Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.16, No.2, 2026, pp.215-232

DOI: https://doi.org/10.51983/ijiss-2026.16.2.23

Multi-Agent Reinforcement Learning for Joint User Association and Resource Allocation in Heterogeneous Cellular Networks

Pothula Pavan Kumar Reddy[1] and Dr.C. Kamalanathan[2*]

[1]Research Scholar, Electrical Electronics and Communication Engineering, GITAM Deemed to be University, Bengaluru Campus, India

[2*]Associate Professor, Electrical Electronics and Communication Engineering, GITAM Deemed to be University, Bengaluru Campus, India

E-mail: [1]kumarreddy279@gmail.com, [2]kamalanadhan@gmail.com

ORCID: [1]https://orcid.org/0009-0001-3686-413X, [2]https://orcid.org/0000-0003-1579-5670

(Received 02 March 2026; Revised 07 April 2026, Accepted 22 April 2026; Available online 05 June 2026)

Abstract - **The proposed model, MARL-UA-RA (Multi-Agent Reinforcement Learning for User Association and Resource Allocation), is designed to overcome the challenges of heterogeneous cellular networks (HetNets) by leveraging Multi-Agent Reinforcement Learning (MARL) to optimize resource allocation and improve energy efficiency. The proposed framework models each base station as an autonomous cooperative agent that perceives strong local state information involving user signal quality metrics, neighbor cell information, and temporal network features, and jointly optimizes a global network objective via a monotonic QMIX mixing network. The model uses a dual-headed Q-network structure with specific association and resource allocation heads, allowing for joint per-user handover and resource level allocation, subject to an ε-greedy exploration policy with experience replay. User association satisfies the handover margin constraint of 2.5 dB and the maximum cell load of 0.85, while resource allocation dynamically adjusts user throughput and spectral efficiency using a proportional resource fraction approach. A composite reward function incorporating spectral efficiency, signal strength, load balancing, handover cost, joint SINR capacity, and global load fairness facilitates the cooperative learning process. Comparative analysis with D3QN, DQN, Q-Learning, Genetic Algorithm, and MRSP-based approaches as baselines clearly shows that the proposed MARL-UA-RA-QMIX outperforms all in terms of achieving the maximum Average System Capacity of 75.70 and Average Network Utility of 34.39, together with a near-optimal fairness index of 0.99 and minimum load imbalance of 0.5132, thus establishing its superiority in jointly optimizing network capacity, spectral efficiency, energy efficiency, and fairness in dynamic heterogeneous cellular networks. The main benefits of the proposed model are its capability to expand to large networks, better energy efficiency, and the possibility to manage the dynamic conditions of the network, and it is well applicable to real-time use in 5G and other applications.**

Keywords: **Multi-Agent Reinforcement Learning (MARL), User Association (UA), Resource Allocation (RA), Energy Efficiency, Heterogeneous Cellular Networks (HetNet's), Reinforcement Learning (RL), Throughput, Fairness**

I. INTRODUCTION

The rapid introduction and development of 5G networks have raised a variety of challenges, and one of the most urgent issues is the efficient use of resources in heterogeneous networks (HetNets). These networks combine macro and small cells to improve network coverage, capacity, and throughput (Sun et al., 2019; Cheng et al., 2020). Nevertheless, the dynamic nature of HetNets, along with varying traffic loads and resource constraints, requires advanced approaches to resource distribution that minimize energy consumption without affecting service quality (Xiao et al., 2023; Rezvani et al., 2025). Energy efficiency has become a critical issue in designing 5G HetNets due to the growing number of connected devices, as well as the high operational costs and environmental impacts associated with energy consumption (Ding et al., 2020; Tilahun et al., 2023). Conventional resource-allocation methods, which in most cases aim to maximize throughput and latency, cannot address the growing concern about energy consumption. In line with this, the need to incorporate energy-saving methods into the mutual distribution of resources, such as power, spectrum, and bandwidth, is growing (Yin & Yu, 2021; Hugh & Soria, 2025). Multi-Agent Parameterized Deep Reinforcement Learning will be an appealing solution in this area. MARL can enable decentralized decision-making by modeling resource allocation as a multi-agent reinforcement learning (MARL) problem, where agents are the network components (base stations or users) (Ramesh et al., 2025; Wang et al., 2025). The parameters of deep learning models enable the system to adapt effectively to dynamic network conditions and to optimize energy over time (Liu & Ma, 2025; Mayilsamy & Rangasamy, 2021). Using deep learning techniques, MARL can learn from extensive amounts of network data and adjust the resource allocation policy based on feedback from the network environment. Not only is it an efficient way to use available resources, but it also makes the HetNet as a whole more energy-efficient (Kim & So, 2025; Tang et al., 2025). Moreover, MARL can support scalable solutions, which are capable of supporting the nature of 5G

networks, which are both heterogeneous and diverse, and the joint process of resource allocation is effective and sustainable.

Key Contribution

- To model user association and resource allocation as a multi-agent reinforcement learning problem in heterogeneous networks.
- To design and implement a MARL-based framework for efficient user association and resource allocation, where base stations (BSs) or access points (APs) act as autonomous agents.
- To evaluate the performance of the proposed framework under varying network conditions, including user density, mobility patterns, and interference.
- To compare the performance of the MARL-based approach with traditional methods, such as centralized optimization techniques and heuristic approaches, in terms of throughput, energy efficiency, and fairness.

The various sections follow this research paper. Section I introduces the topic; Section II provides a literature review of previous papers and identifies the research gap in existing work. Section III describes the proposed methodology, followed by an overall architecture diagram, the working principles of multi-agent reinforcement learning for Heterogeneous cellular networks, and MARL-based resource allocation, including user association and resource allocation. This also includes user allocation, resource allocation, efficient resource distribution, a multi-agent-based resource allocation algorithm, a data flow diagram of the proposed methodology, and descriptions of the proposed algorithms. Section IV explains the dataset description, simulation setup, Parameter initialization, simulation results, and the performance comparison of various metric analyses, and provides a discussion. Section V explained the main summary and key findings of this research.

II. LITERATURE REVIEW

Research on Joint User Association (UA) and Resource Allocation (RA) in heterogeneous cellular networks (HetNets) is now of particular interest owing to the dynamicity and complexity of current communication systems, particularly the introduction of 5G and beyond (Wang et al., 2021; Nasir & Guo, 2019). These issues were traditionally tackled using optimization techniques such as convex optimization and dual ascent, which helped balance network load and reduce resource use. But these approaches are not always successful in large, dynamic networks, as are computationally intensive and do not account for real-time conditions (Salami et al., 2025).

As mobile data traffic and Quality of Service (QoS) demands grow, traditional centralized methods struggle to scale. This has led to the investigation of Reinforcement Learning (RL) as an instrument for adaptive decision-making (Liang et al., 2023; Luo et al., 2023). The user association and resource allocation problems were solved using single-agent RL algorithms such as Q-learning and Deep Q-Networks (DQN), demonstrating the potential to optimize the network. The methods, however, are limited by their inability to scale to large networks, as do not aim to resolve interactions among agents operating in dynamic environments (Shahzadi et al., 2023).

Multi-Agent Reinforcement Learning (MARL) is a new approach of interest to overcome these limitations. In a MARL model, each network element (base station, user, or network) is represented as a separate learning agent, enabling decentralized and cooperative decision-making (Mishra et al., 2025). This architecture allows agents to consider interactions with other agents and respond to network changes, e.g., traffic changes and mobility (Ma et al., 2024).

It was demonstrated that MARL is efficient at solving joint UA and RA problems, as it allows agents to learn optimal strategies both collaboratively and competitively (Niasi et al., 2025). Multi-agent Q-learning, DQN, and Actor-Critic models have been used to maximize a range of goals, including throughput, energy efficiency and fairness, in HetNets (Pindi & Velez, 2025). These models allow the agents to decide based on the network's current state and evolve over time to reach optimal functioning. Besides enhancing network throughput, MARL frameworks have been used to solve multi-objective optimization problems in HetNets (Urooj et al., 2024). Examples of such areas of concern include fairness amongst users and energy-efficient resource distribution. MARL can result in a more sustainable and equitable network resource management by learning policies that strike a balance between these goals (Alhazmi & Arafah, 2025).

Nevertheless, MARL has achieved numerous successes in HetNets, but several challenges remain. There are problems like non-stationarity (the environment is always changing as agents learn), limited observability (agents do not know everything about the network), and convergence stability (do the agent solutions lead to optimal solutions) that still need to be resolved (Wu & Chen, 2025; Lee & Kim, 2024). In an effort to enhance the performance, scalability, and stability of MARL in HetNets, scholars have been exploring hybrid models and more sophisticated learning algorithms, including hierarchical learning, reward shaping, and agent-agent communication (Alablani & Alenazi, 2023). In MARL, which provides considerable benefits to joint UA and RA in dynamic and heterogeneous environments, current research has been aimed at addressing its issues, especially in ensuring scalability, stability, and efficient coordination across a large population of agents when interacting in a complex network context.

Research Gap

Although Multi-Agent Reinforcement Learning (MARL) has shown promising results in solving Joint User Association (UA) and Resource Allocation (RA) problems in heterogeneous cellular networks (HetNets), there remain

research gaps. Among them are the issues of scalability, where the increased number of agents in large networks is associated with higher computational and communication costs, and convergence stability, where dynamic environments can lead to agents converging to suboptimal policies. Also, there is still the problem of partial observability where agents do not always have full or accurate information of the state of the network. Moreover, better agent coordination, particularly in competition, and multi-objective optimization (trade-offs between fairness, throughput, and energy efficiency) are also issues that need to be further developed. The crucial step to achieving this goal is to overcome these gaps to see the full potential of MARL in large-scale and dynamic HetNets.

III. Proposed Methodology

Overall Architecture Diagram

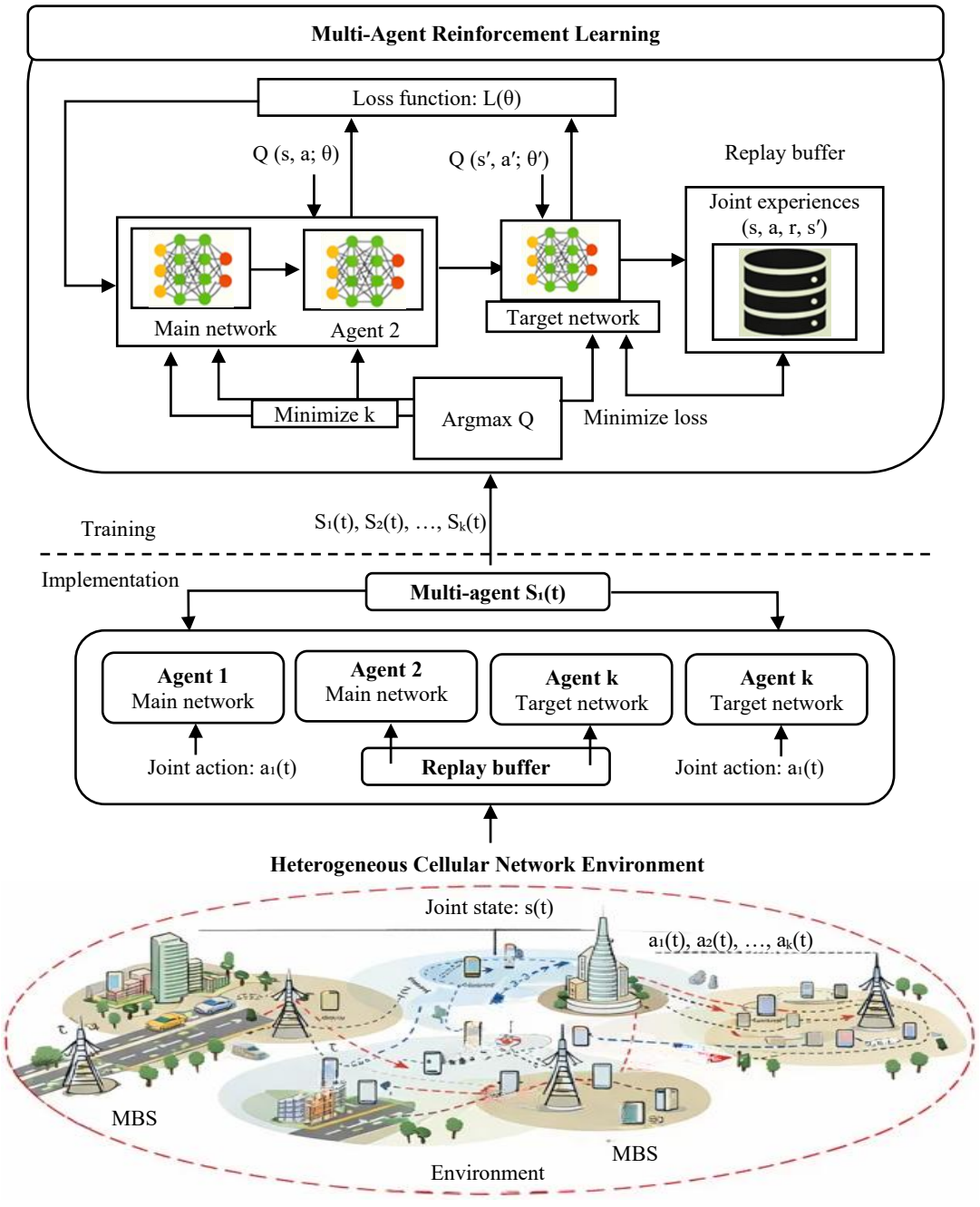

Fig. 1 Overall Architecture Diagram

The Multi-Agent Reinforcement Learning (MARL) architecture shown in this fig. 1 applies to a Heterogeneous Cellular Network (HetNet). In the training step, both a target network and a main network are used by each agent i.e. Agent

1, Agent 2 and so on. The learning and policy refining are done within the main network and the Q-values are measured by the target network, which aids in stabilizing the learning process. The loss is an optimization of the difference between the prediction of the Q-values between a main and target network, which optimizes the decision-making of an agent. A replay buffer is a history of joint experience (states, actions, rewards, and next states) that agents store and sample to update their policies using the argmax Q method. The agents also rely on these experiences to reduce the loss and modify their policies with the help of repeated learning processes. The trained networks are applied to the HetNet environment in the implementation phase where the joint state of the network is observed, and joint actions are chosen to have an impact on user association, resource allocation, and other network parameters. The environment is made up of microcells, picocells, and femtocells, in which there is the interaction between agents and base stations to optimally distribute resources. The process results in a better spectral efficiency, less interference, and better equity of the user, thereby maximizing the overall network performance due to efficient and fairly equitable resource utilization.

Working Principle for Multi-Agent Reinforcement Learning for Heterogeneous Cellular Networks

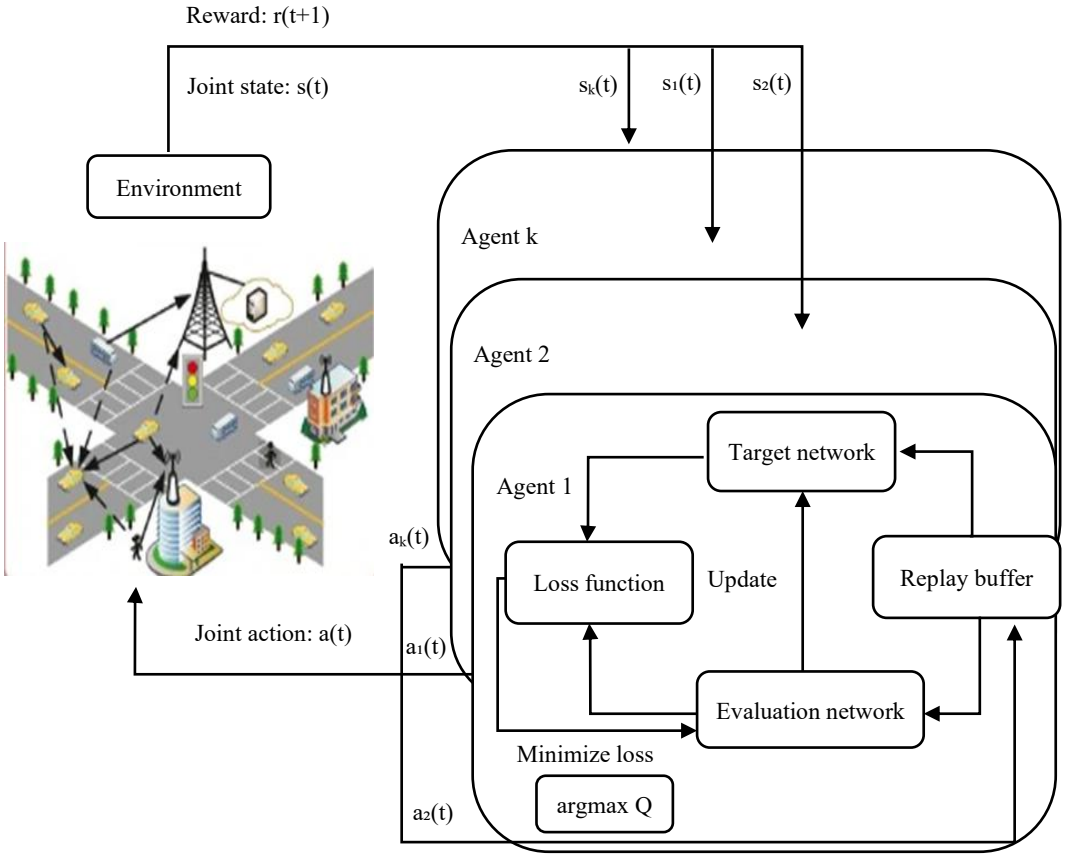

Fig. 2 Working Principle for Multi-Agent Reinforcement Learning for Heterogeneous Cellular Networks

MARL-based resource allocation scheme fig. 2, consists of an overview of RL algorithms, state space, Action space, reward function, and complexity analysis. In RL Algorithm, the agent should learn the best strategy from the reward of the trial-and-error interaction with the environment over time. The MDP process used for to simplify the modeling of RL. At the time of slot t agent should select the action of $a(t)$ from the action space A with the current state of $s(t)$, which is move to the new state of $s(t+1)$ after that immediately receive the reward of $r(t+1)$ from the environments. From the above steps are called as episode, which is recorded without experience of tuple such as $(s(t), a(t), r(t+1), s(t+1))$. In this state the system should be repeats the process of another episode with the next time of slot should keep in the maximum episode is reached.

For the traditional Q-learning the objective algorithm should have the optimal policy π should maximize the following cumulative counted reward at the time of slot t.

$$R(t) = \sum_{\lambda=0}^{\infty} \gamma^{\lambda} \gamma(t+1+\lambda), \gamma \epsilon (0,1)) \qquad (1)$$

From the above equation (1) describes the γ represents the discount factor, γ defined as the set as 0, which is mainly focused on the immediate reward based on greater Y, the rewards should spread over the time.

Action value of function $Q^{\pi}(s,a)$ should represented as the expected reward of to taking the action represents as a and the state as s under the policy as π. Which is can be expressed as

$$Q^{\pi}(s,a) = E^{\pi}[R(t)|s(t) = s, a(t) = a] \quad (2)$$

From the above equation (2) satisfices the Bellman equations.

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s' \varepsilon s} P^{a}_{s,s'} \, maxQ^{\pi*}(s', a') \quad (3)$$

From the above equation (3) describes the $R(s,a) = E[r(t+1)|s(t) = s, a(t) = a]$, which is also expected reward , s' and a' should indicate the state and acton of the next time slot and $P^{a}_{s,s'}$ which is contains the transition probability of performed the action a transferring to the state s' from state s. Q-learning process should learned from the action with the value function should converge the probability as 1 to the optimal as $Q^{\pi*}(s,a)$. The optimal Q function should express as,

$$Q^{\pi*}(s,a) = R(s,a) + \gamma \sum_{S' \epsilon S} P^{a}_{s,s'} \, maxQ^{\pi*}(s', a') \quad (4)$$

From the above equation (4) describes the Traditional Q-learning helps for to make the Q-functions due to the large state space followed by increase the size.

MADRL Algorithm

The proposed MADRL framework for the HVN shown in above fig. 2 in the proposed model, transmitted to V2V links, which is contains the agent and each of them have DQN. To trained this model have the multiple episodes. Each episode has the batch of experiences tuples Dare samples from the experiences replay with buffer fed into the neural network optimizations with the fixed periods. The loss function should be calculated as in equation (17)

$$Loss = \sum_{D} [r' + \gamma maxQ(s', a'; \psi_{target}) - Q(s,a;\psi_{eval}]^2 \quad (5)$$

From the above equation (5) describes r' defined as the reward of the next time slot and ψ_{target} should noted as the parameter of the target. The parameters of the target network are periodically replicated from evaluation network parameter ψeval and updated once after a fixed number of episodes. The ε-greedy strategy is introduced to select actions for exploring the unknown action space. It indicates that each agent has a probability of 1-ε to select action according to,

$$a^{*} = \arg maxQ(s,a) \quad (6)$$

From the above equation (6) describes the probability of ε should have the select action as randomly. Early stage of training which is better to explore the more actions for to reduced after period of training.

1. Algorithm-1 MARL-Based Resource Allocation

Initialization:

$AgentQNetwork - evaluation\ and\ target\ networks$

$QMixerNetwork - evaluation\ and\ target\ networks$

$ReplayBuffer\ (capacity = 50{,}000)$

$\varepsilon \leftarrow \varepsilon\_start = 1.0$

$for\ each\ episode\ do$

$Decay\ exploration\ rate: \varepsilon \leftarrow max(\varepsilon\_end, \varepsilon \times \varepsilon\_decay)$

$for\ each\ time\ step\ t\ do$

$for\ each\ BS\ agent\ k\ do$

$Get\ local\ observation\ S\_k(t)$:

$[Spectral\ Efficiency, Throughput, CQI, SNR, Cell\ Load\ per\ user]$

$Use\ \varepsilon - greedy\ to\ select\ resource\ level\ a\_k(t)$:

$resource\ level \in \{0,1,2,3,4\}$

$(higher\ level \rightarrow larger\ resource\ fraction\ allocated)$

$end\ for$

$Apply\ resource\ actions\ to\ environment$:

$- Scale\ user\ throughput\ by\ resource\ fraction$:

$frac = (resource\_level + 1) / n\_resource\_levels$

$new\_throughput = current\_throughput \times (0.5 + 0.5 \times frac)$

$- Update\ spectral\ efficiency\ accordingly$

$Compute\ reward\ r\_k(t+1)\ for\ each\ agent\ k$:

$r\_k = w1 \cdot r\_spectral\_efficiency + w2 \cdot r\_load\_balance + w3 \cdot r\_fairness\ (Jain's\ Index) + w4 \cdot r\_joint\_capacity + w5 \cdot r\_global\_load\_fairness$

$Store\ (S\_k(t), G(t), a\_k(t), r\_k(t+1),$

$S\_k(t+1), G(t+1), done)\ in\ ReplayBuffer$

$end\ for$

$for\ every\ train\_frequency\ steps\ do$

$Sample\ mini-batch\ (size = 256)\ from\ ReplayBuffer$

$for\ each\ agent\ k\ do$

$Compute\ Q\_k(S\_k, a\_k)\ via\ eval\ ResourceHead$

$Weight\ user\ Q-values\ by\ CQI-based\ importance$

$end\ for$

$Mix\ agent\ Q-values$:

$Q\_tot\ =\ QMixerNetwork(Q\_1, \ldots, Q\_K, G(t))$

$Compute\ target$:

$Q\_k\hat{}max\ =\ max\_a\ Q\_k(S\_k(t+1), a)\ via\ target\ network$

$y = r + \gamma \cdot QMixerNetwork(Q\_1\hat{}max, \ldots, Q\_K\hat{}max, G(t+1)) \cdot (1 - done)$

$Compute\ TD\ loss$:

$L\ =\ mean((y\ -\ Q\_tot)^2)$

$Update\ eval\ networks\ via\ Adam\ with\ gradient\ clipping$

$Soft\ update: \theta\_target \leftarrow \tau \cdot \theta\_eval\ +\ (1-\tau) \cdot \theta\_target$

$end\ for$

$Every\ target\_update\_frequency\ episodes$:

$Hard\ update\ target\ networks$

$end\ for$

2. Algorithm Explanation

The algorithm 1 starts with the initialization of the evaluation and target Q-networks, replay buffer, and exploration rate ε set to its maximum value. At each time step, each base station agent observes its local network state in terms of SNR, CQI, spectral efficiency, and cell load, followed by the selection of a resource level based on an ε-greedy policy. The resource levels are then used to scale the user throughput and spectral efficiency, after which each agent receives a reward based on load balance, fairness, and network capacity. These transitions are then stored in the replay buffer, and mini-batches are sampled periodically to calculate the TD loss, where the Q-values of individual agents are combined using the QMIX network to calculate a global joint Q-value. The evaluation networks are then updated using the Adam optimizer with gradient clipping, while the target networks are softly updated at each step and hard-updated periodically for stability during training. This continues until the policy converges to an optimal resource allocation strategy with a decaying exploration rate over episodes.

A. User Association and Resource Allocation in Heterogeneous Cellular Networks

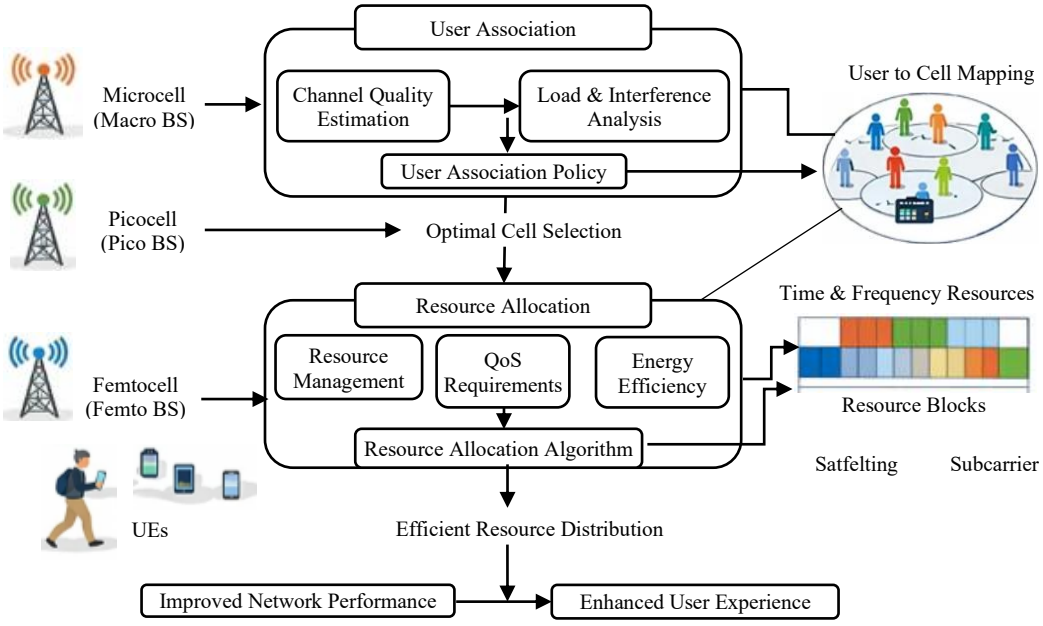

Fig. 3 Working Principles of User Association and Resource Allocation in Heterogeneous Cellular Networks

The fig. 3 shows the MARL-UA-RA model of the joint user association and resource allocation in wireless networks. It starts with User Association, the channel quality estimation and load/interference analysis is used in informing the user association policy, and the best cell is selected by the user. Resource Allocation, is the second level that takes into account the resource management available, QoS needs, and energy efficiency priorities in motivating the resource allocation algorithm, which guarantees fair resource allocation throughout the network. These steps, in combination, lead to the improved system performance due to the improved network throughput and energy-saving, as well as an improved user experience, and they indicate the synergy between optimized resource management and intelligent user assignment in a dynamic wireless environment.

1. User Allocation

The process of assigning a user equipment (UE) to the base station (Macrocell, Picocell, or Femtocell) to which the user equipment (UE) has to be attached is known as user association. Channel quality and load analysis are among the factors that led to this decision. The first one is channel quality estimation where the signal strength between the UE and various base stations is measured. The closer the relationship, the more the user will identify with that base station. Load and interference analysis is then used to assess the overall network conditions, including load and interference from neighboring cells. According to such assessments, the system uses a user association policy that assigns users to the base station that best serves them. This is important because it provides the user with a higher level of connectivity by choosing the base station with the best signal quality and the least interference.

2. Resource Allocation

After the user is associated with the base station, the next step is resource allocation, in which available resources (such as time and frequency blocks) are allocated to users to satisfy their communication requirements. The system has to trade off several things: the resource management will ensure that the available resources are distributed efficiently among the users; the Quality of Service (QoS) constraints will ensure that every user obtains the minimum resources necessary to have a satisfactory experience; the energy efficiency will make sure that the resources are consumed in an environmentally friendly and cost-effective way. An algorithm for resource allocation optimizes resource allocation to meet user requirements and network constraints. This process has ensured that every user receives an equal share of resources without compromising network performance.

3. Efficient Resource Distribution

The last part is efficient resource allocation, in which the system ensures that allocated resources are used in the best possible way. The system allocates resource blocks (time and frequency resources) based on user requirements and preferences, divided into satisfying and subcarrier usage. This implies that users with more bandwidth or greater performance needs receive priority in resource allocation, without compromising the overall network performance. The outcome of efficient resource allocation is better network performance, with resources utilized optimally, resulting in faster, more reliable networks. Also, the different components of this model are derived mathematically after fulfilling the expectations of the users concerning the connectivity and the integrity and efficiency of the whole network. The estimation of the channel quality, user association policy, and resource allocation policy are all parts of this model.

4. Channel Quality Estimation

This can be modeled as a function of the signal-to-noise ratio (SNR) for each UE and each base station.

$$SNR_k = \frac{P_{BS}}{N_0 . d_k^{\propto}} \quad (7)$$

From the above equation (7) describes the P_{BS} should represents the power transmitted by the base station and N_0 is the noise power, d_k is the distance among the UE and base station, $\propto$ is the path loss exponent.

5. User Association Policy

The user is associated with the base station BS_jthat provides the highest SNR in equation (8),

$$BS_j = argmax(SNR_{k,j}) \quad (8)$$

6. Resource Allocation

The resource allocation algorithm divides the time and frequency resources into blocks, ensuring that each user's required QoS is met. This can be done through optimization techniques, such as solving for the optimal allocation of resources $R_{i,j}$ for each UE I and the base station of j.

$$max \sum_i \sum_j R_{i,j} \; subject \; to \; \sum_i R_{i,j} \leq available \; resources \; at \; base \; station \; j \quad (9)$$

From the above equation (9) describes the $R_{i,j}$ should represent the resource allocated to the user I and the base station j, constraints should ensure that the allocation does not exceed the capacity of the base stations.

B. Algorithm -2 Multi-Agent Based User Association

$Initialization$:

$$AgentQNetwork \; - \; evaluation \; and \; target \; networks$$

$$QMixerNetwork \; - \; evaluation \; and \; target \; networks$$

$ReplayBuffer\ (capacity\ =\ 50{,}000)$

$\varepsilon \leftarrow \varepsilon\_start\ =\ 1.0$

$for\ each\ episode\ do$

$Decay\ exploration\ rate: \varepsilon \leftarrow max(\varepsilon\_end, \varepsilon\ \times\ \varepsilon\_decay)$

$for\ each\ time\ step\ t\ do$

$for\ each\ BS\ agent\ k\ do$

$Get\ local\ observation\ S\_k(t)$:

$[RSRP, RSRQ, Path\ Loss, Serving\ Cell\ Distance, Neighbor\ RSRP, Neighbor\ Load, Neighbor\ Distance\ for\ each\ user]$

$Use\ \varepsilon - greedy\ to\ select\ association\ action\ a\_k(t)$:

$association\ \in \{0 = stay, 1 = Neighbor1, 2 = Neighbor2, 3 = Neighbor3\}$

$(masked\ by\ neighbor\ availability)$

$end\ for$

$Apply\ association\ actions\ to\ environment$:

$-\ Execute\ handover\ only\ if$:

$neighbor\ RSRP \geq serving\ RSRP + 2.5\ dB\ (handover\ margin)$

$AND\ neighbor\ load \leq 0.85\ (max\ load\ threshold)$

$-\ Update\ serving\ cell, RSRP, distance, path\ loss$

$Compute\ reward\ r\_k(t+1)\ for\ each\ agent\ k$:

$r\_k\ =\ w1 \cdot r\_signal\ quality\ (RSRP\ gain) + w2 \cdot r\_load\ balance + w3 \cdot r\_handover\ penalty + w4 \cdot r\_global\ load\ fairness$

$Store\ (S\_k(t), G(t), a\_k(t), r\_k(t+1), S\_k(t+1), G(t+1), done)\ in\ ReplayBuffer$

$end\ for$

$for\ every\ train\_frequency\ steps\ do$

$Sample\ mini-batch\ (size = 256)\ from\ ReplayBuffer$

$for\ each\ agent\ k\ do$

$Compute\ Q\_k(S\_k, a\_k)\ via\ eval\ AssociationHead$

$Weight\ user\ Q-values\ by\ RSRP-based\ importance$

$Mask\ invalid\ neighbors\ with\ Q\ =\ -inf$

$end\ for$

$Mix\ agent\ Q-values$:

$Q\_tot\ =\ QMixerNetwork(Q\_1, \ldots, Q\_K, G(t))$

$Compute\ target$:

$Q\_k\hat{}max\ =\ max\_a\ Q\_k(S\_k(t+1), a)\ via\ target\ network$

$y = r + \gamma \cdot QMixerNetwork(Q\_1\hat{}max, \ldots, Q\_K\hat{}max, G(t+1)) \cdot (1 - done)$

$Compute\ TD\ loss$:

$L\ =\ mean((y\ -\ Q\_tot)^2)$

$Update\ eval\ networks\ via\ Adam\ with\ gradient\ clipping$

$Soft\ update: \theta\_target \leftarrow \tau \cdot \theta\_eval\ +\ (1-\tau) \cdot \theta\_target$

$end\ for$

$Every\ target\_update\_frequency\ episodes$:

$Hard\ update\ target\ networks$

$end\ for$

1. Algorithm Explanation

The algorithm 2 initializes the evaluation and target Q-networks, replay buffer, and exploration rate ε to its maximum value. At each time step, each base station agent observes user-level signal features such as RSRP, RSRQ, path loss, and neighbor cell information, and chooses an association action based on an ε-greedy policy. A handover is performed only if the target neighbor provides a sufficient RSRP gain of at least 2.5 dB and is below the maximum load threshold, satisfying both signal and load conditions. The obtained reward reflects signal improvement, load balance among cells, handover cost, and global fairness, which are stored in the replay buffer as transitions. Mini-batches are periodically sampled to calculate the TD loss, where the individual agent Q-values from the association head are

combined via the QMIX network to obtain a cooperative global Q-value. The networks are updated using the Adam optimizer with soft target updates per step, enabling the policy to dynamically learn the best user-to-cell association actions over episodes.

Dataflow Diagram for Proposed Methodology

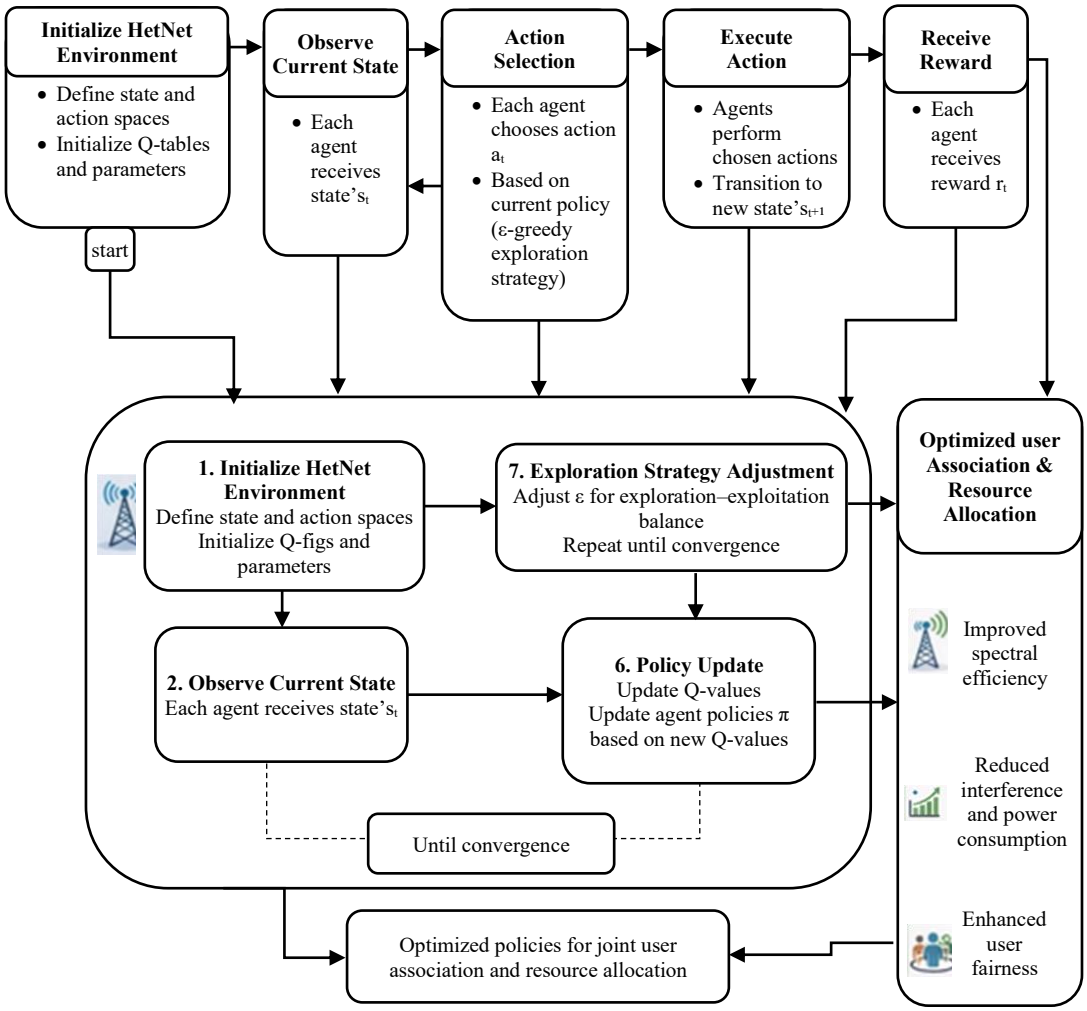

Fig. 4 Dataflow Diagram for Proposed Methodology

Fig. 4 presents an algorithmic framework for Multi-Agent Reinforcement Learning (MARL) used for joint user association and resource allocation in a heterogeneous cellular network. It starts with setting up the HetNet environment, during which the state and action spaces are created and the agents initialize their Q-tables and parameters. Each agent then monitors the current system condition and chooses an action according to an ε-greedy policy. A particular action is performed, as a result of which the agent switches to a new state. The reward is given to the agent based on the outcome of the action. To reconcile between exploration and exploitation in the learning process, the algorithm optimizes the exploration factor (ε). This enables the agents to experiment with new actions and, in the process, correct their policies. The agents' policies are revised after every action and reward cycle using new Q-values based on the rewards obtained, which enhance decision-making in the long run. This is done iteratively until convergence is achieved. In the end, the algorithm delivers user association and resource allocation policies that are optimized to achieve spectral efficiency, minimize interference and power consumption, and increase user fairness. The framework ensures network resources are distributed optimally to maximize performance and user experience in dynamic, heterogeneous environments.

Proposed Algorithm-MARL-UA-RA

Initialization:

$$Evaluation\ network\ parameters\ \theta, target\ network\ parameters\ \theta^{-}$$

$$Replay\ buffer\ D\ with\ capacity\ C\ =\ 50{,}000$$

$$\varepsilon \leftarrow 1.0, \gamma\ =\ 0.95, \tau\ =\ 0.008$$

$$for\ each\ episode\ do$$

$$Sample\ N\ =\ 250\ users\ across\ K\ \geq\ 22\ active\ cells$$

$\varepsilon \leftarrow max(\varepsilon\_min, \varepsilon \times \varepsilon\_decay)$

$for\ each\ time\ step\ t\ do$

$for\ each\ agent\ k\ do$

$Observe\ local\ state$:

$s\_k(t) = [RSRP, RSRQ, SNR, CQI, \eta, T, L\_p, P\_tx, v, d, RSRP\_n, L\_n, d\_n]$

$Observe\ global\ state$:

$g(t) = [\mu\_L, \sigma\_L, \mu\_R, \sigma\_R, \mu\_T, H, \rho\_peak]$

$Select\ joint\ action\ via\ \varepsilon - greedy$:

$a\_k\text{^}assoc(t) \in \{0, 1, 2, 3\}$

$a\_k\text{^}res(t) \in \{0, 1, 2, 3, 4\}$

$end\ for$

$Apply\ association\ action$:

$Handover\ valid\ if$:

$RSRP\_neighbor \geq RSRP\_serving + \Delta\_HO\ (\Delta\_HO = 2.5\ dB)$

$AND\ L\_neighbor \leq L\_max\ (L\_max = 0.85)$

$Update: RSRP, d, L\_p = 46.3 + 33.9 \cdot log_{10}(f) + (44.9 - 6.55 \cdot log_{10}(h\_b)) \cdot log_{10}(d)$

$Apply\ resource\ action$:

$\varphi\_k = (a\_k\text{^}res + 1) / N\_res$

$T\_new = T\_cur \times (0.5 + 0.5 \times \varphi\_k)$

$\eta\_new = \eta\_cur \times (0.5 + 0.5 \times \varphi\_k)$

$Compute\ reward$:

$r\_k(t) = w_1 \cdot tanh(\bar{\eta})$ $[spectral\ efficiency]$

$+ w_2 \cdot tanh(RSRP)$ $[signal\ quality]$

$- w_3 \cdot tanh(|L - L*) \cdot 2)$ $[load\ balance, L* = 0.5]$

$- w_4 \cdot tanh(H\_k \cdot 3)$ $[handover\ penalty]$

$+ w_5 \cdot (J - 0.5)$ $[Jain's\ fairness\ index]$

$+ w_6 \cdot tanh(C\_eff / 5)$ $[joint\ SINR - resource\ capacity]$

$- w_7 \cdot tanh(\sigma\_L \cdot 2)$ $[global\ load\ fairness]$

$where\ C\_eff = log_2(1 + SINR) \times \varphi\_k$

$J = (\Sigma_i T_i)^2 / (n \cdot \Sigma_i T_i^2)$

$Store\ transition\ (s\_k(t), g(t), a\_k(t), r\_k(t), s\_k(t+1), g(t+1))\ in\ D$

$end\ for$

$for\ every\ 2\ time\ steps\ do$

$Sample\ mini - batch\ B\ from\ D$

$Compute\ agent\ Q - values$:

$Q\_k(s\_k, a\_k; \theta) = (Q\_k\text{^}assoc + Q\_k\text{^}res) / 2$

$weighted\ by\ importance: \omega\_i = softmax(RSRP\_i)$

$Compute\ joint\ Q - value$:

$Q\_tot = f\_mix(Q\_1, \ldots, Q\_K, g(t); \theta\_mix)$

$where\ f\_mix\ is\ monotonic: \partial Q\_tot / \partial Q\_k \geq 0$

$Compute\ target$:

$Q\_k\text{^}max = max\_a\ Q\_k(s\_k(t+1), a; \theta^-)$

$y = r + \gamma \cdot f\_mix(Q\_1\text{^}max, \ldots, Q\_K\text{^}max, g(t+1); \theta\_mix^-) \cdot (1 - done)$

$y \leftarrow tanh(y / 5) \times 5$ $[target\ stabilization]$

$Compute\ loss\ and\ update$:

$\delta\_k = y - Q\_tot$ $[TD\ error, clipped\ to\ \pm 5]$

$L(\theta) = 1/|B| \cdot \Sigma\ \delta\_k^2$

$\theta \leftarrow \theta - \alpha \cdot \nabla\_\theta\ L(\theta)$ $[Adam, \alpha = 0.0005]$

$Soft\ update\ target$:

$\theta^- \leftarrow \tau \cdot \theta + (1 - \tau) \cdot \theta^-$

$end\ for$

$Every\ 20\ episodes$:

$\theta^- \leftarrow \theta$ $[hard\ update]$

$Evaluate\ policy\ on\ validation\ set$

$Save\ best\ model;\ early\ stop\ after\ 20\ non - improving\ evaluations$

$end\ for$

1. Algorithm Explanation

The proposed algorithm initializes dual-headed Q-networks for association and resource allocation, a QMIX mixer network, a replay buffer, and exploration parameters before the start of training. At each time step, each base station agent observes a complex local state that includes user signal metrics, neighbor cell details, and temporal information, together with a global state that represents network-wide load and throughput information, and selects an association action and resource level jointly using an ε-greedy policy. The association action initiates a handover only if the target neighbor meets both the RSRP gain margin of 2.5 dB and the maximum load constraint of 0.85, while the resource action adjusts each user's throughput and spectral efficiency proportionally using a resource fraction φ. A compound reward is then calculated for each agent based on spectral efficiency, signal quality, load balance, handover cost, per-cell fairness using Jain's index, joint SINR-based capacity, and global load fairness, with all transitions being stored in the replay buffer. In the training phase, the Q-values of the agents calculated from both the association and resource heads are weighted by the user RSRP importance, normalized over the active agents, and mixed using the monotonic QMIX network to obtain the global joint Q-value, which is optimized by minimizing the TD loss using the Adam optimizer with adaptive gradient clipping and soft target network updates. The cooperative multi-agent training procedure is repeated over episodes with a decaying exploration rate, periodic hard target updates, and early stopping based on validation performance, finally converging to an optimal joint policy for the joint user association and resource allocation problem in the heterogeneous network.

IV. RESULTS AND DISCUSSION

Dataset Description

The Speed Dataset and Bandwidth Dataset for 4G LTE are important for analyzing and optimizing 4G LTE network performance (Raca et al., 2018). The 4G LTE Speed Dataset consists of such attributes as: timestamp, location, cell ID, download/upload speed (Mbps), signal strength (RSSI or SNR), latency, type of network, carrier aggregation (present or absent), and type of user equipment (UE), which are used to assess the performance of speed in practice and simulate the factors of geographical location, signal quality, and user load. The Bandwidth Dataset, in turn, is dedicated to the preferential distribution and use of bandwidth, the attributes of which include the time and date, cell ID, bandwidth allocated (MHz), bandwidth consumed (Mbps), modulation scheme (e.g., the QPSK, 16-QAM, 64) and the transmission power, SNR, user load, and the scheduling algorithm (e.g., the round-robin scheduling algorithm and the priority-based scheduling algorithm). Combined, these datasets can offer information on the network capacity, equitable allocation of bandwidth, and efficacy of LTE network functionality and support improved quality of service (QoS) and resource control.

Simulation Setup

This is simulated using the MARL-UA-RA algorithm on a two-tier small cell network, where one MBS covers 100 antennas, ten SBSs, and 50 UEs, and the MBS is randomly located at 500 m. UEs and SBSs are randomly distributed within the coverage area. It is segmented into clusters and SBS covers the areas. The channels are believed to be slow Rayleigh fading channels; hence no variation is experienced when transferring a block of data. The highest possible transmission power of MBS and SBS is assumed as dBm and dBm respectively. MBS=34+46 log d and SBS=37+30 log d is the path loss factor of the MBS and SBS respectively. In the distance, where would UE and BS. These parameters are only a few but there are other simulation parameters, which are specified in simulation parameter table I

The table I explains local state space encodes detailed per-user signal and mobility information for 30 users per cell with 321 dimensions, while the 24-dimensional global state encodes network-wide load and throughput information for cooperative decision-making. Each agent chooses a joint action consisting of an association choice among 4 options and a resource level among 5 levels per user, yielding a very large but structured action space that is efficiently represented and computed by the dual-headed Q-network. The learning rate of 0.0005 employs cosine annealing to guarantee convergent learning, and the discount factor of 0.95 trades off between short-term and long-term rewards for network performance. Exploration is implemented using ε-greedy with decay from 1.0 to 0.08 over episodes, and training is conducted using mini-batches of 256 uniformly sampled transitions from a replay buffer of 50,000 experiences.

Parameter Initialization

TABLE I PARAMETER INITIALIZATION

Parameter	Values
State Space	Local: [RSRP, RSRQ, SNR, CQI, η, T, L_p, P_tx, v, d] × 30 users + Neighbor features (RSRP, Load, Distance × 3) + Cell features (Load, Load Imbalance, Temporal encodings) = 321 dimensions; Global: g(t) = [μ_L, σ_L, L_max, L_min, Var_L, μ_R, σ_R, μ_T, σ_T, K_active/K, H_rate, ρ_peak] = 24 dimensions
Action Space	Discreate Joint: Association a^assoc ∈ {0,1,2,3} × Resource Level a^res ∈ {0,1,2,3,4} per user; Total = 9 actions per user × 30 users per agent
Learning Rate (α)	0.0005 with Cosine Annealing scheduler (η_min = 8×10^{-5})
Discount Factor (γ)	0.9-0.99
Exploration Rate (ε)	ε_start = 1.0, ε_end = 0.08, ε_decay = 0.955 (per episode)
Batch Size	256
Repay Buffer Size	50,000-100,000
Target Network Update	Soft update: τ = 0.008 per step; Hard update: every 20 episodes
Gradient Clipping	Adaptive clipping based on 95th percentile of recent gradient history; maximum cap = 2.0
Number of Episodes	100
Hidden Dimension	128 (agent network); Mixer embedding = 32
Max Steps per Episode	50 time steps
Users per Episode	250 users across ≥ 25 active cells

Target network stability is guaranteed by soft updates with τ = 0.008 at each step and hard updates every 20 episodes, with adaptive gradient clipping bounded by 2.0 to prevent training instability over the 100 training episodes with early stopping after 20 episodes without improvement.

Metric Evaluation

The performance of the proposed framework was assessed using the following network-level metrics of efficiency, load balancing, fairness, mobility stability, and convergence of learning, as described by equations (10) - (20). These metrics provide a comprehensive measure of the system's throughput, resource utilization, and learning performance.

1. Network Performance Metrics

Equations (10) and (11) combine to assess the overall network performance by averaging the user-level throughput and spectral efficiency over all N users. Although $\bar{T}$ assesses the effective data rate supported by the network, $\bar{\eta}$ assesses the efficiency of bandwidth use. Both equations assess the network capacity and bandwidth use based on the traffic and resource allocation.

- Average Throughput

$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i \qquad (10)$$

- Average Spectral Efficiency

$$\bar{\eta} = \frac{1}{N}\sum_{i=1}^{N} \eta_i \quad (11)$$

2. Load Balance Metric

$$\Delta L = \frac{1}{K}\sum_{k=1}^{K} |L_k - \bar{L}| \quad (12)$$

Equation (12) expresses the amount of variation of individual cell loads L_k from the average network load $\bar{L}$ over K cells. The equation shows the level of balance of traffic in the network. The smaller the value of ΔL, the better the load balancing, the smaller the congestion hotspots, and the more efficient the resource utilization.

3. Fairness Metric

$$J = \frac{(\sum_{i=1}^{n} T_i)^2}{n\sum_{i=1}^{n} T_i^2} \quad (13)$$

Equation (13) represents Jain's Fairness Index, which evaluates how evenly throughput is distributed among $n$users. The metric ranges from $\frac{1}{n}$to 1, where a value close to 1 indicates equal resource allocation among users, and lower values reflect performance imbalance or user starvation within the network.

4. Handover Rate

$$H_{\text{rate}} = \frac{\sum_{i=1}^{N} 1[H_i = 1]}{N} \quad (14)$$

Equation (14) defines the handover rate as the proportion of users experiencing a cell transition among the total Nusers. The indicator function $\mathbf{1}[H_i = 1]$equals 1 if a handover occurs for user i, and 0 otherwise. This metric reflects mobility management stability, where excessively high

values may indicate frequent cell switching and signaling overhead.

5. Energy Efficiency

$$EE = \frac{\bar{T}}{\overline{P_{tx}} + P_0} \quad (15)$$

Equation (15) defines energy efficiency as the ratio of average network throughput $\bar{T}$to the total power consumption, which includes the average transmit power $\bar{P}_{tx}$and fixed circuit power P_0. This metric measures how effectively the network converts energy into useful data transmission, expressed in bits per joule (bps/W). Higher values indicate more sustainable and power-efficient network operation.

6. Training Convergence Metrics

- *TD loss*

$$L(\theta) = \frac{1}{|B|}\sum_{b\in B}\delta_b^2 \ (16)$$

- *TD error*

$$\delta = \text{clip}(y - Q_{tot}, -5,5) \ (17)$$

- *Evaluation Reward*

$$\overline{R_{eval}} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} r_k^m \ (18)$$

The equations (16) - (18) assess the learning stability and convergence properties of the reinforcement learning model. The TD error in equation (16) calculates the mean squared temporal difference error for a mini-batch B, which represents the network's accuracy in approximating the target Q-values. The clipped TD error in equation (17) ensures learning stability by preventing large updates and avoiding gradient explosion. The evaluation reward in equation (18) represents the cumulative performance of the trained policy over several episodes, which directly indicates the quality of the policy

7. Cell Utilization

- *Cell load*

$$\overline{L_k} = \frac{N_k}{N_{max}} (19)$$

- *Average Users Per cell*

$$\bar{U} = \frac{1}{K}\sum_{k=1}^{K} N_k \ (20)$$

The cell-level resource consumption and traffic distribution are expressed by equations (19) and (20). The cell load $\bar{L}_k$ is defined as the ratio of active users N_k to the maximum capacity N_{max} in cell K. The average users per cell $\bar{U}$ give information about the traffic density in K cells.

Results

1. Performance Analysis of QMIX-Based MARL Framework: Reward Convergence, Spectral Efficiency, Fairness, and Energy Efficiency across Training Episodes

TABLE II QUANTITATIVE PERFORMANCE SUMMARY OF THE PROPOSED MARL-BASED JOINT USER ASSOCIATION AND RESOURCE ALLOCATION FRAMEWORK

Metric	Range	HETNET-MARL-UA-RA
Reward	Env-dependent	750 - 950
Fairness	0 - 1	0.995 – 0.999
Loss	Env-dependent	0 – 0.1071
Handover Rate	0 - 1	0.125 – 0.060
Throughput	Env dependent	0.009 – 0.2
Load Imbalance	0 – 1	0.133 – 0.2

In table II the HETNET-MARL-UA-RA model has shown excellent performance with high fairness values (0.995-0.999) and low loss values (0-0.1071), which indicates that the model has high accuracy. The reward value varies between 750 and 950, depending on the environment. Although the handover rate (0.125-0.060) and load imbalance (0.133-0.2) are low, which indicates that the model is performing well, the throughput (0.009-0.2) varies.

Fig. 5 below shows the six key training values of the proposed QMIX-based MARL algorithm over 100 episodes. Subplot 1 plots the total reward per episode, which varies between 500 and 1050 in the initial episodes but settles down between 800 and 850 from episode 40 onwards, with the smoothed plot validating the consistent improvement in the policy and reduction in variance towards convergence. Subplot 2 plots the training loss, which increases rapidly from a negligible value in the first 20 episodes as the replay buffer collects enough experience, and then plateaus at a value of 0.10 from episode 60 onwards, thereby validating that the Q-networks have entered a stable learning phase. Subplot 3 plots the exploration rate ε, which decreases smoothly from 1.0 to its final value of 0.08 around episode 50, thereby validating the transition from random to deterministic exploration of the policy. Subplot 4 plots the evaluation reward calculated on the validation set every 5 episodes, which varies between 860 and 965, with a clear peak at episode 11 followed by gradual re-stabilization, thereby validating that the policy generalizes well to the unseen network conditions despite the fluctuations. From Subplot 5, the rate of handover is observed to be reducing steadily from around 0.10 in the initial episodes to around 0.075 in episode 100, thus confirming that the association policy is progressively learning to reduce unnecessary handovers while ensuring signal quality. From Subplot 6, there is a steady increase in the average throughput from 0.008 Mbps in the initial episodes to around 0.021 Mbps in episode 100, thus directly validating the efficiency of the combined user association and resource allocation strategy learned by the framework.

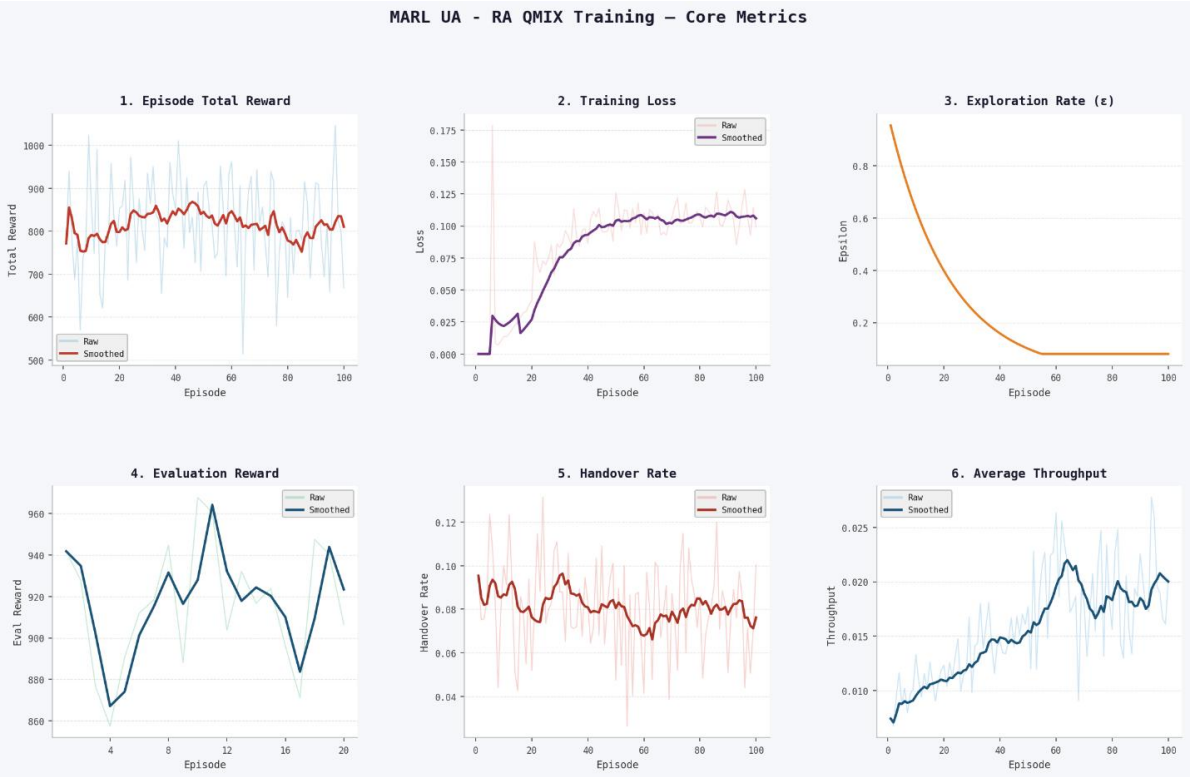

Fig. 5 MARL – UA– RA QMIX Training Performance – Core Metrics

Fig. 6 MARL-UA-RA-QMIX Training – Network Quality Metrics

Fig. 6 illustrates five network quality metrics assessed after 100 training episodes of the proposed QMIX-based MARL framework. Subplot 7 indicates that the load imbalance converges to a stable value of 0.50, which implies that the policy successfully balances users across cells. Subplot 8 indicates that the spectral efficiency increases from 0.008 to 0.020 bits/s/Hz, which confirms that the joint resource allocation policy improves the spectrum efficiency of the network. Subplot 9 indicates that the Jain fairness index consistently stays above 0.998, which validates that the proposed framework maintains near-optimal fairness in the distribution of throughput to all users. Subplot 10 indicates

that the energy efficiency increases from 0.00075 to 0.00200 bits/J and peaks around episode 65 before converging to a stable value, which reflects the framework's capability to provide higher throughput with reduced transmission power. Subplot 11 indicates that the average RSRP converges to 0.20 dBm after the initial variance in the early episodes of training, which suggests that the proposed framework successfully associates users with cells having better signal quality.

2. *Comparative Performance Analysis of MARL-UA-RA-QMIX Against Baseline and Heuristic Methods*

Table III shows a comparison of the proposed MARL-UA-RA-QMIX framework with six baseline approaches in terms of two important performance metrics: Average System Capacity (ASC) and Average Network Utility (ANU). Among the deep reinforcement learning-based baselines, D3QN (GS), D3QN (SS), and DQN all have the same ASC of 52.51 and ANU of 23.53, which implies that single-agent strategies with and without message passing have similar and suboptimal performance in dynamic HetNet settings. Q-Learning has slightly better ASC of 52.83 and ANU of 23.68, while the Genetic Algorithm (GA) has significantly poorer performance of 38.11 and 19.38, which is expected due to the limitations of evolutionary algorithms in real-time adaptive settings.

TABLE III PERFORMANCE COMPARISON OF PROPOSED MARL-UA-RA-QMIX AGAINST PREVIOUS MODELS IN TERMS OF AVERAGE SYSTEM CAPACITY (ASC) AND AVERAGE NETWORK UTILITY (ANU)

Method	ASC	ANU
D3QN (GS)	52.51	23.53
D3QN(SS)	52.51	23.53
DQN	52.51	23.53
Q – learning	52.83	23.68
GA	38.11	19.38
MRSP (Zhao et al., 2019)	68.15	31.71
MARL- UA – RA -QMIX	75.70	34.39

The MRSP baseline proposed by Zhao et al. has significantly better ASC of 68.15 and ANU of 31.71, which serves as a strong benchmark through its specialized resource and association scheme. The proposed MARL-UA-RA-QMIX framework has the best ASC of 75.70 and ANU of 34.39, which are improvements of about 11% and 8.5% over the MRSP baseline, respectively, and thus confirm the efficacy of cooperative multi-agent joint optimization.

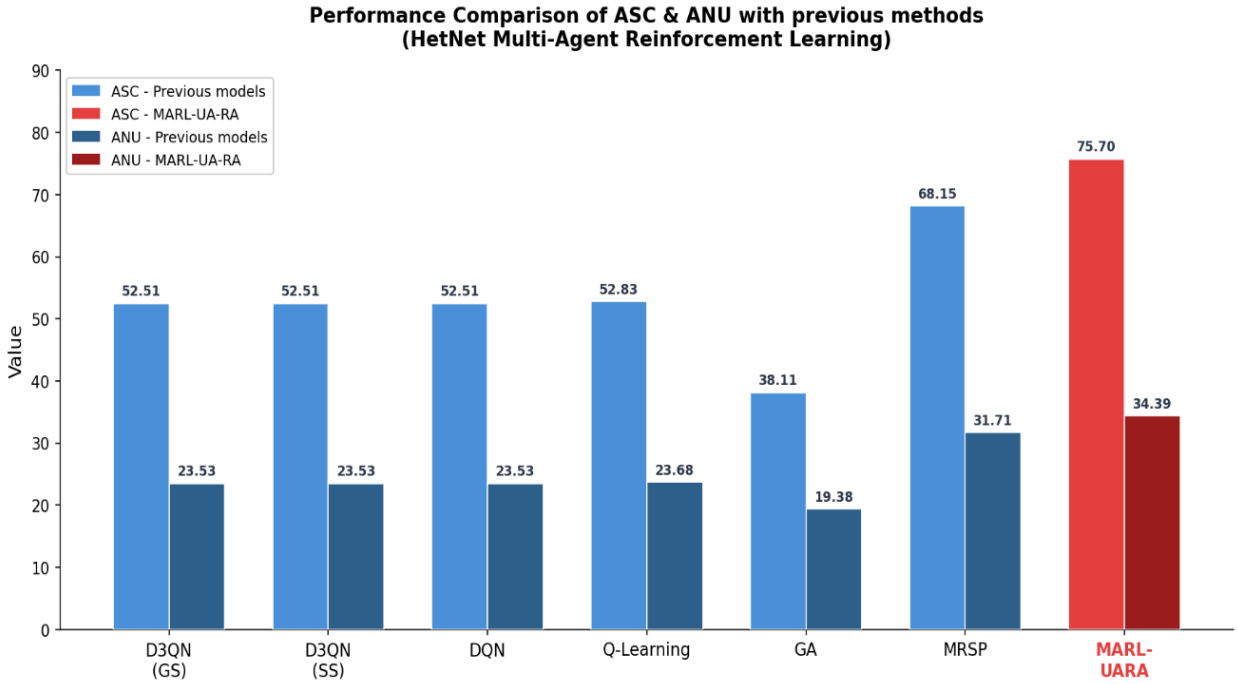

Fig. 7 Bar Chart Comparison of ASC and ANU with Previous Method and Proposed Methods in HetNet Multi-Agent Reinforcement Learning (QMIX)

Fig. 7 again emphasizes the results through a grouped bar chart that compares ASC and ANU for all seven methods. The light blue bars indicating the baseline ASC values are relatively flat at 52 for the D3QN variants, DQN, and Q-Learning algorithms, thus proving that the traditional single-agent RL algorithms reach a similar level of performance regardless of architectural differences. The GA method indicates a considerable decrease in both ASC and ANU, thus proving its inappropriateness for dynamic network optimization. MRSP indicates a considerable increase in ASC to 68.15, thus proving it to be the best among the baselines. However, the proposed MARL-UA-RA-QMIX algorithm clearly outperforms all others with the highest bars in both ASC (75.70) and ANU (34.39), thus proving the superiority of multi-agent cooperative learning in heterogeneous cellular network optimization.

TABLE IV PERFORMANCE COMPARISON OF PROPOSED MARL-UA-RA AGAINST HEURISTIC BASELINE METHODS IN TERMS OF ANU, ASC, FAIRNESS INDEX, AND LOAD IMBALANCE

Method	ANU	ASC	Fairness	Load Imbalance
Random	25.67	45.78	0.8838	0.6175
Greedy	26.47	47.27	0.94	0.6949
Round Robin	25.67	45.83	0.88	0.6202
MARL-UA-RA	34.39	75.70	0.99	0.5132

From table IV and fig. 8, a complete comparison of the proposed MARL-UA-RA framework with three traditional heuristic approaches: Random, Greedy, and Round Robin, is shown for four different criteria: ANU, ASC, Fairness Index, and Load Imbalance. The Random and Round Robin approaches provide almost equal ANU values of 25.67 and ASC values of around 45.78 and 45.83, respectively, thus validating the fact that non-adaptive scheduling algorithms provide negligible differences in performance, as can be seen from the close clustering of dots in fig. 8. The Greedy approach provides slightly better ANU of 26.47 and ASC of 47.27 with a fairness index of 0.94, but at the expense of maximum load imbalance of 0.6949, thus validating the fact that greedy cell association leads to cell overload, as can be seen from the high orange dot in fig. 7

Conversely, the proposed MARL-UA-RA framework, indicated by diamond-shaped markers in the fig. 8, performs best on all four aspects at the same time, with the maximum ANU of 34.39 and ASC of 75.70, indicating an improvement of about 34% and 60% over the best heuristic method, along with the highest fairness index of 0.99 and the lowest load imbalance of 0.5132, thereby comprehensively proving that the cooperative multi-agent learning strategy not only maximizes the network capacity and utility but also guarantees fair resource allocation and balanced cell loading in the heterogeneous network.

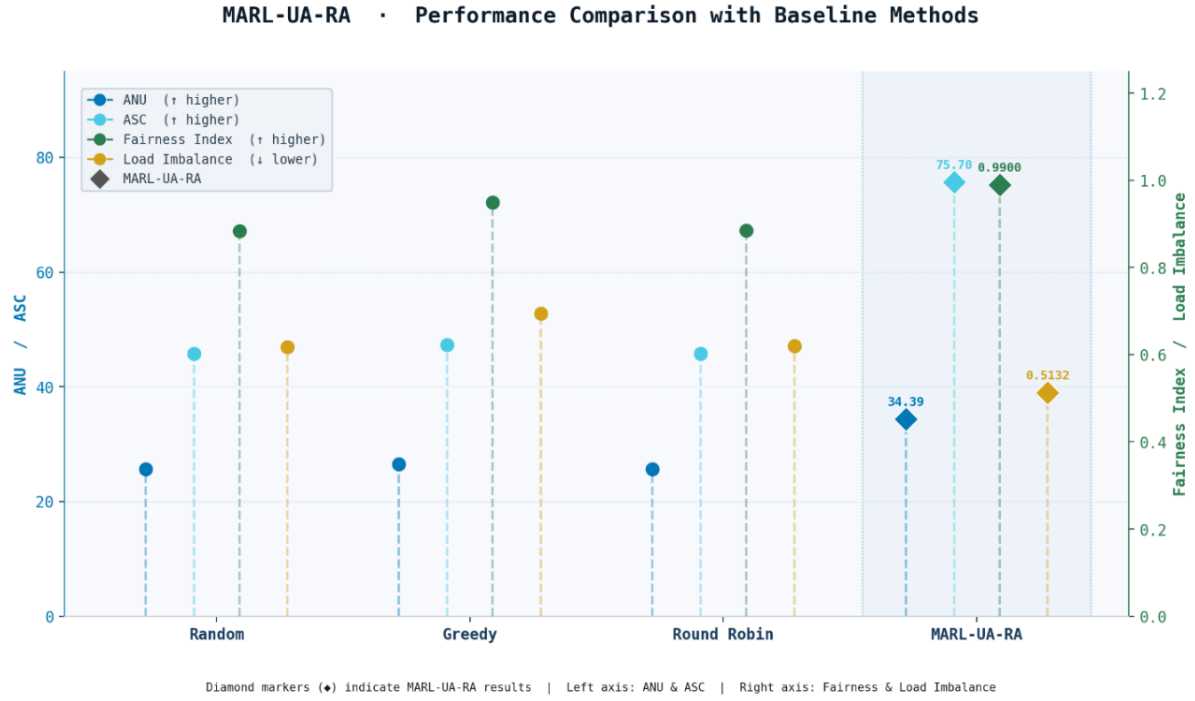

Fig. 8 Multi-Metric Performance Comparison of MARL-UA-RA Against Random, Greedy, and Round Robin Baseline Methods

Discussion

The experimental outcomes show the efficacy of the proposed MARL-UA-RA-QMIX framework in addressing the joint user association and resource allocation problem in Heterogeneous Cellular Networks. The analysis of the training process for 100 episodes shows the convergence of the episode reward to 800-850, with the throughput increasing from 0.008 to 0.021 Mbps and spectral efficiency increasing from 0.008 to 0.020 bits/s/Hz, thus supporting the refinement of the joint policy. The reduction in handover rate from 0.10 to 0.075 supports the conclusion that the framework is capable of learning to reduce unnecessary handovers while preserving signal quality. The close-to-optimal value of Jain's fairness index of 0.99 and the improvement in energy efficiency during training support the conclusion that the framework is capable of addressing multiple network objectives. The comparison of the proposed framework with deep reinforcement learning-based frameworks such as D3QN, DQN, and Q-Learning, and the best-performing MRSP benchmark, shows that MARL-UA-RA-QMIX outperforms the competition with an Average System Capacity of 75.70 and Average Network Utility of 34.39. Moreover, the comparison of the proposed approach with heuristic algorithms such as Random, Greedy, and Round Robin reveals that the cooperative multi-agent learning approach performs better than the non-adaptive approaches in all four aspects. Overall, the above findings prove that the proposed QMIX-based framework is an

efficient and fair solution for dynamic heterogeneous network optimization.

V. CONCLUSION

This work has proposed a QMIX-based Multi-Agent Reinforcement Learning framework for Joint User Association and Resource Allocation (MARL-UA-RA) in Heterogeneous Cellular Networks (HetNets), where each base station is modeled as a self-contained cooperative agent that makes fully decentralized decisions while cooperatively optimizing global network performance. The proposed MARL-UA-RA-QMIX framework utilizes a dual-headed Q-network architecture with separate association and resource allocation heads, coupled with a monotonic QMIX mixing network that aggregates individual Q-values from each agent into a global joint Q-value, facilitating cooperative optimization for all simultaneously active base stations. The experimental analysis has clearly shown that the proposed MARL-UA-RA-QMIX framework substantially outperforms all baseline approaches including D3QN (GS), D3QN (SS), DQN, Q-Learning, Genetic Algorithm, and MRSP, achieving the best Average System Capacity of 75.70 and Average Network Utility of 34.39, which correspond to improvements of about 11% and 8.5% over the best baseline MRSP, respectively. Moreover, the proposed framework has also achieved a near-optimal Jain's fairness index of 0.99 and minimum load imbalance of 0.5132 compared to all baseline heuristic approaches, thus validating its capability to simultaneously optimize network capacity, spectral efficiency, and energy efficiency while ensuring fair resource allocation among users.

Analysis of training data further supports the convergence of reward, throughput, and spectral resources for 100 episodes with a progressively reducing handover rate, thus establishing the robustness and applicability of the obtained joint policy. Although encouraging, the findings are accompanied by several issues associated with scalability in ultra-dense networks, stability of convergence in highly dynamic mobility patterns, and multi-objective optimization involving energy harvesting and interference management.

REFERENCES

[1] Alablani, I., & Alenazi, M. J. (2023). DQN-GNN-based user association approach for wireless networks. *Mathematics*, *11*(20), 4286. https://doi.org/10.3390/math11204286

[2] Alhazmi, A. S., & Arafah, M. A. (2025). Neural Network–Based Adaptive Resource Allocation for 5G Heterogeneous Ultra-Dense Networks. *Sensors*, *25*(24), 7521. https://doi.org/10.3390/s25247521

[3] Cheng, Z., Chen, N., Liu, B., Gao, Z., Huang, L., Du, X., & Guizani, M. (2020). Joint user association and resource allocation in HetNets based on user mobility prediction. *Computer Networks*, *177*, 107312. https://doi.org/10.1016/j.comnet.2020.107312

[4] 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

[5] Hugh, Q., & Soria, F. (2025). VoltSecure: A Secure Federated Learning Model for Decentralized Energy Management Systems. *International Academic Journal of Innovative Research*, *12*(3), 33-42. https://doi.org/10.71086/IAJIR/V12I3/IAJIR1223

[6] Kim, H., & So, J. (2025). Distributed Multi-Agent Deep Reinforcement Learning-Based Transmit Power Control in Cellular Networks. *Sensors*, *25*(13), 4017. https://doi.org/10.3390/s25134017

[7] Lee, I., & Kim, D. K. (2024). Decentralized multi-agent DQN-based resource allocation for heterogeneous traffic in V2X communications. *IEEE Access*, *12*, 3070-3084. https://doi.org/10.1109/ACCESS.2023.3349350

[8] Liang, R., Lyu, H., & Fan, J. (2023). A deep reinforcement learning-based power control scheme for the 5g wireless systems. *China Communications*, *20*(10), 109-119. https://doi.org/10.23919/JCC.ea.2021-0523.202302

[9] Liu, Q., & Ma, Y. (2025). Communication resource allocation method in vehicular networks based on federated multi-agent deep reinforcement learning. *Scientific Reports*, *15*(1), 30866. https://doi.org/10.1038/s41598-025-15982-x

[10] Luo, Y., Wang, Y., Lei, Y., Wang, C., Zhang, D., & Ding, W. (2023). Decentralized user allocation and dynamic service for multi-UAV-enabled MEC system. *IEEE Transactions on Vehicular Technology*, *73*(1), 1306-1321. https://doi.org/10.23919/JCC.ea.2021-0523.202302

[11] Ma, J., Gao, H., Guo, L., & Li, H. (2024). Energy-efficient joint resource allocation for heterogeneous cellular networks with wireless backhauls. *AEU-International Journal of Electronics and Communications*, *176*, 155170. https://doi.org/10.1016/j.aeue.2024.155170

[12] Mayilsamy, J., & Rangasamy, D. P. (2021). Enhancement of Energy Efficient Routing Scheduling Algorithm based on SDN Using IoT. *International Academic Journal of Science and Engineering*, *8*(1), 10-18.

[13] Mishra, D., Traversi, E., Trotta, A., Raut, P., Galkin, B., Di Felice, M., & Natalizio, E. (2025). Network slicing in aerial base station (UAV-BS) towards coexistence of heterogeneous 5G services. *Computer Networks*, *261*, 111146. https://doi.org/10.1016/j.comnet.2025.111146

[14] Nasir, Y. S., & Guo, D. (2019). Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. *IEEE Journal on selected areas in communications*, *37*(10), 2239-2250. https://doi.org/10.1109/JSAC.2019.2933973

[15] Niasi, K. S. K., Shanthi, R., Hossain, R. R., & Rama Sree, S. (2025). *Reinforcement learning-based power control in dense WLAN environments*. *Journal of Internet Services and Information Security, 15*(3), 362–376. https://doi.org/10.58346/JISIS.2025.I3.025

[16] Pindi, N. R., & Velez, F. J. (2025). Traffic Scheduling and Resource Allocation for Heterogeneous Services in 5G New Radio Networks: A Scoping Review. *Smart Cities*, *8*(5), 168. https://doi.org/10.3390/smartcities8050168

[17] Raca, D., Quinlan, J. J., Zahran, A. H., & Sreenan, C. J. (2018, June). Beyond throughput: A 4G LTE dataset with channel and context metrics. In *Proceedings of the 9th ACM multimedia systems conference* (pp. 460-465). https://doi.org/10.1145/3204949.3208123

[18] Ramesh, P., Bhuvaneswari, P. T. V., Dhanushree, V. S., Gokul, G., & Sahana, S. (2025). User association-based load balancing using reinforcement learning in 5G heterogeneous networks. *The Journal of Supercomputing*, *81*(1), 328. ttps://doi.org/10.1007/s11227-024-06788-1

[19] Rezvani, A., Mirzaei, A., Mikaeilvand, N., Nouri-Moghaddam, B., & Gudakahriz, S. J. (2025). A Novel Framework for Enhancing Data Collection Macro-Strategies in Heterogeneous IOT Networks using Advanced Mathematical Modeling. *Archives for Technical Sciences/Arhiv za Tehnicke Nauke*, (33). https://doi.org/10.70102/afts.2025.1833.001

[20] Salami, Z. A., Bhaskaran, B., Mary, I. T. B., Ugli, I. S. S., Sripavithra, C. K., & Kalidoss, D. (2025). Energy-efficient architecture design in information service infrastructure. *Indian Journal of Information Sources and Services*, *15*(2), 168-173. https://doi.org/10.51983/ijiss-2025.IJISS.15.2.23

[21] Shahzadi, R., Ali, M., & Naeem, M. (2023). Combinatorial resource allocation in UAV-assisted 5G/B5G heterogeneous networks. *IEEE Access*, *11*, 65336-65346. https://doi.org/10.1109/ACCESS.2023.3285827

[22] Sun, Y., Peng, M., & Mao, S. (2019). A game-theoretic approach to cache and radio resource management in fog radio access networks. *IEEE Transactions on Vehicular Technology*, *68*(10), 10145-10159. https://doi.org/10.1109/TVT.2019.2935098

[23] Tang, Q., Sun, W., Liu, Z., Li, Q., & Yuan, X. (2025). Multi-agent reinforcement learning based dynamic self-coordinated topology optimization for wireless mesh networks. *Journal of Network and Computer Applications*, *239*, 104177. https://doi.org/10.1016/j.jnca.2025.104177

[24] Tilahun, F. D., Abebe, A. T., & Kang, C. G. (2023). Multi-agent reinforcement learning for distributed resource allocation in cell-free massive MIMO-enabled mobile edge computing network. *IEEE Transactions on Vehicular Technology*, *72*(12), 16454-16468. https://doi.org/10.1109/TVT.2023.3290954

[25] Urooj, S., Arunachalam, R., Alawad, M. A., Tripathi, K. N., Sukumaran, D., & Ilango, P. (2024). An effective model for network selection and resource allocation in 5G heterogeneous network using hybrid heuristic-assisted multi-objective function. *Expert systems with applications*, *248*, 123307. https://doi.org/10.1016/j.eswa.2024.123307

[26] Wang, J. H., He, H., Cha, J., Jeong, I., & Ahn, C. J. (2025). Multi-agent reinforcement learning for efficient resource allocation in internet of vehicles. *Electronics*, *14*(1), 192. https://doi.org/10.3390/electronics14010192

[27] Wang, Z., Zong, J., Zhou, Y., Shi, Y., & Wong, V. W. (2021). Decentralized multi-agent power control in wireless networks with frequency reuse. *IEEE Transactions on Communications*, *70*(3), 1666-1681. https://doi.org/10.1109/TCOMM.2021.3135540

[28] Wu, G., & Chen, G. (2025). Task offloading and resource allocation in cellular heterogeneous networks for NOMA-based mobile edge computing. *Ad Hoc Networks*, *169*, 103742. https://doi.org/10.1016/j.adhoc.2024.103742

[29] Xiao, Y., Song, Y., & Liu, J. (2023). Collaborative multi-agent deep reinforcement learning for energy-efficient resource allocation in heterogeneous mobile edge computing networks. *IEEE Transactions on Wireless Communications*, *23*(6), 6653-6668. https://doi.org/10.1109/TWC.2023.3335597

[30] Yin, S., & Yu, F. R. (2021). Resource allocation and trajectory design in UAV-aided cellular networks based on multiagent reinforcement learning. *IEEE Internet of Things Journal*, *9*(4), 2933-2943. https://doi.org/10.1109/JIOT.2021.3094651

[31] Zhao, N., Liang, Y. C., Niyato, D., Pei, Y., Wu, M., & Jiang, Y. (2019). Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, *18*(11), 5141-5152. https://doi.org/10.1109/TWC.2019.2933417