d_n$ is defined as ratio of distance of $n^{th}$ SBS to the MUD normalized by the constant $d_{th}$. The parameter $d_{th}$ helps to determine whether $n^{th}$ SBS is near or far from MUD. For instance, if distance is less than $d_{th}$, it indicates that the interference from the $n^{th}$ SBS has a greater impact on the MUD compared to any other SBS located farther away where the distance is more than $d_{th}$. The total reward of SBS can be calculated as:

$$r^n(t) = \sum_{m=1}^{M} r^n_m(t) \quad (8)$$

### 3.2 Q-Value, Update Rule, and Learning Rate Parameters

Building on the previously defined states, actions, and reward functions, the Q-value for the proposed power allocation algorithm [27–30] can be formulated. At time t, given the state $S_t \in \boldsymbol{S}$ and action $A_t \in \boldsymbol{A}$, the $Q(S_t, A_t)$, represents the expected value for all N SBSs

$$Q_{t+1}(S_t, A_t) = Q(S_t, A_t) + \alpha \left[ r_{t+1}\, \lambda \max_{A_{t+1}} Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \right] \quad (9)$$

The values of α and λ are provided in Table 1. Solving Equation (9) requires Bellman optimality principle which involves exploring a large search space and hence increases the complexity. Unlike

standard Q-learning that uses a single agent, distributed Q-learning employs multiple agents collaborating to achieve a common goal. This framework allows each agent to benefit from shared information thus improving the overall outcome. Moreover, interaction and cooperation among agents help overcome individual weaknesses, leading to quicker and more efficient decision-making.

Next, we investigate two types of learning rates, denoted as $\alpha$, in Eq. (9). Specifically:

1. Fixed learning rate: When $\alpha$ is constant, learning process is more suited for dynamic environments. Here, we consider $\alpha = 1/2$.
2. Dynamic learning rate: In this case we consider $\alpha = 1/(N_{S,A} + 1)$, where $N_{S,A}$ represents the frequency of occurrence of the state action pair $(S, A)$. By reducing $\alpha$ at an appropriate rate we can achieve convergence to correct $Q$ function.

### 3.3 Collaboration Strategies in Multi-Agent Reinforcement Learning

The interaction among multiple SBSs in the decision-making process constitutes a multi-agent reinforcement learning scenario, and we examine different learning approaches within this context:-

• Independent learning (IL): In independent learning, each agent focuses solely on maximizing its own reward without accounting for the performance or rewards of other agents. Accordingly, the decision-making process for the $n^{th}$ SBS is based entirely on its individual learning and experience:

$$A_n = \arg\max_{A \in \boldsymbol{A}} Q_n(S_n, A) \qquad (10)$$

Cooperative learning (CL): In contrast to independent learning, cooperative learning considers the actions of other SBSs, which helps to improve both learning efficiency and system performance. Within this approach, each SBS makes decisions by factoring in the behavior of others, leading to faster convergence and better outcomes. The cooperative action selection is defined accordingly:

$$A_n = \arg\max_{A \in \boldsymbol{A}} \sum_{1 \le i \le N} Q_i(S_i, A) \qquad (11)$$

Docitive learning (DL): Docitive Learning focuses on maximizing the collective Q-value of all SBSs by allowing agents to share knowledge, thereby accelerating convergence and improving performance. This collaborative objective can be mathematically formulated as:

$$A_n = \arg\max_{A \in \boldsymbol{A}} \left[ \max_{i \in N} Q_i(S_i, A) \right] \qquad (12)$$

Algorithm 1: Distributed $Q$-learning Based Power Allocation.

1. **Input:** $M$, $N$ and position of MBS, SBSs, MUDs, SUDs; gains $g_{mbs,mud}$ , $g_{sbs\,,mud}$, $g_{mbs,sud}$, $g_{sbs\,,sud}$ and $\sigma^2$ ; transmission power $p_t^M$, $p_{min}$ and $p_{max}$; QoS parameters $\Gamma_{MUD}$, $\Gamma_{SUD}$, $SINR_{min}^{MUD}$ and $SINR_{min}^{SUD}$; other variables $X$, $Y$ , $Z$, $\lambda$ $\alpha$ and $\varepsilon$ etc.
2. **Initialization:** Initialize $iter = 0$, $Q = 0$, $N_{S,A} = 0$ and temp $Q$-table $Q_{\sum\,,t} = 0$.
3. **for** $iter \in [1, iter_{max}]$
4. Calculate $Q_{\sum\,,t} = \sum_{n \in N} Q_n$
5. **for** all SBS $n \in [1, N]$
6. Generate random $temp$ value
7. **if** $iter < \theta * iter_{max}$ and $temp < \varepsilon$ then
8. Randomly selects an action.
9. **else**
10. Selects an action for given SBS according to learning mode.
11. **end if**
12. **end for**
13. Calculate SUDs and MUDs capacity.
14. **for** all SBS $n \in [1, N]$
15. Calculate total reward.
16. Update the $Q$ table.
17. Observe new state.
18. **end for**
19. **if** $\left| \sum_{n \in N} Q_j - Q_{\sum\,,t} \right|_2 < \beta$
20. Capture results and exit.
21. **end if**
22. **end for**
23. **End**.

## 4. Numerical Results and Discussions

### 4.1. Simulation Parameters

We consider a heterogeneous IoT wireless network as illustrated in Fig. 1, comprising a single macro base station (MBS), one MUD (M = 1) and number of SBSs ($N = 16$) each serving one SUD. To simulate a densely deployed scenario with significant interference, the MUD is positioned centrally among the SBSs and SUDs, following the layout described in [13]. The MBS transmitted power $p_t^M$ and noise power $\sigma^2$ is set to be $50dBm$ and $-120dB$ respectively. The transmit power of all SBSs varies between a minimum of $-20dBm$ and a maximum of $25dBm$, with a step of $1.5\,dBm$ resulting in 31 possible power levels ($N_p = 31$). The simulation is structured around three concentric layers centered on the MBS and MUD. Both MUD and SUD require a minimum data rate of $2\ b/s/Hz$, represented by thresholds $\Gamma_{MUD}$ and $\Gamma_{SUD}$ The minimum SINR requirements i.e. $SINR_{min}^{MUD}$ and $SINR_{min}^{SUD}$ are set to $5\,dB$. Additional simulation parameters include i.e. $\alpha = 0.5$, $\lambda = 0.9$, $\varepsilon = 0.1$ and $iter_{max} = 50000$.

In this study, we assume that both the macro and small base stations operate on a common carrier frequency of $f = 0.9GHz$. The signal attenuation between MBS and the MUD, as well as the link between the SBS and its serving SUD, is given as $PL = 62.3dB + 40log10(d/5)$. Likewise, the signal attenuation between the MBS and a SUD, as well as between the $i^{th}$ SBS and $j^{th}$ SUD is determined as $PL = 62.3 + 32log10(d/5) + PL_i$. Here $PL_i$ is an additional loss component calculated using an empirical model as given in [25] with the number of walls $N_W = 2$ considered in the scenario.

### 4.2 Fairness

In a multi-agent framework, maintaining fairness across several devices plays a vital role in improving network efficiency. To assess this aspect in our simulations, we adopt Jain's fairness index [26] as a standard metric. This index enables us to quantify how evenly the resources are distributed among SUDs and facilitates resource allocation strategies in terms of fairness. The mathematical expression of the fairness index is given as follows:

$$f(c_1, c_2, \dots\dots c_N) = \frac{(\sum_{n=1}^{N} c_n)^2}{N \sum_{n=1}^{N} c_n^2} \qquad (13)$$

Where $c_n$ denotes the capacity allocated to the $n^{th}$ SBS among total SBS in the network.

### 4.3. Performance comparison of proposed reward function with other reward functions.

We evaluate and assess the effectiveness of the proposed reward function (PRwF) by comparing its performance against existing reward functions defined in [11, 12, 13] and we denote them as RwF1, RwF2 and RwF3 respectively. To assess the performance we consider fixed learning rate ($\alpha$ = 0.5) and cooperative learning structure. The results presented in Fig. 2 are obtained by averaging the outcomes over 100 independent simulation runs to ensure reliability and consistency in the comparison. Fig. 2(a) illustrates that with increasing numbers of SBSs, the average capacity for the MUD drops below the required threshold when using RwF1 and RwF2, indicating their limitations in preserving QoS under dense conditions. Conversely, RwF3 and the PRwF maintain MUD capacity above the threshold, showcasing greater resilience to interference and better resource management. This confirms the capability of PRwF to sustain QoS and enhance network performance in high-density scenarios Fig. 2(b) presents the comparison of various reward functions in meeting the minimum capacity needs of SUDs as there is increase in SBS. RwF1 and RwF2 fail to maintain the required capacity once the SBS count exceeds 6, indicating their inefficiency in dense deployments. RwF3 performs better and sustains capacity up to 12 SBSs, while the PRwF extends this support up to 16 numbers of SBS. These findings underline the superior capability of PRwF in effectively controlling interference and allocating resources, enabling reliable performance even under high network density. Fig. 2(c) depicts the aggregate capacity of SUDs, which is the main goal of the optimization. As the number of SBSs grows, the PRwF consistently delivers greater aggregate capacity than the other reward functions. When the number of SBS reaches to maximum, PRwF achieves about 5% higher aggregate capacity as compared to RwF3, showing a clear enhancement. The improvement is even more pronounced against RwF1 and RwF2, emphasizing PRwF's stronger capability in optimizing resource distribution. Fig. 2(d) presents a comparison of the fairness index among SUDs, showcasing how well different reward functions allocate resources fairly. The results reveal that the PRwF attains a higher fairness index than the others, promoting a more equitable capacity distribution among SUDs. When the number of SBS reaches to maximum, RwF3

achieves a fairness index of 0.9183, whereas PRwF outperforms with a value of 0.9499, reflecting improved fairness. Conversely, RwF1 and RwF2 struggle to maintain fairness as network density increases, resulting in greater capacity imbalance. These findings highlight that PRwF not only enhances overall capacity but also supports fair resource sharing, making it ideal for dense network environments. Fig. 2(e) shows the likelihood of SUDs experiencing an SINR below the required threshold, indicating potential outages. The data clearly indicate that the PRwF significantly reduces outage probability compared to other reward functions. While RwF1 and RwF2 have an average outage rate of 50%, meaning half of the SUDs fail to meet the minimum SINR as the number of SBSs grows, PRwF outperforms RwF3 by lowering outage probability by 60% when there are 16 SBSs. This highlights PRwF's effectiveness in sustaining reliable connections and enhancing QoS, especially in dense network scenarios. Fig. 2(f) and Fig. 2(g) present the number of iterations and the computation time required for different reward functions during performance evaluation. While the proposed PRwF outperforms RwF1, RwF2, and RwF3 in terms of capacity, fairness, and outage probability, it requires slightly more iterations and longer calculation time to converge. This increase is due to large negative constant included in reward function which lead to slower convergence because the optimization algorithm may oscillate between feasible and infeasible regions and which focuses on achieving better resource allocation and QoS. Despite the marginal increase in computational time, the superior performance gains in key metrics such as capacity, fairness, and reliability make PRwF a favorable choice for dense network scenarios where optimized performance is critical. Additionally, Table 2 provides a concise comparison of the performance outcomes for various reward functions when there are maximum number of SBSs considered for simulation.

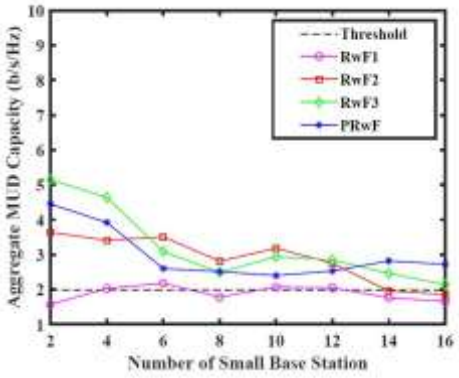


**Figure 2.** *Average MUD Capacity*

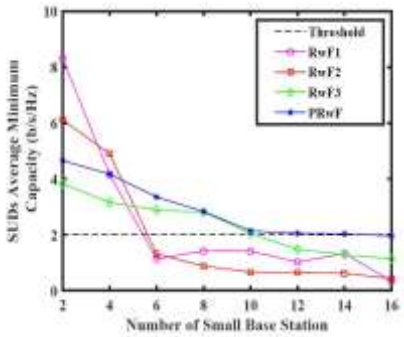


***Figure 3.*** *SUDs Average Minimum Capacity*

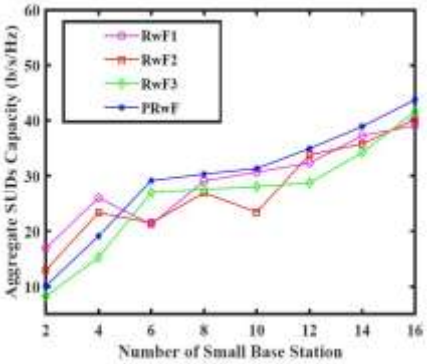


**Figure 4.** *Aggregate SUDs Capacity*

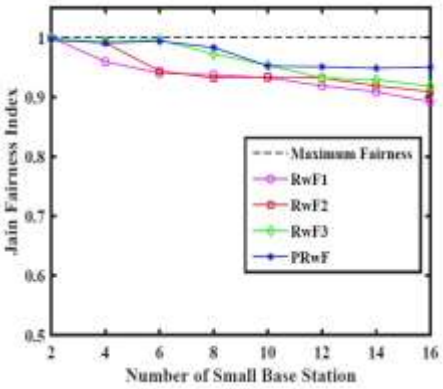


**Figure 5** *Jain Fairness Index*

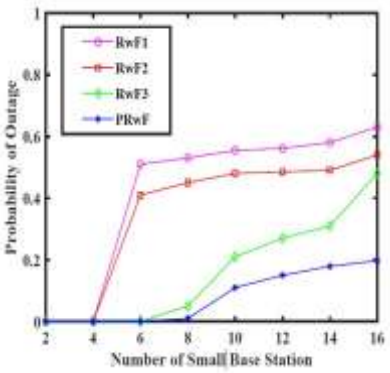


**Figure 6.** *Probability of Outage*

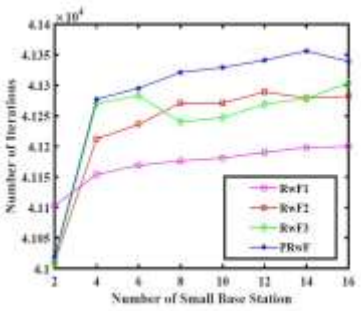


***Figure 7.*** *Number of Iterations*

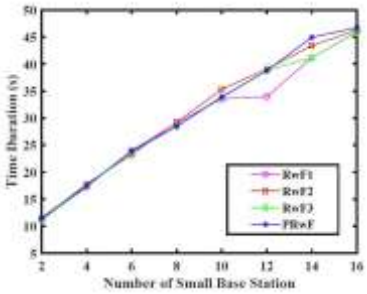


**Figure 8.** *Time Duration*

***Table 2.*** *Comparison of Proposed Reward Function with other Reward Functions*

| Reward Function | QoS guarantee for both MUDs and SUDs | Fairness | Outage Probability | Reliability | Convergence |
|---|---|---|---|---|---|
| RwF1 | No | 0.8918 | High | No | Fast |
| RwF2 | Yes | 0.9086 | High | No | Slow |
| RwF3 | No | 0.9183 | High | No | Slow |
| PRwF | Yes | 0.9499 | Low | Yes | Slow |

### 4.4. Performance comparison of different learning modes

We investigate the performance of our PRwF with various learning modes along with fixed and dynamic learning rate. The performance evaluation of PRwF highlights that CL consistently outperforms DL and IL across multiple metrics, primarily due to its coordinated decision-making and resource allocation. Various performance measures for different learning modes are shown in Fig. 3. From Fig. 3(a) we can conclude that DL shows some advantages in specific scenarios, such as higher capacity because, its ability to maintain consistent QoS across the network. In Fig. 3(b) and Fig. 3(c) CL achieves superior capacity management by leveraging global network information, which allows it to optimize resource distribution more effectively, ensuring better QoS compliance even as the number of SUDs and SBSs increases. Additionally, from Fig. 3(d) we can conclude that CL demonstrates improved fairness in resource allocation by minimizing disparities in service quality, which is a critical factor in maintaining network stability and user satisfaction for both fixed and dynamic learning rate. The lower outage probability observed in Fig. 3(e) further emphasizes CL's ability to ensure reliable connectivity, reducing the likelihood of SUDs experiencing insufficient SINR. However for both fixed and dynamic learning rate CL will have 60% to 70% less outage probability than DL and IL when number of SBS is 16. In contrast, IL struggles due to its lack of coordination, leading to higher iteration counts as shown in Fig. 3(f) and longer convergence times as shown in Fig. 3(g), which can hinder real-time network performance. Furthermore, we can conclude that the coordinated approach of CL offers a more robust solution for managing network resources efficiently, making it the most reliable learning mode for achieving optimal performance.

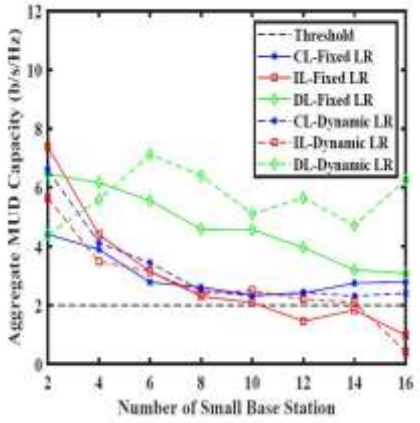


**Figure 9.** *Average MUD Capacity*

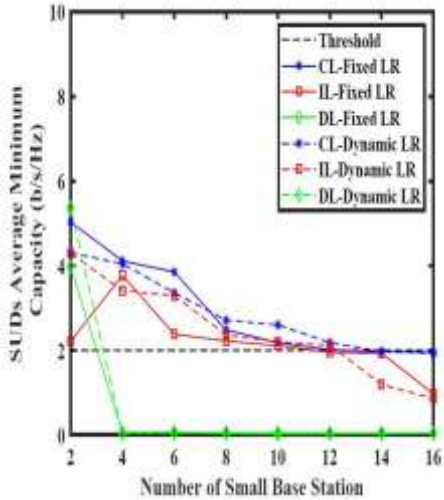


**Figure 10.** *SUDs Average Minimum Capacity*

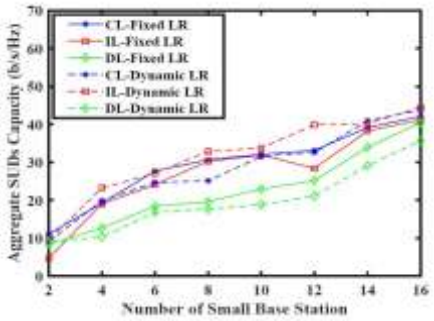


***Figure 11.*** *Aggregate SUDs Capacity*

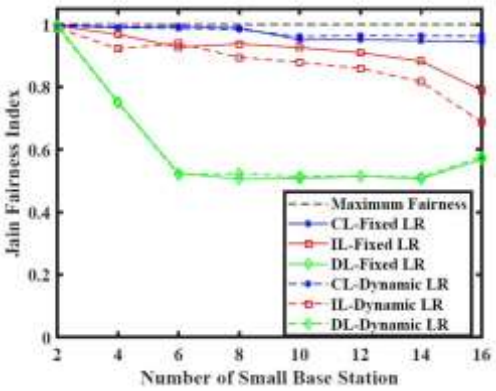


***Figure 12.*** *Jain Fairness Index*

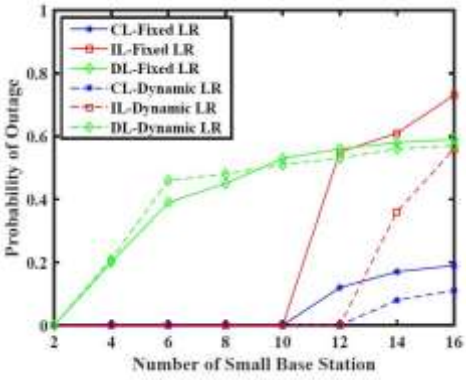


***Figure 13.*** *Probability of Outage*

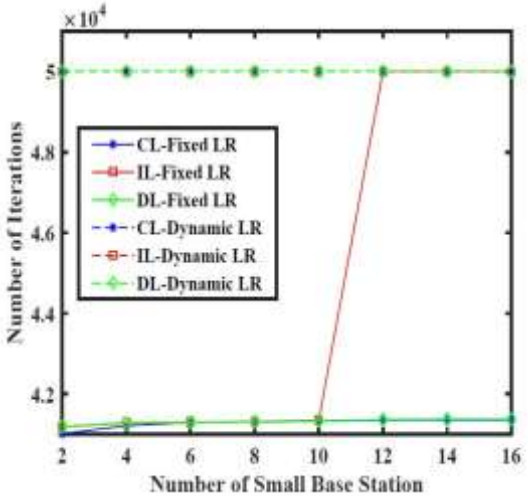


**Figure 14.** *Number of Iterations*

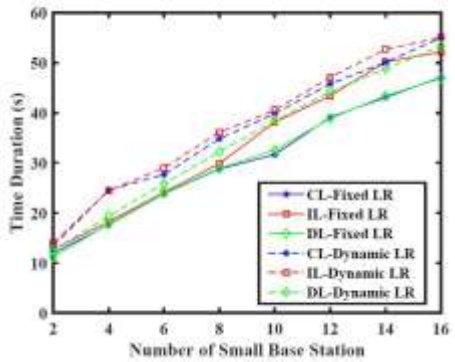


**Figure 15.** *Time Duration*

### 4.5. Convergence Analysis

In our simulation we have assumed the number of iterations for our algorithm to be 50000, although it consistently converges well before reaching this limit. Fig. 2(f) illustrates the convergence behavior of the proposed work in relation with the number of SBSs. The order of iterations needed by the proposed algorithm is around $4 \times 10^4$ which is approximately in between $2^{15}\ to\ 2^{16}$ and it is significantly a smaller fraction of the total iterations required by an exhaustive search, which would require $32^{16} = 2^{80}$ iterations. This demonstrates that the proposed algorithm's efficiency, offering a considerable reduction in computational complexity while still achieving convergence in large-scale network scenarios.

### 4.6. Complexity Analysis

In our work, we combine our power allocation algorithm with Independent Learning (IL), Cooperative Learning (CL), and Docitive Learning (DL) to enhance the performance of heterogeneous IoT systems. The algorithm is summarized in Algorithm 1 along with simulation parameters. In the initial loop stage, conducting a comprehensive search, containing $Z^N$ elements in action set $\boldsymbol{A}$, leads to a complexity of $O(Z^N)$ for each SBS. Subsequently, in subsequent loop stage, updating $Q$-values also requires a complexity of $O(Z^N)$. Assuming maximum iteration to be $iter_{max}$, hence the overall complexity can be expressed as $O(iter_{max} \times N \times Z^N)$.

## 5. Conclusion and Future Work

This study focuses on optimizing power allocation in dense heterogeneous IoT networks with the goal of maximizing the total capacity of smallcell user devices (SUDs) while meeting their QoS needs. To this end, we introduced a reward function that

promotes fair resource distribution and ensures the minimum SINR for stable connections. Our approach uses a distributed Q-learning algorithm incorporating different learning strategies, which effectively improves network performance and reliability in interference-prone settings. The proposed method supports various industrial applications by efficiently handling power in complex environments. Future work will explore scalability in denser deployments and investigate cognitive radio integration for further enhancement of the model.

## Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## References

[1] Zhou, I., Li, X., Chen, Y., & Zhang, H. (2021). Internet of Things 2.0: Concepts, applications, and future directions. *IEEE Access*, *9*, 70961–71012. https://doi.org/10.1109/ACCESS.2021.3078549

[2] Palattella, M. R., Dohler, M., Grieco, L. A., Boggia, G., & Mendes, P. (2016). Internet of Things in the 5G era: Enablers, architecture, and business models. *IEEE Journal on Selected Areas in Communications*, *34*(3), 510–527. https://doi.org/10.1109/JSAC.2016.2525418

[3] Peng, M., Zhang, K., Jiang, J., Wang, J., & Wang, W. (2015). Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *IEEE Transactions on Vehicular Technology*, *64*(11), 5275–5287. https://doi.org/10.1109/TVT.2014.2379922

[4] Zhang, H., Huang, S., Jiang, C., Long, K., Leung, V. C. M., & Poor, H. V. (2017). Energy-efficient user association and power allocation in millimeter-wave-based ultra-dense networks with energy harvesting base stations. *IEEE Journal on Selected Areas in Communications*, *35*(9), 1936–1947. https://doi.org/10.1109/JSAC.2017.2720898

[5] Bin Farooq, U., Sajid Hashmi, U., Qadir, J., Imran, A., & Mian, A. N. (2018). User transmit power minimization through uplink resource allocation and user association in HetNets. *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, 1–6. https://doi.org/10.1109/GLOCOM.2018.8647409

[6] Song, L., Chai, K. K., Chen, Y., Schormans, J., Loo, J., & Vinel, A. (2017). QoS-aware energy-efficient cooperative scheme for cluster-based IoT systems. *IEEE Systems Journal*, *11*(3), 1447–1455. https://doi.org/10.1109/JSYST.2015.2465292

[7] Wang, J., Jiang, C., Wei, Z., Pan, C., Zhang, H., & Ren, Y. (2019). Joint UAV hovering altitude and power control for space-air-ground IoT networks. *IEEE Internet of Things Journal*, *6*(2), 1741–1753. https://doi.org/10.1109/JIOT.2018.2875493

[8] Bakht, K., Ahmed, R., Ijaz, S., Ahmed, A., Jabeen, F., & Khwaja, A. S. (2019). Power allocation and user assignment scheme for beyond 5G heterogeneous networks. *Wireless Communications and Mobile Computing*, 1–9. https://doi.org/10.1155/2019/2472783

[9] Ha, V. N., & Le, L. B. (2014). Fair resource allocation for OFDMA femtocell networks with macrocell protection. *IEEE Transactions on Vehicular Technology*, *63*(3), 1388–1401. https://doi.org/10.1109/TVT.2013.2284572

[10] Kai, C., Li, H., Xu, L., Li, Y., & Jiang, T. (2019). Joint subcarrier assignment with power allocation for sum rate maximization of D2D communications in wireless cellular networks. *IEEE Transactions on Vehicular Technology*, *68*(5), 4748–4759. https://doi.org/10.1109/TVT.2019.2903815

[11] Saad, H., Mohamed, A., & Elbatt, T. (2012). Distributed cooperative Q-learning for power allocation in cognitive femtocell networks. *Proceedings of the IEEE Vehicular Technology Conference (VTC-Fall)*, 1–5. https://doi.org/10.1109/VTCFall.2012.6399230

[12] Tefft, J. R., & Kirsch, N. J. (2013). A proximity-based Q-learning reward function for femtocell networks. *Proceedings of the IEEE Vehicular Technology Conference (VTC-Fall)*, 1–5. https://doi.org/10.1109/VTCFall.2013.6692057

[13] Amiri, R., Mehrpouyan, H., Fridman, L., Mallik, R. K., Nallanathan, A., & Matolak, D. (2018). A machine learning approach for power allocation in HetNets considering QoS. *Proceedings of the IEEE International Conference on Communications (ICC)*, 1–6. https://doi.org/10.1109/ICC.2018.8422864

[14] Zhang, G. A., Gu, J. Y., Bao, Z. H., Xu, C., & Zhang, S. B. (2014). Joint routing and channel assignment algorithms in cognitive wireless mesh networks. *Transactions on Emerging Telecommunications Technologies*, *25*(3), 294–307. https://doi.org/10.1002/ett

[15] Zhang, Y., Kang, C., Ma, T., Teng, Y., & Guo, D. (2018). Power allocation in multi-cell networks using deep reinforcement learning. *Proceedings of*

*the IEEE Vehicular Technology Conference (VTC-Fall)*, 1–6. https://doi.org/10.1109/VTCFall.2018.8690757

[16] Meng, F., Chen, P., Wu, L., & Cheng, J. (2020). Power allocation in multi-user cellular networks: Deep reinforcement learning approaches. *IEEE Transactions on Wireless Communications*, *19*(10), 6255–6267. https://doi.org/10.1109/TWC.2020.3001736

[17] Ding, H., Zhao, F., Tian, J., Li, D., & Zhang, H. (2020). A deep reinforcement learning for user association and power control in heterogeneous networks. *Ad Hoc Networks*, *102*, 102069. https://doi.org/10.1016/j.adhoc.2019.102069

[18] Giannopoulos, A., Oikonomou, K., Koutsopoulos, I., & Antoniou, E. (2021). Demand-driven power allocation in wireless networks with deep Q-learning. *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 248–251. https://doi.org/10.1109/WoWMoM51794.2021.00045

[22] Huangi, R., Si, J., Shi, J., & Li, Z. (2021). Deep-reinforcement-learning-based resource allocation in ultra-dense networks. *Proceedings of the IEEE Wireless Communications and Signal Processing (WCSP)*, 1–5. https://doi.org/10.1109/WCSP52459.2021.9613186

[23] Zhao, Y., Peng, T., Guo, Y., & Wang, W. (2021). Energy-efficient uplink power allocation in ultra-dense networks through multi-agent reinforcement learning. *Proceedings of the IEEE Vehicular Technology Conference (VTC-Fall)*, 1–7. https://doi.org/10.1109/VTC2021-Fall52928.2021.9625554

[24] Sun, M., Mei, E., Wang, S., & Jin, Y. (2023). Joint DDPG and unsupervised learning for channel allocation and power control in centralized wireless cellular networks. *IEEE Access*, *11*, 42191–42203. https://doi.org/10.1109/ACCESS.2023.3270316

[25] Valcarce, A., & Zhang, J. (2010). Empirical indoor-to-outdoor propagation model. *IEEE Access*, *9*, 682–685. https://doi.org/10.1109/ACCESS.2010.5643660

[26] Bin Sediq, A., Gohary, R. H., Schoenen, R., & Yanikomeroglu, H. (2013). Optimal and low-complexity power allocation for OFDMA-based relay systems. *IEEE Transactions on Wireless Communications*, *12*(1), 171–184. https://doi.org/10.1109/TWC.2012.112112.110796

[27] Khenak, F. (2010). V-learning. *Proceedings of the 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM)*, 228–232. https://doi.org/10.1109/CISIM.2010.5643660

[28] Smith, J. Q., & White, D. J. (1994). Markov decision processes. *Mathematics of Operations Research*, *157*(1), 1–25.

[29] Hierons, R. (1999). *Machine learning* (T. M. Mitchell). McGraw-Hill.

[19] Wang, H., Fan, Y., & Yang, L. (2021). A power allocation algorithm for underwater acoustic communication networks based on reinforcement learning. *Proceedings of the IEEE International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, 986–990. https://doi.org/10.1109/ICCASIT53235.2021.9633599

[20] Jiang, S., & Zheng, J. (2021). A Q-learning based dynamic power control algorithm for D2D communication underlaying cellular networks. *Proceedings of the IEEE Wireless Communications and Signal Processing (WCSP)*, 1–5. https://doi.org/10.1109/WCSP52459.2021.9613167

[21] Anzaldo, A., & Andrade, A. G. (2022). Deep reinforcement learning for power control in multi-tasks wireless cellular networks. *Proceedings of the IEEE Mediterranean Communication and Networking Conference (MeditCom)*, 61–65. https://doi.org/10.1109/MeditCom55741.2022.9928617

[30] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.