









Machine Learning-Based Strategy for Joint User Association and Resource Allocation in Next-Generation Networks

Matheus Alves  [Universidade Federal do Sul e Sudeste do Pará | mathsalves@unifesspa.edu.br]
Gustavo Broechl  [Universidade Federal do Sul e Sudeste do Pará | gustavo.broechl@unifesspa.edu.br]
Luna Loyolla  [Universidade Federal do Sul e Sudeste do Pará | luna.loyolla@unifesspa.edu.br]
Warley Junior  [Universidade Federal do Sul e Sudeste do Pará | wmvj@unifesspa.edu.br]
Marcela Alves  [Universidade Federal do Sul e Sudeste do Pará | marcela.alves@unifesspa.edu.br]
Hugo Kuribayashi   [Universidade Federal do Sul e Sudeste do Pará | hugo@unifesspa.edu.br]

 FL 17 QD 04 LT Especial S/N, Nova Maraba, Maraba-PA, CEP 68505-080, Universidade Federal do Sul e Sudeste do Pará.

Received: 03 September 2024 • Accepted: 09 March 2025 • Published: 02 May 2025

Abstract This study presents an approach based on Reinforcement Learning (RL) to optimize the orchestration of User Association and Resource Allocation (UARA) mechanisms in next-generation heterogeneous networks, focusing on maximizing user satisfaction. The proposed strategy aims to improve the efficiency of these networks by overcoming operational challenges through user-centered adaptive algorithms. RL algorithms are utilized to rebalance the network load and optimize the distribution of radio resources among User Equipments (UEs), ultimately leading to improved service conditions. The results suggest that the strategic application of RL algorithms can lead to significant improvements compared to traditional methods, such as Max-SINR and Cell Range Expansion (CRE), reaching over 90% user satisfaction, highlighting the relevance of this research for next-generation networks.

Keywords: User Association, Resource Allocation, Machine Learning, Reinforcement Learning.

1 Introduction

The exponential increase in mobile data consumption has driven a growing demand for high-speed data services. Ensuring reliable Quality of Service (QoS) becomes a significant challenge for mobile network operators. There is an inevitable need to efficiently handle this massive volume of data while maintaining high data transfer rates. Given this scenario, the search for innovations in resource management becomes imperative to meet user expectations and ensure a superior connectivity experience (Jayaraman *et al.*, 2023).

Resource management in mobile networks faces complexity due to increasing demand and consumer expectations. Operators are exploring new approaches to improve operational efficiency by focusing on management processes that go beyond conventional practices. These efforts aim to meet the massive demand for high-speed data and ensure consistent QoS delivery under all operating conditions. Optimization, in this context, becomes a priority in order to cope with the demands imposed by the exponential increase in data traffic. Innovative resource allocation and load balancing strategies become crucial to ensure network efficiency, especially during peak demand times (Zhang *et al.*, 2019).

The related literature (Alhashimi *et al.*, 2023; Paixão *et al.*, 2023; Zhao *et al.*, 2019) indicates that efficient orchestration schemes for network management and operation mechanisms become crucial while also representing challenges, given the complex nature of such processes. Furthermore, the heterogeneity and densification of these networks translate into the need for interoperable and (ultra) flexible infrastructure to orchestrate the various processes involved.

Furthermore, integrating communications on different scales and various radio access technologies represents a highly heterogeneous and dynamic challenge, making it a task of significant complexity in terms of management and orchestration within Next-Generation Networks (NGNs). Emerging use cases and applications have stringent requirements for reliability, latency, and throughput (Shen *et al.*, 2020). Such requirements pose new challenges to network design, resource management, and device orchestration in NGNs. Network densification via deploying Ultra-Dense Small Cell Networks (UDNs) represents a promising approach and can improve network capacity (Tanveer *et al.*, 2022). With UDNs, it is possible to expand the network coverage area closer to the end-user devices and thus improve the spectral efficiency, performance, and service capacity. However, it does not provide a solution to scalable management of these networks.

Although crucial, it is challenging to manage UDN resources due to the non-uniform characteristics of user devices in space and time (Alzubaidi *et al.*, 2022). Therefore, allocating and adaptively orchestrating resources becomes the primary challenge since resource allocation mechanisms affect network performance and user experience. On the other hand, the user association strategy based on the maximum received Signal-to-Interference-plus-Noise Ratio (SINR) tends to unbalance the load on the network service. Even under Small Base Station (SBS) coverage, User Equipment (UE) still perceives the highest downlink signal from Macro Base Station (MBS) layer. Thus, MBSs tend to remain overloaded, while SBS-based layers remain idle, with a low service level (Adedoyin and Falowo, 2020). Therefore,

UEs should be reassociated with less overburdened SBSs to exploit a potentially greater availability of radio resources.

The concept of “load” in heterogeneous mobile networks has several definitions, often correlating with the number of UEs associated to a Base Station (BS). When the maximum capacity of UEs that a BS can support fluctuates, while the overall number of utilized radio Resource Block (RB) remains constant, the traffic load experienced by the UEs is directly proportional to the number of devices attached to the BS Bikram Kumar *et al.* (2019). Consequently, resource allocation and load balance become critical challenges in NGNs. The resource allocation problem is further complicated by the need to meet the varying QoS demands of UEs, while simultaneously striving to maximize the efficiency of available resources. This results in a complex task that must consider factors such as UE density, traffic dynamics, and coverage conditions within the network.

Furthermore, User Association and Resource Allocation (UARA) problems are essential and should be handled and optimized jointly. Joint UARA mechanisms could significantly enhance the network service capacity and help meet QoS requirements (Kim *et al.*, 2024). However, network conditions may change rapidly due to user mobility, time-varying channel conditions, and other spatio-temporal changes. Nevertheless, managing NGNs requires scalable and adaptable to adapt to dynamic network conditions. Hence, adaptive and flexible network orchestration resources become crucial, considering the accelerated growth in the demand for data traffic projected for the coming years.

Several approaches have been proposed in the literature to solve the challenges related to the optimization of UARA mechanisms (Jain *et al.*, 2021; Wang *et al.*, 2019; Xu *et al.*, 2021; Zhao *et al.*, 2019). Among the adopted techniques, it is possible to highlight the use of combinatorial optimization techniques, game theory approaches, stochastic geometry methods, and other heuristic-based methods. However, considering the nonconvex and combinatorial characteristics of this context, it is challenging to obtain a globally optimal joint optimization strategy. These approaches require nearly complete information, which may not be available, making the computation of the optimal strategy intractable. Moreover, considering a UDN-based network, the high density of BSs may represent a limiting barrier towards unmanageable levels of computational complexity.

In this challenging scenario, Machine Learning (ML)-based methods emerge as a promising strategy to facilitate resource allocation, load balancing, and user association processes. These techniques offer the possibility of learning complex and dynamic patterns, allowing for an adaptive response to changing network conditions. However, despite the growing interest in ML-based techniques in recent years, it is crucial to recognize that these approaches have limitations in specific real-world scenarios. For example, in densely populated urban environments characterized by high variability in signal conditions and interference, ordinary ML-based algorithms may struggle to generalize from training data, resulting in suboptimal resource allocation decisions that compromise service quality, particularly when faced with sparse or inaccurate data typical of real-world scenarios.

On the other hand, Reinforcement Learning (RL) can obtain the optimal policy to solve the intelligent decision problem by interacting with the environment. This approach has a low prerequisite for prior knowledge and is also a kind of online learning method, which has been extensively studied in ML (Naderializadeh *et al.*, 2021). In a RL environment, RL agents consider the maximum long-term rewards, rather than simply getting the current optimal rewards. This is important for time-varying dynamic systems (Zhao *et al.*, 2019, 2018).

This paper proposes an RL-based approach for the joint adjustment of the user association and resource allocation processes. With their continuous learning and adaptability characteristics, RL techniques offer a perspective to address the specific challenges of this complex network infrastructure. This mechanism can be designed using automated techniques to develop strategies that lead to efficient solutions in real time instead of conventional management proposals based on manual operation approaches or the solution of classical optimization problems.

Furthermore, RL facilitates the use of intelligent agents that interactively learn to adjust network operations with a specific objective, such as maximizing user experience. Multiple agents can collaborate or compete in this task. In addition, these agents can also retrieve knowledge by transferring learning from historical instances when the network conditions reached previously observed states.

Given this, this work establishes a RL-based strategy for the joint orchestration of UARA mechanisms, capable of assisting the management process of a mobile network. The proposed scheme must also seek individualized fulfillment of UEs traffic requirements, following the expected NGN trends and requirements. The main features of the proposed scheme include the following:

- Introduction of a RL strategy that maximizes the QoS levels of user equipment (UEs), demonstrating promising improvements in the satisfaction of these UEs;
- Analytical and technical schemes aimed at system modeling and partial automation of the functioning of UARA mechanisms, in accordance with the context of the presented problem.

The remainder of this paper is organized as follows. Section 2 presents the related work. Our proposed strategy to orchestrate the UARA mechanisms are detailed in Section 3. Section 4 presents the problem formulated based on maximization of UE satisfaction and presents the theoretical foundations involved in the study. Finally, Section 5 describes the experiments applied and the results obtained, while Section 6 concludes the study by making some final considerations.

2 Related Work

Several systematic work reviews have explored UARA mechanisms in NGNs in recent years (Xu *et al.*, 2021; Alhashimi *et al.*, 2023; Attiah *et al.*, 2020; Adedoyin and Falowo, 2020; Yazici *et al.*, 2023). There are also surveys to analyze the potential application of ML-based approaches to highlight the main methods and open challenges (Wang

et al., 2020; Wang *et al.*, 2021). In this context, we highlight some of the main works in UARA related to this work to synthesize their respective contexts, applied techniques, and potential performance gaps.

In (Zhao *et al.*, 2019), the authors propose a distributed RL method to solve the UARA problem in Heterogeneous Networks (HetNets). Given the non-convex and combinatorial nature of the problem, the method uses a multi-agent RL approach to maximize the long-term downlink utility while ensuring the QoS requirements of users. The results demonstrate that the method tends to outperform other optimization approaches in terms of system capacity and computational efficiency. However, the computational complexity of the study and the reliance on distributed strategies limit its applicability, and its validity remains unproven in UDN scenarios.

Labana and Hamouda (2020) address the problem of maximizing the efficiency of wireless communication networks through a Coordinated Multi-Point (CoMP) transmission approach. A convex optimization approach solves the problem, divided into three sub-problems: user association, resource allocation, and power allocation. The simulation results show significant improvements in network efficiency when the proposed CoMP approach is used, especially in resource-strapped scenarios and high-capacity fronthaul links. However, limitations of the work include the reliance on perfect channel state information and the assumption of ideal network conditions, which may not hold in real-world scenarios, affecting the practicality of the proposed solutions.

The study conducted by Kim *et al.* (2023) addresses the UARA challenges in wireless networks with high user mobility. Due to the complexity of these problems, especially in applications such as the Internet of Things, traditional solutions face difficulties. The study proposes a deep RL approach to address these issues. The proposed solution involves formulating the optimization problem where the main objective is to maximize the user data rates while minimizing the number of cell switches (handovers). The simulation results show that the proposed approach offers a faster convergence rate and superior performance in reducing cell swaps compared to conventional methods (a reduction of 58%). However, UEs can be associated with a maximum of one BS at a time, and while the reward function aims to maximize the total downlink rate, it does not assess whether certain UEs are being served disproportionately, nor does it take into account the QoS requirements.

Mahbub *et al.* (2021) addresses maximizing user association probability and the average number of associated user devices in HetNets. The proposed solution addresses the optimal allocation of resources under network constraints, such as power, layer density, and UE density. They formulate the problem as an optimization problem and the results reveal several significant trends. With increasing transmission power of the specific layer, the UE association probability increases, especially for a lower SINR threshold. However, higher SINR targets tend to reduce the probability of user association. The study's findings may be limited by the absence of influence of thermal noise, the reliance on user and base station distributions assumed to be random, and the lack of consideration of UDN scenarios.

In (Jain *et al.*, 2021), the authors address the challenges

of 5G networks, which will be extremely dense and heterogeneous due to the increase in the number of users, BSs, and various types of applications. The study introduces a joint optimization scheme that formulates the user association strategy as a Mixed Integer Linear Program (MILP), aiming to maximize the total network throughput while optimizing bandwidth allocation and BS selection. The results show that the proposed solution can significantly improve network performance in terms of throughput and system fairness compared to the baseline scenario. Nevertheless, the work employs approximation techniques to derive a convex (and simplified) version of the actual objective function of the problem to ensure the application of combinatorial optimization techniques. However, the formulation of a simplification version tends to generate a problem far from the real problem.

Zhai *et al.* (2024) jointly optimize user association, spectrum allocation, and power allocation in the context of Unmanned Aerial Vehicle (UAV) communications. The main objective is to maximize the sum-log-rate of all UEs in two adjacent BSs. The authors develop a genetic algorithm to optimize UAV positioning, followed by a theoretical analysis to create a low complexity branch and bound algorithm for optimal user association and spectrum allocation. Although promising, the model does not account for channel capacity limitations and user heterogeneity, which can lead to congestion during high-demand scenarios.

Several RL algorithms are explored, either in conjunction with CoMP or in a multiagent strategy. Therefore, this work seeks to address some of the gaps in these works, such as the simplification of the optimization problem. Unlike the papers found in the related works, this work seeks to present a strategy using RL to handle the UARA orchestration complexity in NGNs, considering UE requirements and optimization issues. Thus, our scheme differs from those of other studies in the following ways:

1. Although many previous works focus on RL algorithms in more general contexts, our research specifically addresses the complexity of UARA orchestration in UDN environments. This consideration is crucial, as the unique characteristics of UDNs, such as the high density of BSs, require adaptive and efficient strategies that meet the specific needs of UEs. Thus, our proposal seeks to simplify the optimization problem in a more targeted solution to the challenges faced by UDNs;
2. Most of the reviewed works primarily aim to maximize the total downlink rate without necessarily considering the distribution of this data rate in a dense user scenario. This focus on aggregate performance often overlooks the critical aspect of how the available resources are allocated among users, which can lead to imbalances in service quality and user experience, particularly in environments characterized by high user density. Consequently, there is a need for approaches that enhance the overall downlink capacity and ensure an equitable distribution of resources among users to meet their individual requirements effectively;
3. This work does not adopt a combinatorial optimization strategy. Instead, it employs alternative method-

ologies that focus on developing a more tractable optimization mechanism. By avoiding reliance on approximation techniques to derive a convex and simplified version of the actual objective function, our approach aims to maintain fidelity to the complexities inherent in the problem. This decision allows for a more accurate representation of the real-world scenarios encountered in UDNs, facilitating the development of solutions that are both effective and applicable in practice;

4. Finally, this study adopts a strategy that incorporates a minimum traffic requirement for each UE, thereby acknowledging the distinction between ordinary and priority UEs. By establishing a baseline level of service for all UEs, the proposed approach ensures that the needs of priority users, who are assigned a greater weight in resource distribution, are consistently met. This differentiation is crucial in environments with varying service demands, as it allows a more nuanced allocation of resources to consider that priority users receive the necessary resources to fulfill their QoS requirements.

3 Fundamental Premises

In current generations of mobile networks, service conditions can change rapidly due to UE mobility, time-varying channel and interference conditions, or content popularity issues. Network densification reinforces the existing challenges towards scalable management issues, infrastructure implementation costs, and spectral efficiency. The achievement of scalability and management intelligence tends to be a fundamental goal, as the design of architectures so that NGNs can handle multiple services remains a persistent challenge.

Hence, adaptable and flexible network orchestration capabilities become crucial, considering the accelerated growth in the data traffic projected for the coming years. Conventional resource orchestration tends to become increasingly inadequate in favor of autonomous approaches, given the different levels of accuracy and computational complexity involved. Standard approaches that rely on instantaneous network information, such as Channel State Information (CSI), and focus on optimizing an instantaneous performance metric, can become impractical as real-time computing these metrics becomes challenging.

The promise behind NGN future deployment is the ability to provide flexibility, reconfigurability, programmability to support fine granularity and vast and heterogeneous use cases. By decoupling network functionalities from the underlying hardware, softwarization and virtualization are two disruptive paradigms considered to be the basis of the design process of NGNs. From these technologies, it is possible to slice the resources of the network infrastructure, where each slice has a particular behavior shaped according to the requirements demanded in the slice (Wijethilaka and Liyanage, 2021). Network slicing provides the potential to create custom on-demand slices for different services with heterogeneous QoS requirements. These slices may also be modified or canceled as needed, increasing the flexibility and adaptability of network management processes.

The adaptive behavior of the network slice based on traf-

fic dynamics can be challenging, since resource allocation between different slices aims at resource isolation. Nevertheless, Software-Defined Network (SDN) controllers can operate cooperatively to detect regions with critical traffic conditions for UEs and guarantee improved UE satisfaction levels.

Figure 1 represents a network with a typical Mobile Edge Computing (MEC) scenario. MEC scenarios are promising for adopting UARA orchestration mechanisms, given the potential availability of manageable network assets made possible by SDN technology. In addition, they represent a network architecture with decentralized computing infrastructure, where computing resources are available at the network edge. As shown in Figure 1, the Central Office (CO) shelters the primary MEC node of the network and has the highest processing and storage capacity compared to the other nodes. On the other hand, the local MEC nodes are those closest to UEs and BSs, although they have lower computational capabilities than those observed in the primary MEC node.

As shown, regions of the network with UEs in critical service conditions can be detected so that local SDN controllers can act cooperatively to improve the traffic conditions of these UEs. Considering the representation in Figure 1, the primary MEC node can determine which local SDN nodes should cooperate to mitigate traffic conditions of the affected UEs. The SDN controller may instantiate an ML-powered slice to modify the service conditions by adjusting the UARA mechanisms. One or multiple RL agents can cooperate or compete against each other, rebalancing the network load to fully explore the available network resources while considering user satisfaction as the primary goal.

3.1 UARA Mechanisms

HetNets enable the deployment of SBSs, which have limited transmission power and coverage. Such SBSs can be organized in layers, with different levels of deployment density, transmission power, and service capacity, with advantages related to ease of deployment and reduced operating and maintenance costs (Kuribayashi *et al.*, 2020).

Furthermore, HetNets has the potential to reduce load imbalance among different BSs, given the different service capacities and transmission power options. On the other hand, the cell selection strategy based on the Max-SINR ratio unbalances the load concentration across the network. Even under SBSs coverage, UEs still perceive the most prominent downlink signal as being from MBSs. Eventually, even if a UE is under the coverage area of a SBS, the SINR of the MBS experienced by this UE may be higher than that perceived by the nearby SBS.

On scale, this causes MBSs to remain overloaded, while the layers composed of SBSs remain idle, with a low level of UEs in service. To make better use of the HetNets infrastructure, UEs should be reassigned with less overloaded SBSs so that their QoS requirements are better met, given the potentially higher availability of radio resources. Hence, if a BS is associated with many UEs, considering an equal allocation of resources, the radio resources are divided equally, leading to reduced data rates to the UEs (Gomez *et al.*, 2018).

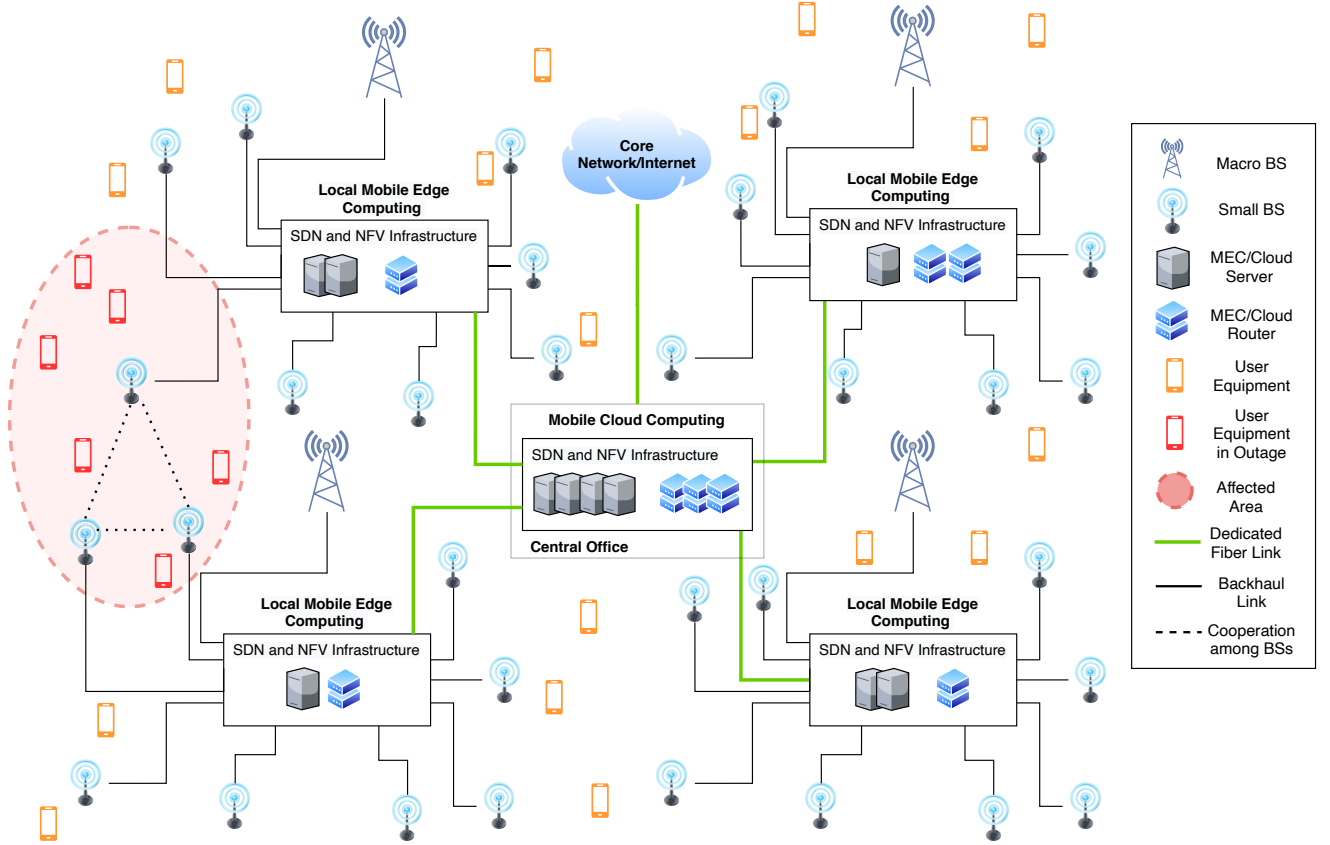


Figure 1. Representation of the application scenario of the proposed scheme.

Consequently, several mechanisms for the joint orchestration of UARA processes for HetNets have been proposed, considering performance metrics such as blocking/coverage probability, spectral efficiency, energy efficiency, asymmetry between *downlink* and *uplink*, among others. Achieving a satisfactory degree of coordination is not an easy task, which requires intelligent mechanisms that consider the traffic load of the network and the network conditions related to the channel, signal, and interference. Such issues represent a task of considerable complexity, given the need for real-time operation and the density of BSs.

Traditional approaches to resource allocation in NGNs have shown limitations due to their lack of intelligence and adaptability Jiang *et al.* (2023); Khani *et al.* (2024). The proposed solutions have been too simplistic or too complex, failing to account for the dynamic nature of NGNs environments in UDN scenarios. Hence, this work evaluates the adoption of an RL-based strategy to guide the orchestration of UARA mechanisms. These techniques are chosen because of their relatively low complexity, which is facilitated by recursive learning based on feedback and local interactions. This characteristic makes them particularly suitable as they can adaptively respond to the dynamic conditions of the network environment.

To address the increasing complexity and variability inherent in NGN environments, the proposed mechanism aims to use adaptive learning principles to optimize performance metrics such as user satisfaction. An RL-based approach has the ability to learn from continuous interactions with the network, which allows the development of strategies that evolve over time, ensuring that resource allocation remains

effective even as user demands and environmental conditions fluctuate. This adaptability is crucial in UDN settings, where traditional static methods struggle to accommodate rapid changes.

Hence, based on the premises and opportunities exposed, we believe that it is possible to establish a ML-based strategy capable of orchestrate UARA mechanisms in NGNs.

4 Problem Formulation

The formulation of the problem is based on the maximization of the degree of satisfaction of the UEs, more precisely when a UE reaches a data rate r_i that meets its minimum traffic requirements r_i^{min} ($r_i^{min} \leq r_i$). The set of BSs is defined by \mathcal{B} , while the set of UEs is denoted by \mathcal{U} . Furthermore, we assume BSs are orchestrable agents of the system and can be used to execute actions that meet the traffic requirements of the UEs satisfactorily met. Thus, the degree of satisfaction of all UEs associated with the j -th BS is denoted by $\Gamma_j(t)$, as:

$$\Gamma_j(t) = \frac{\sum_{\forall i} x_{ij}(t) \varphi_i \psi_i}{\sum_{\forall i} \psi_i}, \forall j \in \mathcal{B}, \quad (1)$$

where $x_{ij}(t)$ represents the binary association state between the i -th UE and the j -th BS at time t . If $x_{ij}(t) = 1$, there is an association between UE and BS, while the opposite case is indicated by $x_{ij}(t) = 0$. Similarly, φ_i represents a binary variable that illustrates the eventual fulfillment of traffic requirements of the i -th UE. When $r_i^{min} \leq r_i$, $\varphi_i = 1$, and otherwise $\varphi_i = 0$.

Table 1. Symbols and Acronyms of the System Model.

Parameter	Description [measurement unit]
i	i -th UE
j	j -th BS
r_i	Downlink rate of the i -th UE [Mbps]
r_i^{\min}	Minimum Downlink rate of the i -th UE [Mbps]
\mathcal{B}	Set of BSs
\mathcal{U}	Set of UEs
$\Gamma_j(t)$	Satisfaction level of UEs associated with the j -BS [%]
$x_{ij}(t)$	Association between the i -th UE and the j -th BS
φ_i	Fulfillment of traffic requirements of the i -th UE
ψ_i^{ordinary}	Weight factor for ordinary users
ψ_i^{priority}	Weight factor for priority users
$\Phi(t)$	Cost Function at time t
Υ	Negative reward factor
ρ	Positive reward factor
$R_j(t)$	Reward associated with the j -th BS at time t
\mathcal{R}_t	Weighted sum of the rewards obtained at time t
γ_t	Discount rate at time t
M_{BS}	Maximum number of associated BSs per UE
M_{UE}	Maximum number of associated UEs per BS
n_i^{RRB}	Number of RBs received by a single UE
T_B	RB threshold parameter
K	Number of independent layers in HetNet
ζ_{ij}	Downlink SINR at i -th UE from j -th BS
P_j^k	Transmit power of the BS j at layer k [dBm]
h_{ij}	Effective gain channel between i -th UE and j -th BS [dBi]
P_N	Thermal power noise [dBm/Hz]
R_i	Per-channel downlink rate at i -th UE from j -th BS [Mbps]
e_ℓ	Per-subcarrier efficiency
n_{sc}	Number of subcarriers per channel
n_{sym}	Number of OFDM symbols per subframe
T_{subframe}	Subframe duration [ms]

Furthermore, ψ_i denotes a weighting factor in calculating the degree of satisfaction when considering priority and common UEs. Hence, we assume that there are certain UEs that have the priority of service. This priority has a direct effect on the resource allocation mechanisms, as well as on the calculation of the degree of satisfaction of the UEs. In particular, condition $\psi_i^{\text{priority}} > \psi_i^{\text{ordinary}}$ is considered, where ψ_i^{priority} denotes the priority of privileged UEs and ψ_i^{ordinary} represents the priority of ordinary UEs. Thus, the balance between priority and common UEs has implications for the computation of $\Gamma_j(t)$, given that $\Gamma_j(t)$ is directly proportional to the term $\varphi_i \psi_i$.

Therefore, local network controllers through BSs can take actions that alter the UARA mechanisms. In this context, although the general objective of the orchestration is to improve traffic conditions, eventually a set of actions can lead to a degradation of the degree of satisfaction of the UEs involved. Thus, the proposed methodology establishes a cost function $\Phi(t)$, which seeks to assign a penalty in such situations:

$$\Phi_j(t) = \begin{cases} -\Upsilon, & \text{Se } \Gamma_j(t) < \Gamma_j(t-1), \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where Υ represents a constant that acts as a negative reward when there is a decrease in the satisfaction of the UEs. From the above, a utility function $R_j(t)$ is considered that represents the reward associated with the j -th BS at time t :

$$R_j(t) = \sum_j \rho \Gamma_j(t) + \Phi_j(t), \quad (3)$$

where ρ represents a reward factor associated with the degree of satisfaction of the UEs. Furthermore, the proposed mechanism

computes the reward obtained by a network controller considering the rewards obtained by all BSs associated with this controller. Thus, from Eq. (3), a weighting factor of the rewards obtained at each instant t is defined, and finally, the objective function presented by Eq. (4) is formulated. Thus, the proposed mechanism seeks to maximize the rewards \mathcal{R}_t , obtained in the long term by the network controller through the weighted sum of the rewards obtained at each instant t :

$$\text{Maximize } \mathcal{R}_t = \sum_t \sum_j \gamma_t R_j(t), \quad (4)$$

where γ_t represents a discount rate that determines the weight of future rewards. Thus, network controllers should choose actions that maximize long-term function (4). This definition of discounted pay-off causes controllers to value immediate rewards more highly than future rewards since γ_t is defined in the range $[0, 1]$. Although the controller considers the rewards it expects to receive in the future, the more immediate rewards have more influence when deciding which action to take. Finally, the objective function defined by Eq. (4) is subject to the following constraints:

$$\sum_{j \in \mathcal{B}} x_{ij} = M_{BS}, \forall i \in \mathcal{U}, \quad (5)$$

$$\sum_{i \in \mathcal{U}} x_{ij} \leq M_{UE}, \forall j \in \mathcal{B}, \quad (6)$$

$$n_i^{RRB} \geq T_B, \forall i \in \mathcal{U}. \quad (7)$$

The restriction (5) ensures that each UE is associated with at most M_{BS} BSs simultaneously. On the other hand, the restriction (6) seeks to ensure that each BS serves at most M_{UE} UEs at the same time. Finally, the constraint (7) ensures that the number of RBs received by a single UE must be greater than a minimum threshold T_B , in order to simulate the concept of a transport block.

4.1 System Model

We employ the distance-based path loss model and simulation parameters recommended by 3GPP (3GPP TR 36.814 V9.0.0, 2010). Performance variations are observed based on Monte Carlo simulations aimed at representing the long-term behavior of the proposed scenario. The system model adopts a HetNet, formed by K independent layers of BSs. The set of all BSs is defined by \mathcal{B} , while the set of all UEs is denoted by \mathcal{U} . By the association criterion based on the highest received signal power (Max-SINR), the i -th UE associates with the j -th BS, starting from the highest SINR value ζ_{ij} received, such that $j = \arg \max(\zeta_{ij}), \forall j \in \mathcal{B}$, as follows:

$$\zeta_{ij} = \frac{P_j^k h_{ij}}{\sum_{q \in \mathcal{B}, q \neq j} P_q^k h_{iq} + P_N}, \forall j \in \mathcal{B}, \quad (8)$$

where P_j^k denotes the transmit power of BS j at layer k , h_{ij} represents the effective channel gain between UE and BS, while P_N represents the thermal noise power. From Eq. (8), the achievable per-channel downlink rate at the i -th UE from the j -th BS can be expressed as:

Table 2. Threshold SINR (dB) to Efficiency e_ℓ (Bits/Symbol).

SINR \leq	-6.5	-4.0	-2.6	-1.0	1.0	3.0	6.6	10.0	11.4	11.8	13	13.8	15.6	16.8	17.6
e_ℓ	0.15	0.23	0.38	0.6	0.88	1.18	1.48	1.91	2.41	2.73	3.32	3.9	4.52	5.12	5.55

$$R_i = e_\ell \cdot \frac{n_{sc} \cdot n_{sym}}{T_{subframe}} \quad (9)$$

where e_ℓ represents an efficiency function per subcarrier in terms of bits per Orthogonal Frequency-Division Multiplexing (OFDM) symbol for a given threshold SINR $e_\ell(\zeta_{ij})$. We use the 15-rate Modulation Coding Scheme (MCS) available in Long Term Evolution (LTE), as shown in Table 2, to parameterize the rate function e_ℓ , according to Equations (8) and (9). The terms n_{sc} , n_{sym} , and $T_{subframe}$ represent the number of subcarriers per channel, number of OFDM symbols and duration of a subframe, respectively. Thus, the data rate obtained by the i -th UE from the j -th BS can be computed as:

$$r_{i,j} = e_\ell \cdot n_{i,j}^{RB} \cdot x_{i,j} \cdot \frac{n_{sc} n_{sym}}{T_{subframe}}, \forall j \in \mathcal{B}, \quad (10)$$

where $n_{i,j}^{RB}$ represents the amount of RBs available by the j -th BS to the i -th UE. Taking into account a fair resource allocation scheme, in which the total number of RBs is equally divided between the associated users, the total number of RBs obtained by the i -th UEs from the j -th can be expressed as

$$n_{i,j}^{RB} = \left\lfloor \frac{n_j^{RB}}{L_j} \right\rfloor, \quad (11)$$

where n_j^{RB} represents the total number of RBs available at j -th BS, while L_j denotes the load of the j -th BS, i.e., the total number of UEs associated with the j -th BS. The notation $\lfloor \cdot \rfloor$ represents the floor function that gives as its output the greatest integer less than or equal to $\frac{n_j^{RB}}{L_j}$ to ensure that the number of RBs ($n_{i,j}^{RB}$) is an integer value.

We assume UEs are classified into common and priority ones. In this case, priority UEs have a greater weight for RBs allocation in a BS than common UE. Additionally, a common UE can have its minimum QoS requirements relaxed, while priority UE must always have their QoS requirements met. However, the traffic conditions of a priority UE depend directly on the load of the BS, as well as the number of other priority UE associated with the same BS. To facilitate understanding, Table 1 summarizes the key symbols and acronyms utilized throughout this work.

4.2 RL-Based UARA Mechanism

The main objective of this work is to optimize the objective function defined by Eq. (4). To this end, we first formalize the problem through a stochastic game and then present the RL strategy considered in this work.

Game Formulation. Through a local SDN controller (as shown in Fig. 1), we consider it possible to consolidate the satisfaction level of all UEs involved in an area with critical service conditions. In this context, we consider that BSs must act cooperatively to maximize the rewards obtained in

the long term, according to Eq. (4). Then, we compute the reward matrix experienced by each BS, considering the satisfaction value of the UEs.

Thus, at each instant t , the reward of each BS is based on the current state of the system (satisfaction of UEs) and on the actions of the other BSs. Consequently, the next state of the network assumes a stochastic behavior since, at instant $t+1$, the new state of the network is influenced by the previous state and the actions taken by the BSs. Consequently, we can formalize the optimization problem as a stochastic game in terms of a Markov Decision Process (MDP) with finite states, defined by the tuple $(\mathcal{S}_m, \mathcal{A}_j, \mathcal{P}_{ss'}(\vec{A}_t^m), \mathcal{R}_t)$ as follows:

- \mathcal{S}_m is the discrete set of possible states of the m cluster, where $\mathcal{S}_m = (s_1, s_2, \dots, s_{s'})$ and s' denote the total number of possible states. Since the module aims to maximize the UE satisfaction level, the set \mathcal{S}_m should be defined from the possible values of R_t . Therefore, there is a need to discretize the R_t values in s' intervals;
- \mathcal{A}_j is the set of possible actions of the j th BS, where $\mathcal{A}_j = (a_1, a_2, \dots, a_n)$, and n represents the total of possible actions for each BS;
- \vec{A}_t^m denotes a vector of joint actions of all BS in m cluster, at time t ;
- $\mathcal{P}_{ss'}(\vec{A}_t^m)$ represents the probability of transition from s state to s' , through joint actions \vec{A}_t^m ;
- \mathcal{R}_t is the reward associated with each BS after the transition of system states;

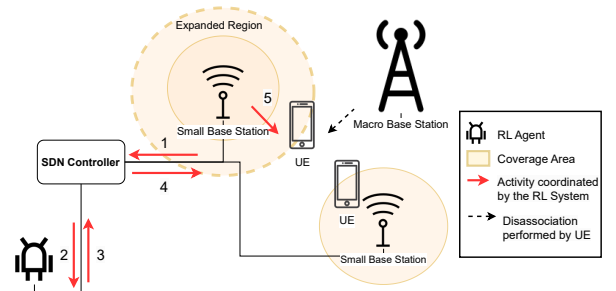
**Figure 2.** Representation of the proposed RL mechanism.

Figure 2 illustrates the application of the RL-based UARA mechanism. Step 1 represents the collection of network information, such as UE satisfaction and the network load on each SBS. This information is used to compute the current system state \mathcal{S}_m . In Step 2, the SDN controller forwards this information to the RL agent, which computes the joint action vector \mathcal{A}_j to redistribute the UE load within the network (Step 3). The reference implementation of this work employs Cell Range Expansion (CRE) to adjust the user association. Hence, the joint action vector \mathcal{A}_j consists of various CRE bias values that modify the signal reception perceptions by the UEs. However, future work may eventually adopt other techniques to parameterize \mathcal{A}_j . During Step 4, the action

vector \mathcal{A}_j is applied to the network, resulting in changes in coverage areas. Step 5 illustrates that a particular UE, previously associated with a MBS, becomes associated with a SBS due to the SINR received from this SBS, along with the potentially greater availability of radio resources. Finally, the process continues iteratively until the average satisfaction of the UEs reaches a predefined minimum threshold established by the network.

Reinforcement Learning Strategy. At each instant t , the j -th BS determines its next action $a_{n,t}$, based on the current state of the network $s_{s',t}$, and through a policy π , such that $a_{n,t} = \pi(s_{s',t})$. From the joint effect of the actions of all BSs \vec{A} , rewards or penalties are assigned to the BSs as a state-action function $Q^\pi(s_{s',t}, a_{n,t})$, expressed as follows:

$$Q^\pi(s_{s',t}, a_{n,t}) = E^\pi \left[\sum_t \gamma_t \mathcal{R}_j(t | s_{s',t}, a_{n,t}) \right], \quad (12)$$

where the term $E^\pi[\cdot]$ represents the expectation operator, while γ_t represents the discount factor at time t . Furthermore, the choice of BSs' actions cannot be completely determined at time t , given the spatio-temporal variation of the UEs. Thus, $Q^\pi(s_{s',t}, a_{n,t})$ represents a long-term mathematical approximation of the rewards and penalties obtained by the BSs, given the stochastic behavior of the system.

Thus, the proposed objective is to discover an optimal action policy that maximizes the rewards obtained in the long term. For all actions selected by the strategy $\pi(s_{s',t}) = \arg \max Q(s_{s',t+1}, a_{n,t+1})$, the term $a_{n,t} = \pi(s_{s',t})$ tends to maximize the values of the function $Q^\pi(s_{s',t}, a_{n,t})$. In order to solve the proposed MDP, the Q-learning technique has been widely used in the literature, given the difficulty in obtaining information about the transition probabilities $\mathcal{P}_{ss'}(\vec{A}_t^m)$. Using Q-learning, the function $Q(s_{s',t}, a_{n,t})$ is now recursively computed as follows:

$$Q(s_{s',t}, a_{n,t}) = Q(s_{s',t}, a_{n,t}) + \lambda [\mathcal{R}_{t+1} + \gamma \max Q(s_{s',t+1}, a_n) - Q(s_{s',t}, a_{n,t})], \quad (13)$$

where λ denotes the learning rate, \mathcal{R}_{t+1} represents the expected reward at time $t+1$, and $\max Q(s_{n,t+1}, a)$ denotes an estimate of the optimal values of the function Q for state $s_{n,t+1}$.

On the other hand, as the dimensionality of the set $\mathcal{S}_m \times \mathcal{A}_j$ increases, the search for optimal strategies tends to become more intensive. In order to achieve convergence, a considerable number of iterations are required to populate the table Q . We consider using deep learning-based reinforcement learning methods to address this issue. In such methods, Deep Neural Network (DNN) is commonly used to represent the state-action space, such that the function $Q(s_{s',t}, a_{n,t})$ is approximated by the DNN, as follows:

$$Q(s_{s',t}, a_{n,t}) \approx Q^*(s_{s',t}, a_{n,t}, \theta). \quad (14)$$

In Eq. (14), the term θ represents a DNN weight parameter. Furthermore, since $Q^*(s_t, a_t, \theta)$ represents an approximate form of $Q(s_t, a_t)$, it is necessary to train the DNN to

minimize the loss function $L(\theta)$ of this training process by updating the weight θ . Consequently, the term $L(\theta)$ can be expressed as:

$$L(\theta) = E[(y^{DQN} - Q^*(s_t, a_t; \theta))^2], \quad (15)$$

where $y^{DQN} = \mathcal{R}_t + \gamma \max Q(s_{t+1}, a_{t+1}; \theta^-)$, and the term θ^- denotes the weight parameter of the target network. In this case, we consider the target network to be a duplicate of $Q(s_t, a_t)$, but with constant weights for a certain number of iterations. Furthermore, to balance the exploratory process, we adopted a simple ϵ -greedy policy, with the ϵ factor decreasing as a function of the maximum number of iterations of the mechanism.

Algorithm 1: Algorithm for Joint User Association and Resource Allocation

```

1 Input: Set of possible actions  $\mathcal{A}_j$ ;
   Result: Optimal action sequence/Policy  $\pi$  for
           achieve the QoS requirements of all UEs.
2 begin
3   Initialize the policy network  $Q(s, a, \theta)$  with
     random weights  $\theta$ ;
4   Clone the policy network, to the target network
      $Q^*(s_t, a_t, \theta^-)$ ;
5   Initialize the system environment, and receives
     the initial state by considering the UE's
     satisfaction index;
6   for  $episode \leftarrow 1$  to  $MaxEpisodes$  do
7     Each BS selects an action  $a_n^j$  via exploration
       or exploitation, using  $\epsilon$ -greedy policy from
        $Q(s, a, \theta)$ ;
8     Local SDN controller consolidate the joint
       actions vector  $\mathcal{A}_j$ , and executes the selected
       actions;
9     Each BS in the clusters reports the service
       condition of its associated UEs, to the local
       SDN controller;
10    Local SDN controller computes the UE's
       satisfaction index, observe the next state;
11    Calculate loss between output Q-values and
       target Q-values, using Eq. (15);
12    Updates weights in the policy network to
       minimize  $L(\theta)$ ;
13    After  $t_{pass}$  time steps, weights in the target
       network are updated to the weights in the
       policy network;
14  end
15 end

```

Algorithm 1 describes the operation of the RL mechanism proposed in this work. Lines 3-5 initialize the proposed mechanism, while the main loop is executed between lines 6-16. The loop (line 6) iterates the algorithm up to a maximum number of episodes $MaxEpisodes$. For each step, the BSs must execute actions based on $Q(s, a)$ to adjust the UARA mechanisms. The ϵ -greedy policy is used to select actions, as per line 7. The local SDN controller consolidates all necessary actions and orchestrates the BSs to execute the

selected actions (line 8). Eventually, there may be reassociation of UEs to new BSs, consequently, a change in the network load balance. Thus, each BS reports to the SDN controller the service conditions of its associated UEs (line 9). In line 11, the SDN controller computes the degree of satisfaction of the UEs involved, thus determining the next state of the network. Between lines 11-13, the weights of the policy and target network are updated to minimize $L(\theta)$. At the end of the main loop execution, the algorithm returns an optimal action selection policy based on the current state.

5 Simulations

In this section, we present the simulation models used for performance evaluations. Then, we describe the numerical results obtained.

5.1 Simulations Models

The simulations were performed using Python 3.7.17, taking advantage of its robust programming capabilities, which facilitate seamless integration with RL algorithms, thereby minimizing the dependency on specialized network simulation tools. All RL implementations were based on the use of the stable-baselines3 library (Raffin *et al.*, 2021), which provides a comprehensive suite of functions and frameworks tailored for developing and optimizing various RL models.

We consider a two-layer HetNet model ($K = 2$), with BSs uniformly distributed over a scenario of 1.0 km^2 , to simulate a deployment strategy driven by the mobile network operator. In particular, the simulation scenario has only 1 MBS and 10 SBSs. We also consider up to 300 UE/ km^2 , whose positions are randomly generated by independent samples of a non-homogeneous poisson point process function, to simulate the real positioning of UEs. Furthermore, MBS, UEs, and SBSs have fixed heights of 30.0, 1.5 and 10.0 meters, respectively. We assume UEs positions as fixed, and thus there is no mobility of this UEs in the simulation scenario.

Each UE has a traffic requirement r_{min} of 2.0 Mbps. Furthermore, it is assumed that a proportion of 20% UEs have priority traffic requirements, while the others assume an ordinary traffic profile. The weights for calculating UEs satisfaction are $\psi_i^{priority} = 1.5$ and $\psi_i^{ordinary} = 1.0$, for the priority and common UEs respectively. These priority UEs are chosen randomly among the UEs that compose the simulation.

The configured transmission power in the system is 46.0 dBm for MBS, 23.0 dBm for SBSs and -174.0 dBm/Hz for thermal noise. In addition, the coding and modulation scheme available in LTE is considered, according to the data presented in Table 2, which presents the efficiency values e_ℓ as a function of the SINR values ζ_{ij} . Finally, Tables 3 and 4 summarize the major simulation parameters.

From the proposed scenario, UEs associate with at most $M_{BS} = 2$ BSs simultaneously based on the Max-SINR criterion, while a BS must serve up to $M_{UE} = 20$ UEs simultaneously. Once this process is completed, the load of each BS and the data rate obtained by each UE (according to Eq. 10) are computed, as well as the satisfaction level of each

Table 3. Physical-layer parameters.

Parameter	Value
MBS power transmission	46.0 dBm
SBS power transmission	23.0 dBm
Carrier frequency	2.0 GHz
$T_{subframe}$	1 ms
n_{sc}	12
n_{sym}	14
n_j^{RB}	100
MBS antenna gain	15.0 dBi
Noise power	-174.0 dBm/Hz
Subchannel BW	180 kHz
UE antenna gain	0.0 dBi
SBS antenna gain	5.0 dBi
MBS-UE path loss	$128.0 + 37.6 \log_{10}(\max(d, 35)/1000)$
SBS-UE path loss	$140.7 + 36.7 \log_{10}(\max(d, 10)/1000)$

Table 4. Simulation parameters.

Parameter	Value
Area	1.0 km^2
Number of simulations	100
Number of tiers (K)	2
MBS density	$1.0/\text{km}^2$
SBS density	$10.0/\text{km}^2$
UE density	up to $300.0/\text{km}^2$
Height of antenna for MBS, SBS and UE	30.0/10.0/1.5 meters

UE. Finally, in this reference implementation, the restriction defined by Eq. 7 was relaxed ($T_B = 0$). Relaxation was adopted to facilitate the theoretical analysis of the model, allowing a clearer understanding of the fundamental interactions within the networks before introducing additional complexities that could obscure the initial results.

At each instant t , the reward \mathcal{R}_t of each BS is based on the current state \mathcal{S}_m . Thus, in the context of RL, the next state of the network assumes a stochastic behavior, since at instant $t + 1$, the new state of the network is influenced by the previous state and by the actions \mathcal{A}_j . The mechanism computes these actions based on the application of CRE bias in the continuous interval of $[20.0, 80.0]$ db. Thus, the j -th BS determines its next action $a_{n,t}$ ($a_{n,t} \in \mathcal{A}_j$), based on the current state of the network $s_{s',t}$, and through a policy π , such that $a_{n,t} = \pi(s_{s',t})$. From the joint effect of the actions of all BSs \vec{A} , rewards are assigned to the BSs, as a state action function $Q^\pi(s_{s',t}, a_{n,t})$. In all actions selected by $\pi(s_{s',t}) = \arg \max Q(s_{s',t+1}, a_{n,t+1})$, $a_{n,t} = \pi(s_{s',t})$ tends to maximize the function $Q^\pi(s_{s',t}, a_{n,t})$.

To approximate a solution for $\pi(s_{s',t})$, this work adopts the algorithms of Advantage Actor Critic (A2C), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), Twin Delayed DDPG (TD3) and Soft Actor-Critic (SAC). This work selected these algorithms due to the continuous nature of the action set \mathcal{A}_j , thus excluding algorithms that operate solely with discrete action sets, such as Deep Q-Networks (DQN).

Furthermore, although initial evaluations indicated suboptimal results from A2C and PPO, our intention was to assess their performance in this specific context. This foundational analysis will inform future research directions, allowing us to refine our approach and explore alternative algorithms to enhance performance outcomes in the orchestration of UARA mechanisms. In addition, each algorithm was run in a set-up of 100 repetitions to facilitate a comprehensive and robust

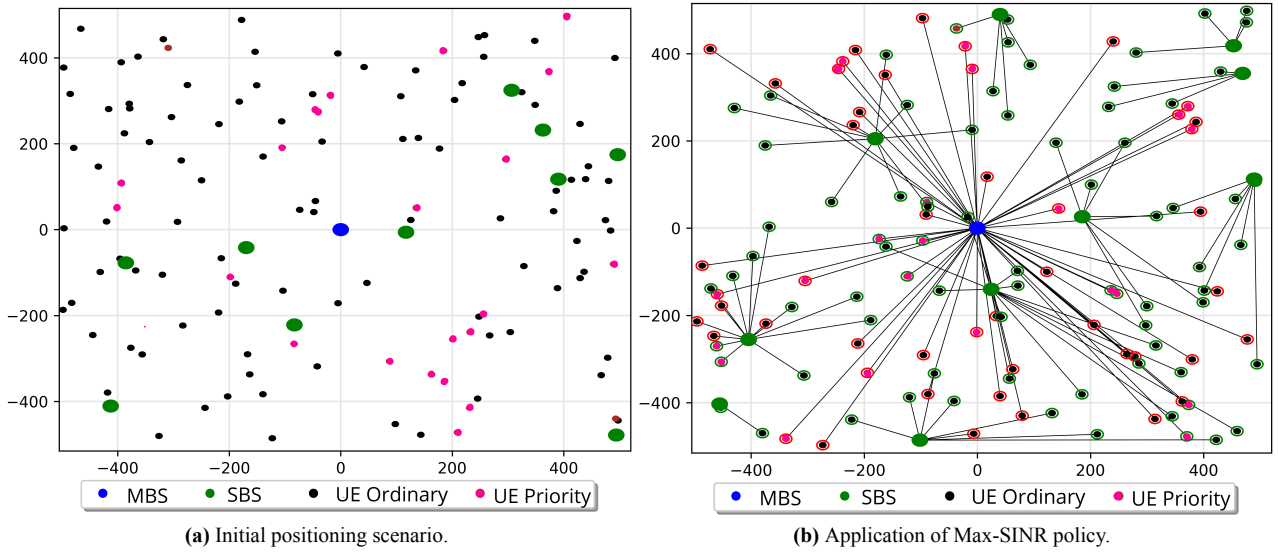


Figure 3. Initial user distribution and association scenarios after applying the Max-SINR policy.

analysis of the results obtained, approximating a stationary distribution. Finally, the main hyperparameters used in the simulations were $t = 100000$ (total iterations), $\gamma = 0.9$ (discount factor), $\lambda = 0.001$ (learning rate).

5.2 Numerical Results

Figure 3 complements the description of the simulations performed in this work to illustrate the simulation scenario. In this representation, UEs are identified by black or red circles (when prioritized), while SBSs is identified by green circles and MBS is represented by a blue circle. Figure 3a presents a representation of the main elements of the simulation and their spatial distribution.

Furthermore, Figure 3b visually represents the association and load balancing status after implementing the 3GPP metric based exclusively on Max-SINR. In this context, UEs can be marked with green or red borders. When a UE reaches a data rate higher than the minimum required ($r_i^{min} \leq r_{ij}$), there are green borders indicating this. In contrast, circles with red borders represent UEs that were unable to meet their minimum traffic requirements ($r_i^{min} > r_{ij}$). Thus, circles with red borders are undesirable and highlight situations where the system infrastructure could not meet the specific traffic requirements of the UEs.

From the results presented in Figure 3b, it becomes evident that there is a significant number of UEs without their QoS requirements met, totaling an average value of 56.10% over the 100 independent simulations. Furthermore, even with the presence of nearby SBSs, it is possible to see that some UEs end up associating with the MBS, overloading it and thus reducing the average RBs per UE. Consequently, in this Max-SINR scenario, an average 48.5% of UEs with met traffic requirements is recorded.

On the other hand, Figure 4a shows the result of applying a unified Bias CRE of 40.0 db to all SBSs. In this scenario, the SBSs behaves as a homogeneous layer in such a way as to force the association of UEs to this layer, thus reducing the load on the MBS layer. In this scenario, it is possible

to visualize a significant reduction of UEs associated with the MBS, causing that, on average, 79.91% have their QoS requirements met, according to the data in Figure 5d.

Furthermore, Figures 4b and 4c present the applications of the A2C and PPO algorithms, respectively. It is important to note that these algorithms present lower results than those observed in the unified CRE bias strategy. Even with the evaluation of various configuration parameters or adjustments in the training timestamps, these algorithms cannot approximate an optimal solution for Eq. 4, being the most inferior among all RL algorithms. Both methods aim to improve the efficiency and stability of training by balancing exploration and exploitation, using actor and critic policies to update the action policies incrementally. This trend may indicate the infeasibility of using these policy-based algorithms in the scenario proposed in this work.

Figure 4d presents the association results for the TD3 algorithm. In this scenario, it is possible to verify the significant number of UEs with met traffic requirements and the intense offload of the MBS towards the SBSs. It is possible to verify that the SBS layer is responsible for serving most of the UEs and not by the MBS, where an average value of 99.20% of UEs with met traffic requirements. Only 15.87% (average) of the UEs are associated with the MBS.

In addition, Figures 5a and 5b present the results of the SAC and DDPG algorithms, respectively. These algorithms share several fundamental similarities with the TD3 algorithm since the latter employs deterministic policy gradient methods, which aim to learn deterministic policies to directly map states to actions, which is especially effective in continuous action spaces (as defined by the set \mathcal{A}_j).

Furthermore, Figures 5c and 5d present the results of applying the RL algorithms in terms of satisfaction and average UEs per SBS and a boxplot graph with the percentage of satisfaction of the UEs in each of the scenarios/models evaluated in this work. Based on Figure 5c, it is possible to verify that the schemes that have promising levels of UEs satisfaction are those that have the highest average percentages of UEs per SBS, in such a way as to alleviate the load of the MBS.

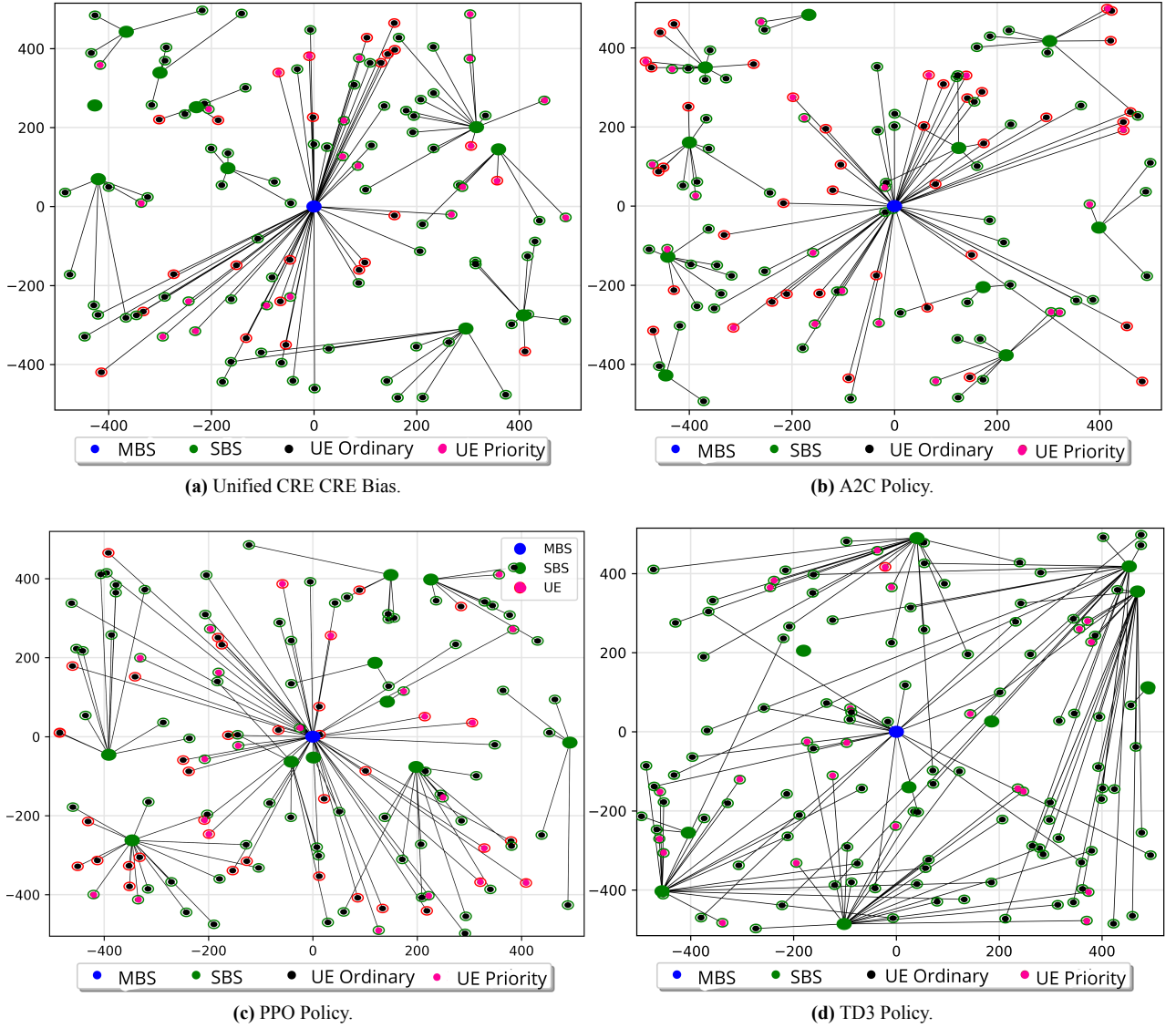


Figure 4. User distribution and association scenarios after applying CRE, A2C, PPO, and TD3 techniques.

According to Figure 5d, it is possible to observe that using a unified CRE bias, set at 40.00 dB, resulted in satisfaction rates between 67% and 94%. While A2C obtained satisfaction rates between 71% and 87%, and PPO, rates ranging from 55% to 75%. On the other hand, the algorithms DDPG, SAC, and TD3 stood out significantly from the others, presenting the best results. These algorithms demonstrated user satisfaction rates consistently above 90% and, in some cases, up to 100%. These results emphasize the promising effectiveness of RL-based approaches, indicating their potential for orchestrating UARA mechanisms in NGNs.

On the other hand, when analyzing Figure 5d, it becomes evident that the proposed methodology is promising in orchestrating UARA processes since all RL-based approaches presented higher satisfaction values than those observed in the Max-SINR approach. The results obtained indicate that reducing the total number of associations with the MBS paves the way to this improvement, allowing the SBSs to assume the responsibility of serving a larger contingent of UEs. However, future studies should clarify the reason for the low efficiency of algorithms based on actors and critics.

Finally, Table 5 complements the results previously presented by showing the cumulative data rate achieved by all UEs for each algorithm, taking into account the average value in all simulations conducted. Consistent with the patterns observed in the earlier results, the SAC, DDPG, and TD3 algorithms demonstrate the highest data rate values. This characteristic can be attributed to the superior load balance performed by these algorithms, which more effectively exploit the availability of RBs in the given scenario.

Table 5. Average cumulative data rate of UEs [Mbps].

	Mean	Variance	Std. Deviation
Max-SINR	698.095	4794.577	69.242
Bias	711.785	4762.151	69.008
A2C	682.043	5120.421	71.557
PPO	688.009	4893.055	69.95
DDPG	731.218	10599.837	102.955
SAC	757.62	6201.82	78.751
TD3	838.346	4417.32	66.462

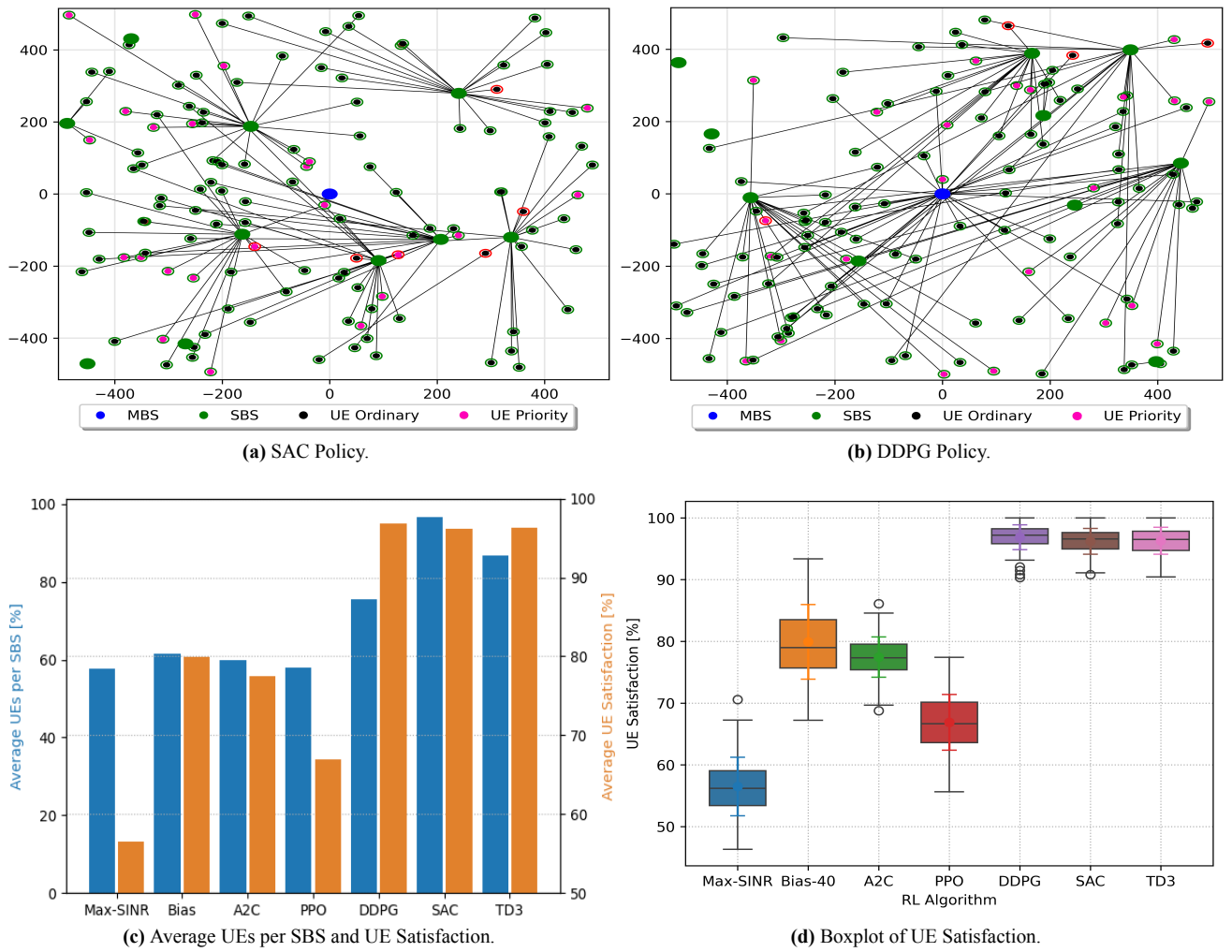


Figure 5. UE satisfaction analysis by RL algorithm.

The proposed approach differs substantially from conventional CRE strategies, which generally adopt a unified bias for SBSs layers. In this work, the proposed methodology performs an individualized association adjustment for each SBSs, allowing UEs to be reassigned with specific (and less overloaded) SBSs, taking into account the resource allocation of each BS. Consequently, UEs are associated with specific SBSs, not specific layers.

6 Conclusion and Future Work

This study investigates the application of algorithmic strategies based on RL with the primary objective of significantly improving the quality of traffic for end users. In particular, this study focuses on the adaptive capacity of these algorithms, specifically on the refined tuning of the UARA mechanisms. This focus becomes particularly crucial when considering the landscape of NGNs, where the increasingly prominent presence of ML techniques highlights the need for innovative operational strategies.

The methodological approach used in this work aims to understand the practical applicability of RL algorithms in the context of next-generation HetNets. The preliminary results collected in this work suggest that the strategic application

of RL algorithms represents a viable path to overcome operational challenges and leverage efficiency in the NGNs, with better results than those obtained by traditional methods (Max-SINR and CRE). The relevance of this research transcends the purely theoretical scope, highlighting the pressing need for adaptive and user-centered strategies in the dynamic scenario of future communications networks.

However, future work will involve the development of a hierarchical approach that enhances the scalability of our stochastic game methodology, enabling more efficient management of high-density user networks by distributing decision-making processes across multiple layers of the network architecture. In addition, this framework will incorporate the evaluation of scenarios with user mobility, which is essential to address the dynamic nature of network environments. By facilitating localized optimization, the hierarchical framework tends to improve overall network performance and adaptability to varying user demands, including those arising from mobile users.

Acknowledgements

Funding

The authors thank the Brazilian National Council for Scientific

and Technological Development (CNPq), the Amazon Foundation for Studies and Research Support (FAPESPA), and the Pro-Rectorate for Graduate Studies, Research, and Technological Innovation (PROPIT) of Unifesspa for the partial financial support granted to carry out this research.

Authors' Contributions

MA is the primary contributor and writer of this manuscript. GB and LL contributed to the implementation of models. MaA and WJ were involved in the conceptualization of this study. HK is responsible for supervision and funding acquisition. All authors read and approved the final manuscript.

Competing interests

The authors have no competing interests in the research developed in the paper.

References

- 3GPP TR 36.814 V9.0.0 (2010). Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9). 3GPP. Available at: https://www.etsi.org/deliver/etsi_tr/136900_136999/136913/09.00.00_60/tr_136913v090000p.pdf.
- Adedoyin, M. A. and Falowo, O. E. (2020). Combination of ultra-dense networks and other 5g enabling technologies: A survey. *IEEE Access*, 8:22893–22932. DOI: 10.1109/ACCESS.2020.2969980.
- Alhashimi, H. F. et al. (2023). A Survey on Resource Management for 6G Heterogeneous Networks: Current Research, Future Trends, and Challenges. *Electronics*, 12(3). DOI: 10.3390/electronics12030647.
- Alzubaidi, O. T. H., Hindia, M. N., Dimyati, K., Noordin, K. A., Wahab, A. N. A., Qamar, F., and Hassan, R. (2022). Interference challenges and management in b5g network design: A comprehensive review. *Electronics*, 11(18). DOI: 10.3390/electronics11182842.
- Attiah, M. L., Isa, A. A. M., Zakaria, Z., Abdulhameed, M. K., Mohsen, M. K., and Ali, I. (2020). A survey of mmWave user association mechanisms and spectrum sharing approaches: an overview, open issues and challenges, future research trends. *Wireless Networks*, 26(4):2487–2514. DOI: 10.1007/s11276-019-01976-x.
- Bikram Kumar, B., Sharma, L., and Wu, S.-L. (2019). Online distributed user association for heterogeneous radio access network. *Sensors*, 19(6). DOI: 10.3390/s19061412.
- Gomez, C. A., Shami, A., and Wang, X. (2018). Machine Learning Aided Scheme for Load Balancing in Dense IoT Networks. *Sensors*, 18(11). DOI: 10.3390/s18113779.
- Jain, A., Lopez-Aguilera, E., and Demirkol, I. (2021). User association and resource allocation in 5g (aura-5g): A joint optimization framework. *Computer Networks*, 192:108063. DOI: <https://doi.org/10.1016/j.comnet.2021.108063>.
- Jayaraman, R. et al. (2023). Effective Resource Allocation Technique to Improve QoS in 5G Wireless Network. *Electronics*, 12(2). DOI: 10.3390/electronics12020451.
- Jiang, W., Feng, D., Sun, Y., Feng, G., Wang, Z., and Xia, X.-G. (2023). Joint computation offloading and resource allocation for d2d-assisted mobile edge computing. *IEEE Transactions on Services Computing*, 16(3):1949–1963. DOI: 10.1109/TSC.2022.3190276.
- Khani, M., Sadr, M. M., and Jamali, S. (2024). Deep reinforcement learning-based resource allocation in multi-access edge computing. *Concurrency and Computation: Practice and Experience*, 36(15):e7995. DOI: <https://doi.org/10.1002/cpe.7995>.
- Kim, D. U. et al. (2023). Resource Allocation and User Association Using Reinforcement Learning via Curriculum in a Wireless Network with High User Mobility. In *2023 International Conference on Information Networking (ICOIN)*, pages 382–386. DOI: 10.1109/ICOIN56518.2023.10048927.
- Kim, Y., Jang, J., and Yang, H. J. (2024). Distributed resource allocation and user association for max-min fairness in hetnets. *IEEE Transactions on Vehicular Technology*, 73(2):2983–2988. DOI: 10.1109/TVT.2023.3316610.
- Kuribayashi, H. P. et al. (2020). Particle Swarm-Based Cell Range Expansion for Heterogeneous Mobile Networks. *IEEE Access*, 8:37021–37034. DOI: 10.1109/ACCESS.2020.2975981.
- Labana, M. and Hamouda, W. (2020). Joint User Association and Resource Allocation in CoMP-Enabled Heterogeneous CRAN. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–6. DOI: 10.1109/GLOBECOM42002.2020.9322501.
- Mahbub, M. et al. (2021). Maximizing the Probability of User Association of a Tier of a Multi-Tier Heterogeneous Network by Optimal Resource Allocation. In *2021 Emerging Technology in Computing, Commun. and Electronics (ETCCE)*, pages 1–6. DOI: 10.1109/ETCCE54784.2021.9689907.
- Naderializadeh, N., Sydir, J. J., Simsek, M., and Nikopour, H. (2021). Resource management in wireless networks via multi-agent deep reinforcement learning. *IEEE Transactions on Wireless Communications*, 20(6):3507–3523. DOI: 10.1109/TWC.2021.3051163.
- Paixão, E. R. et al. (2023). Multilayer Framework for Resource Orchestration in Next Generation Networks. *Journal of Communication and Information Systems*, 38:1–8. DOI: 10.14209/jcis.2023.1.
- Raffin, A. et al. (2021). Reliable Reinforcement Learning Implementations. *Journal of Mach. Learning Research*, 22(268):1–8. Available at: <http://jmlr.org/papers/v22/20-1364.html>.
- Shen, X., Gao, J., Wu, W., Lyu, K., Li, M., Zhuang, W., Li, X., and Rao, J. (2020). AI-Assisted Network-Slicing Based Next-Generation Wireless Networks. *IEEE Open Journal of Vehicular Technology*, 1:45–66. DOI: 10.1109/OJVT.2020.2965100.
- Tanveer, J., Haider, A., Ali, R., and Kim, A. (2022). An overview of reinforcement learning algorithms for handover management in 5g ultra-dense small cell networks. *Applied Sciences*, 12(1). DOI: 10.3390/app12010426.
- Wang, C., Renzo, M. D., Stanczak, S., Wang, S., and Larsson, E.

- E. G. (2020). Artificial intelligence enabled wireless networking for 5g and beyond: Recent advances and future challenges. *IEEE Wireless Communications*, 27(1):16–23. DOI: 10.1109/MWC.001.1900292.
- Wang, L., Ai, Y., Liu, N., and Fei, A. (2019). User association and resource allocation in full-duplex relay aided noma systems. *IEEE Internet of Things Journal*, 6(6):10580–10596. DOI: 10.1109/JIOT.2019.2939875.
- Wang, S., Balarezo, J. F., Kandeepan, S., Al-Hourani, A., Chavez, K. G., and Rubinstein, B. (2021). Machine learning in network anomaly detection: A survey. *IEEE Access*, 9:152379–152396. DOI: 10.1109/ACCESS.2021.3126834.
- Wijethilaka, S. and Liyanage, M. (2021). Survey on network slicing for internet of things realization in 5g networks. *IEEE Communications Surveys & Tutorials*, 23(2):957–994. DOI: 10.1109/COMST.2021.3067807.
- Xu, Y., Gui, G., Gacanin, H., and Adachi, F. (2021). A survey on resource allocation for 5g heterogeneous networks: Current research, future trends, and challenges. *IEEE Communications Surveys & Tutorials*, 23(2):668–695. DOI: 10.1109/COMST.2021.3059896.
- Yazici, I., Shayea, I., and Din, J. (2023). A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems. *Engineering Science and Technology, an International Journal*, 44:101455. DOI: <https://doi.org/10.1016/j.jestch.2023.101455>.
- Zhai, D., Li, H., Tang, X., Zhang, R., and Cao, H. (2024). Joint position optimization, user association, and resource allocation for load balancing in uav-assisted wireless networks. *Digital Communications and Networks*, 10(1):25–37. DOI: <https://doi.org/10.1016/j.dcan.2022.03.011>.
- Zhang, L. et al. (2019). 6G Visions: Mobile Ultra-broadband, Super Internet-of-Things and Artificial Intelligence. *China Communications*, 16(8):1–14. DOI: 10.23919/JCC.2019.08.001.
- Zhao, N., Liang, Y.-C., and Pei, Y. (2018). Dynamic contract incentive mechanism for cooperative wireless networks. *IEEE Transactions on Vehicular Technology*, 67(11):10970–10982. DOI: 10.1109/TVT.2018.2865951.
- Zhao, N. et al. (2019). Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Cellular Networks. *IEEE Transactions on Wireless Communications*, 18(11):5141–5152. DOI: 10.1109/TWC.2019.2933417.