Multi-Agent Reinforcement Learning for Joint Spectrum and Energy Optimization in CR-NOMA Enabled Internet of Unmanned Agents

Saleha Ahmed, *Student Member, IEEE,* Muhammad Uzair, *Student Member, IEEE,* Syed Asad Ullah, *Student Member, IEEE,* Kapal Dev, *Senior Member, IEEE,* Aamir Mahmood, *Senior Member, IEEE,* Mikael Gidlund, *Senior Member, IEEE* and Syed Ali Hassan, *Senior Member, IEEE*

***Abstract*—With the rapid growth of Internet-of-Things (IoT) devices and unmanned agents (UAs), there is a rising need for energy- and spectrum-efficient wireless networks that can support large-scale, resource-constrained deployments. To meet this demand, integration of deep reinforcement learning (DRL), non-orthogonal multiple access (NOMA), and energy harvesting (EH) offers a promising approach to enhance energy efficiency (EE) and spectrum utilization in future sixth-generation (6G) networks, particularly for sustainable Internet of UA (IUA) communications. In this paper, we investigate an IUA network where multiple low-power secondary users (SUs), equipped with radio frequency energy harvesting (RF-EH) antennas, use a cognitive radio NOMA (CR-NOMA) scheme to share uplink channels with nearby primary users (PUs). We formulate a joint transmit power control and EH scheduling problem to maximize the long-term EE of the SUs and spectrum utilization of the network, subject to quality-of-service (QoS) constraints. To address the decentralized nature of the problem, we model the environment as a multi-agent system where each SU independently optimizes its transmission and EH strategies. A range of DRL and non-DRL algorithms is then applied to solve this optimization problem. We also explore different RF-EH diversity combining techniques to further boost system performance. Simulation results highlight the impact of these techniques on EE of SU, offering insights for optimizing performance under dynamic EH conditions.**

***Keywords*– Internet of unmanned agents (IUA), deep reinforcement learning (DRL), non-orthogonal multiple access (NOMA), energy harvesting (EH), and energy efficiency (EE).**

I. INTRODUCTION

THE next generation of wireless networks, envisioned as sixth-generation (6G), is expected to enable massive-scale connectivity, ultra-low latency, and ubiquitous intelligence across a wide range of applications, including autonomous systems, smart environments [1], and the Internet of unmanned agents (IUA) [2]. An IUA network refers to a networked ecosystem of autonomous, unmanned entities that operate with minimal or no human intervention. These agents interact, collaborate, and communicate wirelessly to perform distributed sensing, monitoring, and data transmission tasks across diverse environments. It is forecasted that the number of connected Internet-of-things (IoT) devices will reach 125 billion by 2030, generating an unprecedented demand for spectral resources and energy consumption [3]. As the IoT ecosystem continues to expand, maintaining the energy sustainability of low-power devices has become increasingly critical. This challenge underscores the importance of developing spectrally efficient, energy-aware, and self-sustaining wireless-powered communication networks (WPCNs) to support environmentally friendly IoT deployments [4].

Meanwhile, cognitive radio-inspired non-orthogonal multiple access (CR-NOMA) has attracted significant research attention as a promising technology for enhancing spectral efficiency (SE) and supporting massive connectivity in next-generation wireless networks [5], [6]. In CR networks, unlicensed secondary users (SUs) are permitted to opportunistically access the licensed spectrum of primary users (PUs), provided that interference and collision with the primary transmissions are effectively mitigated [6]. NOMA, recognized as a key multiple access scheme for fifth-generation (5G) and beyond communication systems, enables multiple users to share the same time-frequency resources by assigning different power levels and employing successive interference cancellation (SIC) at the receivers [7]. In recent years, considerable efforts have been made to explore CR-NOMA. For instance, in [8], the authors elaborated on the benefits of employing CR-NOMA, such as enhanced SE, large-scale connectivity, low latency, and improved fairness. However, these works typically rely on static system assumptions and centralized optimization, which limit scalability and adaptability in dynamic networks.

Energy harvesting (EH) technology has emerged as a transformative enabler toward building green, sustainable, and intelligent wireless networks. EH technologies, particularly radio-frequency EH (RF-EH), empower IoT devices to recharge from ambient or dedicated wireless signals, offering a self-sustainable communication paradigm that alleviates reliance on battery replacements [9], [10]. Although RF-EH offers a viable solution to address energy limitations in IUAs, its real-world performance is hindered by the non-linear and fluctuating energy conversion efficiency of practical EH cir-

Corresponding Author: S. A. Hassan is with the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), 44000, Islamabad, Pakistan, and also with the Department of Electronic Engineering, Kyung Hee University, Yongin 17104, South Korea (email: ali.hassan@seecs.edu.pk).

S. Ahmed, M. Uzair, and S. A. Ullah are with the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), 44000, Islamabad, Pakistan (email: {sahmed.bese21seecs, muzair.bese21seecs, sullah.phdee21seecs}@seecs.edu.pk).

S. A. Ullah, is also with the Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS), 87300, Quetta, Pakistan, (email: syed.asad@buitms.edu.pk).

A. Mahmood and M. Gidlund are with the Department of Computer and Electrical Engineering, Mid Sweden University, 851 70 Sundsvall, Sweden (e-mail: {aamir.mahmood, mikael.gidlund}@miun.se).

K. Dev is with Adapt and the Department of Computer Science, Munster Technological University, Ireland, (email: kapal.dev@ieee.org)

cuits, which varies with the input power level from ambient RF sources. Several studies have explored techniques to enhance RF-EH capabilities. For instance, the authors in [11] proposed a novel RF-EH circuit aimed at improving the operational efficiency of battery-less IoT devices through advanced circuit design. In [12], the use of microwave antennas for EH applications was investigated. To efficiently capture arbitrarily polarized signals, a dual-band aperture-coupled patch rectenna with high conversion efficiency was introduced in [13]. Additionally, [14] presented the design and analysis of a hexagonal fractal antenna array (HFAA) tailored for next-generation wireless communication. In [15], a triple-band antenna was developed to effectively harvest RF energy from commonly used frequency bands, including those of cellular and Wi-Fi networks. While these works enhance hardware-level RF-to-DC conversion efficiency, they primarily address the physical layer and do not consider the efficient EH creating a gap between circuit-level EH advancements and intelligent resource management strategies. To enhance the sustainability of EH-enabled IUAs, it is essential to harvest sufficient energy from ambient RF signals. This necessitates the investigation of diversity combining techniques at the receiver side, which can significantly improve the efficiency of RF-EH. A detailed discussion of these techniques is provided in the subsequent sections.

The convergence of EH, CR, and NOMA, collectively termed as EH-CR-NOMA, offers a promising solution to several critical challenges in 5G and beyond wireless networks. These include improving SE, minimizing energy consumption, and enabling massive device connectivity, all while ensuring reliable communication between PUs and SUs. Extensive research has explored various combinations of these technologies. For example, in [16], integrated EH-CR systems have been investigated, focusing on mechanisms such as RF-EH from PUs and opportunistic data transmission in idle channels. In [17], the authors introduced a 3D matching framework for machine-to-machine EH-CR networks to enhance EE, while a Lyapunov-based optimization scheme was proposed in [18] for EH-CR wireless sensor networks (WSN). Simultaneous wireless information and power transfer (SWIPT), a specialized form of EH, has also been studied in NOMA-aided networks. In [19], an optimized user pairing and power allocation approach was proposed for SWIPT-NOMA systems, targeting improvements in both spectral and EE. Resource management remains a focal challenge in EH-CR-NOMA systems, particularly the trade-off between EH and data transmission. For instance, [20] proposed a two-layer optimization method leveraging Dinkelbach's algorithm for joint time and power control under equal switching time constraints across users. Other works, such as [21] and [22], examined EH-CR-NOMA systems with greedy energy usage policies and overlay CR modes, respectively, using bisection search and Dinkelbach methods for resource allocation. However, these are often constrained to single PU-SU scenarios or idealized assumptions like perfect spectrum sensing. Additional contributions in EH-enabled WSNs [23], heterogeneous cellular networks [24], and more highlight limitations in assuming perfect channel knowledge and deterministic energy arrival conditions rarely met in practical scenarios.

In EH-CR-NOMA-based systems, the integration of IUA introduces new challenges, including dynamic mobility patterns, intermittent connectivity, heterogeneous EH capabilities, and stringent latency/quality-of-service (QoS) constraints. These factors complicate the joint resource optimization process, demanding more adaptive and intelligent strategies such as deep reinforcement learning (DRL) to achieve sustainable and efficient system performance. To efficiently exploit the potential of EH and CR-NOMA in IUA networks, it is imperative to design intelligent resource allocation mechanisms that can dynamically adapt to the time-varying wireless environment. Particularly, low-cost UAs equipped with EH capabilities must judiciously manage their harvested energy and transmission opportunities while ensuring the QoS requirements are met [25]. Traditional optimization-based methods often assume perfect channel state information (CSI) and static environmental models, making them unsuitable for practical deployments characterized by uncertainty, variability, and incomplete information.

Given the uncertainty and dynamics of wireless environments and energy availability, DRL has emerged as a viable approach for adaptive decision-making. DRL's ability to operate in unknown and time-varying conditions makes it particularly suitable for tasks such as resource allocation, dynamic spectrum access, network security, and data offloading [26] [27]. Recent works have integrated DRL into EH-enabled wireless communication systems for optimizing long-term metrics like throughput and EE. For example, [28] applied RL to power control in underwater EH relay networks, and [29] proposed a DRL-based centralized resource scheduling policy for EH-IoT systems. Other efforts include DRL-based access control in wireless local area networks (WLANs) [30], energy management using deep deterministic policy gradient (DDPG) [31]–[33], and offloading decisions in EH-mobile edge computing (MEC) systems [34]. Notably, DDPG has shown strong performance in continuous action space problems like joint time and power allocation, as studied in [31], [35]. The work in [36] introduces a distributed multi-agent framework for EH-CR-NOMA using DDPG with an action adjustment mechanism to improve convergence, however, remains limited in scope. The reliance on a single DRL algorithm overlooks more advanced mutli-agent DRL (MADRL) methods.

Overall, current DRL-based approaches predominantly rely on centralized scheduling or focus on single-user scenarios, failing to address the stringent demands of practical multi-user EH-CR-NOMA systems. In such settings, distributed decision-making and minimal communication overhead are not just desirable but essential, making the limitations of existing methods a critical barrier to real-world deployment.

Motivated by these challenges and opportunities, in this paper, a decentralized DRL-based resource management framework for an EH-IUA network with multiple devices (or agents) is proposed. We formulate a MADRL-based resource allocation scheme that provides scalable and adaptive performance under stochastic environmental dynamics, aligning with the vision of intelligent and self-organizing IUA networks. In the

proposed MADRL framework, each EH IoT device independently learns its optimal strategy for transmit power control and EH scheduling to maximize its long-term energy efficiency (EE) while satisfying QoS constraints. Our approach enables fully decentralized learning and decision-making among the UAs, leveraging the strengths of EH for energy sustainability, CR for spectrum access, and NOMA for efficient user multiplexing. Additionally, the proposed framework incorporates diversity combining techniques to enhance RF signal reception and employs a set of state-of-the-art DRL algorithms, along with baseline non-DRL algorithms, to evaluate and identify a robust algorithm well-suited for diverse wireless environments. Accordingly, the key contributions of our work can be summarized as follows.

- We develop a practical EH-enabled CR-NOMA IUA framework that supports timely data transmission for multiple energy-constrained SUs in the presence of multiple PUs. To ensure this, each time slot is dynamically partitioned into two phases: (i) a data transmission phase, where the SU utilizes the energy stored in its rechargeable battery to send data to the BS, and (ii) an EH phase, during which the SU harvests and stores RF energy from nearby PUs for future transmissions.
- We propose a multi-agent DRL framework to maximize the EE of the IUA network, where each SU independently learns its optimal strategy for transmit power control and EH scheduling to maximize its long-term EE while satisfying QoS constraints. The proposed approach enables fully distributed learning and decision-making among the UAs, leveraging the strengths of EH for energy sustainability, CR for spectrum access, and NOMA for efficient user multiplexing.
- We incorporate three RF-EH diversity combining techniques—equal gain combining (EGC), maximum ratio combining (MRC), and selection combining (SC) [37]—to improve RF signal reception, while explicitly accounting for the power consumption associated with signal processing and RF circuitry at the SU.
- To enable robust performance under varying wireless environments, we evaluate a set of DRL algorithms, including DDPG [38], proximal policy optimization (PPO) [39], soft actor critic (SAC) [40] and twin delayed DDPG (TD3) [41], along with baseline non-DRL strategies such as the random and greedy methods [6]. Furthermore, we analyze the impact of key system parameters, such as the number of RF-EH antennas and the transmit power of PUs, on EH performance and the achievable sum rate, and report the SU's time slot utilization patterns under different combining techniques.

The remainder of the paper has the following organization: Section II introduces the system model, develops an in-depth justification for the application of DRL, and presents a detailed discussion of the diversity combining techniques employed in the paper. In Section III, we formulate the optimization problem at hand, followed by Section IV, which elaborates on the fundamental principles of the implemented DRL and non-DRL algorithms and provide a detailed overview of the simulation setup. Section V presents comprehensive simulation results, whereas Section VI concludes the paper.

TABLE I: Summary of key notations.

Symbol	Description
J	Total number of primary users (PUs), indexed by j
K	Total number of secondary users (SUs), indexed by k
N	Number of time slots per episode, indexed by n
τ	Duration of a single time slot (in seconds)
$E_{k,n}$	Available energy of SU k at the start of slot n
Ω_j	Transmit power of PU j
$\tilde{\Omega}_{k,n}$	Transmit power selected by SU k in slot n
$\zeta_{k,n}$	Time allocation factor for SU k in slot n (transmission)
ρ_k	Static circuit power consumption of SU k
$h_{j,k,q}$	Channel gain between PU j and antenna q of SU k
$\tilde{h}_k$	Channel gain between SU k and the base station
Q	Number of RF-EH antennas per SU
η	RF-to-DC energy conversion efficiency
d	Distance between PU and SU
γ	Path loss exponent
$R_{k,n}$	Data rate of SU k in time slot n
ϕ	Minimum QoS threshold (required rate)
ψ	Discount factor for long-term EE objective
$E_{\max}$	Maximum battery capacity of each SU

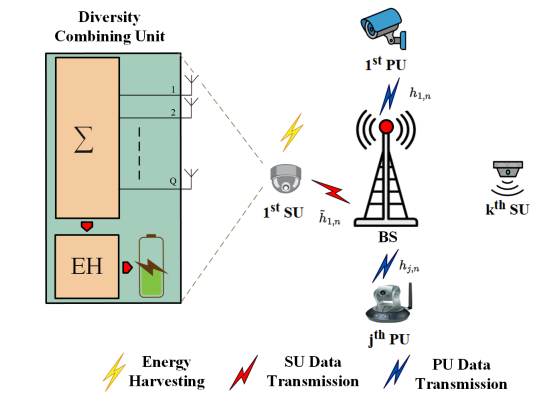
II. System Model Description

As illustrated in Fig. 1a, we consider a WPCN composed of a centralized base station (BS), a set of PUs $\{P_j\}_{j=1}^{J}$, and a set of SUs $\{S_k\}_{k=1}^{K}$, where J and K denote the total number of PUs and SUs, respectively. Each SU is equipped with RF-EH capability and a rechargeable energy storage unit. Table I presents a detailed summary of the key notations used.

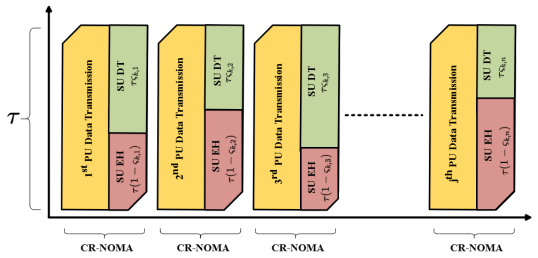
The PUs communicate with the BS in a time-division multiple access (TDMA) manner. The communication timeline is organized into episodes, with each episode consisting of N timeslots of equal duration of τ seconds, resulting in a total episode duration of $N\tau$ seconds. It is assumed that $N \geq J$, ensuring that every PU obtains at least one transmission opportunity within an episode. The transmission schedule is cyclic i.e. in the n-th timeslot, the transmitting PU is determined by $j = ((n-1) \oplus J)+1$, where $n \in \{1, 2, \ldots, N\}$ and $\oplus$ denotes the modulo operation. For example, when $J = 2$ and $N = 4$, PU P_1 transmits in timeslots t_1 and t_3, while PU P_2 transmits in t_2 and t_4. Each PU transmits to the BS with a constant transmit power Ω_j.

The SUs opportunistically access the spectrum using CR-NOMA without violating the QoS requirements of the PUs. Each SU S_k operates under stringent energy constraints, relying on ambient RF energy to replenish its battery. During each timeslot n, an SU S_k can dynamically split the slot between data transmission and EH activities. Let $\mathcal{E}_{k,n}$ denote the available energy level of S_k at the beginning of timeslot n. Based on its energy level and channel conditions, SU S_k chooses a time-splitting coefficient $\varsigma_{k,n} \in [0, 1]$, such that $\varsigma_{k,n}\tau$ seconds are allocated for data transmission, and $(1-\varsigma_{k,n})\tau$ seconds are used for harvesting energy.

Each SU is equipped with Q RF-EH antennas. Let $\tilde{h}_{j,k,q}$ denote the channel gain between PU P_j and the q-th antenna of SU S_k, where $1 \leq q \leq Q$. The received RF signals during the EH phase are combined using diversity techniques to maximize the harvested energy. The overall channel gain between the j^{th}



(a)



(b)

Fig. 1: Overview of considered IoT (or WPCN) network: (a) system model design, (b) representation of TDMA-based PU transmissions and CR-NOMA implementation for SD transmissions.

PU and k^{th} SU is denoted by $\tilde{h}_{j,k}$. During its transmission phase, SU S_k communicates with the BS using a transmit power $\tilde{\Omega}_{k,n}$ over $\varsigma_{k,n}\tau$ seconds. The channel gain between SU S_k and the BS is denoted by $\tilde{h}_k$. Each SU must ensure that its transmit power $\tilde{\Omega}_{k,n}$ satisfies its energy availability constraint. In this work, we assume that each SU has perfect knowledge of its CSI and battery state. Although this may not always hold in practice, it is a common assumption in DRL-based wireless studies to evaluate upper-bound performance. Based on this information, they independently determine their optimal $\varsigma_{k,n}$ and $\tilde{\Omega}_{k,n}$ policies.

In contrast to the conventional single-agent design, this work considers a decentralized MADRL framework, where each SU S_k is modeled as an independent learning agent. Each agent observes its local state, comprising its residual battery energy and channel gains, and takes actions by selecting its time-sharing and power allocation decisions. The agents aim to maximize their long-term expected cumulative throughput while maintaining energy sustainability. Through this decentralized learning framework, the system achieves enhanced spectral and EE compared to traditional single-SU architectures and can adaptively respond to varying network dynamics such as changing PU activity patterns and channel fluctuations.

A. Motivation for Employing DRL

The optimization problem addressed in this study inherently involves sequential decision-making under uncertainty, where agents must occasionally sacrifice immediate performance to achieve improved long-term outcomes. Such temporal trade-offs naturally motivate the application of DRL, which is well-suited for learning dynamic, long-horizon strategies in complex wireless environments.

To illustrate this rationale, consider a representative scenario involving two PUs, denoted as P_1 and P_2, and two SUs, say S_1 and S_2, wherein each SU is equipped with a single RF-EH antenna. Let the instantaneous data rate achieved by these SUs during the n-th and $(n+1)$-th timeslot be expressed as

$$\tilde{\mathcal{R}}_n = \varsigma_{1,n} \log_2 \left(1 + \frac{\tilde{\Omega}_{1,n}|\tilde{h}_1|^2}{1 + \Omega_1|\tilde{h}_{1,1}|^2}\right), \tag{1}$$

and

$$\tilde{\mathcal{R}}_{n+1} = \varsigma_{2,n+1} \log_2 \left(1 + \frac{\tilde{\Omega}_{2,n+1}|\tilde{h}_2|^2}{1 + \Omega_2|h_{2,2}|^2}\right), \tag{2}$$

respectively, where S_1 and P_1 are active in n-th timeslot and S_2 and P_2 are active in $(n+1)$-th timeslot.

Assuming that P_1 is transmitting in a specific timeslot and it exhibits strong channel conditions to both the BS and the SU. In such a setting, if a SU performs transmission, it would result in severe co-channel interference from a strong channel of P_1, significantly degrading the achievable rate, as evident in (1). Instead, the SU can opportunistically choose to harvest energy from the strong RF emissions of P_1, leveraging its high channel gain to accumulate energy for future use. Thus, EH becomes a more advantageous action in the short term, even if it forgoes immediate data transmission. In contrast, consider the case where P_2 is characterized by weak channels to both the BS and the SU. The RF signals emitted by P_2 offer negligible EH potential. Simultaneously, the weak interference from P_2 enables the SU to achieve a relatively high data rate through CR-NOMA. Therefore, in such scenarios, data transmission is the preferred action.

The central challenge lies in designing a control policy that dynamically selects between EH and data transmission actions based on time-varying and uncertain channel states. DRL offers a principled framework for learning such a policy by interacting with the environment and discovering optimal state-action mappings that maximize cumulative performance. Additionally, DRL algorithms provide real-time advantages for IUA networks. Since actions are taken once per slot ($\tau = 1$ s), inference must finish within this deadline. This is feasible because each SU's policy depends only on its local state as discussed in Section II-C, enabling lightweight feedforward execution on embedded hardware. Decentralized operation further reduces coordination latency and scales efficiently with the number of SUs.

In subsequent sections, we demonstrate how a DRL agent effectively captures these nuanced trade-offs and adapts to diverse wireless contexts, leading to robust system performance across varying interference and EH conditions.

B. Diversity Combining Techniques for RF-EH

This subsection provides a comprehensive overview of diversity combining techniques employed within RF-EH systems. These techniques are incorporated into the system model, where each SU is equipped with Q RF antennas and an EH chain

consisting of a diversity combining unit, an EH circuit, and an energy storage module.

Diversity combining aims to leverage spatial diversity by combining multiple versions of the same signal with varying fadings received at multiple antennas. While traditionally used to enhance SNR in fading environments, these techniques can be adapted for RF-EH to maximize the total harvested power instead. In the proposed model, ambient RF signals received at the Q antennas are combined and then used to harvest energy, which is then stored in a battery. Assuming flat fading channels and Q antennas, the harvested power from the j-th PU at an SU using a generic diversity combining strategy is given as [37], [42]

$$P_h = \eta\tau d^{-\gamma}\Omega_j \left| \sum_{q=1}^{Q} e_q \tilde{h}_{j,k,q} \right|^2, \tag{3}$$

here, $\eta \in [0, 1]$ is the RF-to-DC conversion efficiency, T is the EH duration, d is the PU–SU distance, and γ is the path loss exponent. Ω_j denotes the transmit power of the j-th PU, $\tilde{h}_{j,k,q}$ is the complex channel gain to the q-th antenna of the k-th SU, and e_q is the complex combining weight for that antenna.

1) MRC: MRC uses the complex conjugate of the channel gains as weights to coherently combine signals, maximizing received energy through constructive addition. The harvested power at the SU in the n-th time slot using MRC is given by [37]

$$P_{n,\text{mrc}} = \eta\tau d^{-\gamma}\Omega_j \sum_{q=1}^{Q} |\tilde{h}_{j,k,q}|^2 - \theta_{\text{mrc}} - \tilde{\theta}_{\text{mrc}}, \tag{4}$$

where θ_{mrc} is representative of the MRC weight as a result of imperfect combining and $\tilde{\theta}_{\text{mrc}}$ denotes the total power consumed by the MRC circuitry.

2) EGC: The RF receiver with EGC combines signals from all antennas using equal weighting, aiming to improve the received SNR without applying complex gain adjustments, offering less complexity compared to MRC. The harvested power at the SU using EGC in the n-th slot is [37]

$$P_{n,\text{egc}} = \frac{\eta\tau d^{-\gamma}\Omega_j}{Q} \left(\sum_{q=1}^{Q} |\tilde{h}_{j,k,q}| \right)^2 - \theta_{\text{egc}} - \tilde{\theta}_{\text{egc}}, \tag{5}$$

where θ_{egc} and $\tilde{\theta}_{\text{egc}}$ denote the weight inefficiency and total circuit power consumption for EGC, respectively.

3) SC: SC is the simplest form of diversity, where only the antenna with the highest instantaneous received power is selected for EH. While it eliminates the complexity of signal combining, it forgoes the energy benefits offered by multipath diversity. The harvested power at the n-th time slot under SC is given by

$$P_{n,\text{sc}} = \max_{m} \left(\eta\tau d^{-\gamma}\Omega_j |\tilde{h}_{j,k,q}|^2 \right) - \tilde{\theta}_{\text{sc}}, \tag{6}$$

where $\tilde{\theta}_{\text{sc}}$ denotes the power overhead of implementing SC.

These diversity combining techniques vary significantly in terms of performance, complexity, and energy consumption. This directly impacts their suitability for real-time operation. SC offers the lowest complexity but limited efficiency, EGC provides a balanced trade-off with moderate overhead, and MRC achieves the highest efficiency at the cost of increased processing burden. In Section V, we assess their influence on the overall performance of the SUs.

C. Best SU Selection Criterion

To select the optimal SU for transmission, a quality score $Q(k)$ is computed for each SU k using a centralized external policy that is defined as follows

$$Q(k) = \sum_{x \in \mathcal{X}} \frac{x_k - \min(x_k)}{\max(x_k) - \min(x_k)}, \tag{7}$$

where $\mathcal{X} = \{\tilde{h}_{j,k},\ \tilde{h_k},\ \mathcal{E}_k\}$ denotes the set of considered metrics for each SU. Based on the computation $Q(k)$, the SU selection is performed according to

$$\text{SU}^* = \begin{cases} random_{k\in\{1,\ldots,K\}}, & \text{with probability } \alpha \\ argmax_{k\in\{1,\ldots,K\}} Q(k), & \text{with probability } 1-\alpha \end{cases} \tag{8}$$

where K denotes the total number of available SUs.

III. Problem Formulation

In this section, we present and formulate the MADRL optimization problem such that each SU jointly optimizes its time-sharing coefficient and transmit power to maximize long-term throughput while satisfying energy sustainability and interference constraints. Consequently, the overall EE of the network is maximized; thereby, the instantaneous EE of k-th SU S_k in timeslot n is defined as

$$EE_{k,n}(\varsigma_{k,n}, \tilde{\Omega}_{k,n}) = \frac{\varsigma_{k,n} \log_2\left(1 + \frac{\tilde{\Omega}_{k,n}|\tilde{h}_{k,n}|^2}{1+\Omega_{j,n}|h_{j,k}|^2}\right)}{(\tilde{\Omega}_{k,n} + \rho_{k,n})\tau}, \tag{9}$$

here, $\rho_{k,n}$ captures static power consumption associated with the circuit operations and signal processing used by the SU in the n-th time slot; including $(\tilde{\Omega}_{k,n} + \rho_{k,n})\tau$ in the denominator ensures the objective penalizes energy usage, aligning the learned policy with energy-aware behavior. In the proposed multi-agent framework, each SU aims to maximize its EE within its designated transmission time slot, based on its individual transmission capabilities. Consequently, the problem is formulated as the maximization of the average EE attained per time slot within the EGC diversity environment, which is given by

$$\underset{\varsigma_{k,n},\ \tilde{\Omega}_{k,n}}{\text{maximize}}\ \mathbb{E}\left\{ \sum_{n=1}^{N} \psi^{n-1} \frac{\varsigma_{k,n} \log_2\left(1 + \frac{\tilde{\Omega}_{k,n}|\tilde{h}_{k,n}|^2}{1+\Omega_{j,n}|h_{j,k}|^2}\right)}{(\tilde{\Omega}_{k,n} + \rho_{k,n})\tau} \right\} \tag{P1}$$

$$\text{s.t.}\quad R_{k,n} \geq \varphi, \tag{P1a}$$

$$0 \leq \varsigma_{k,n} \leq 1, \tag{P1b}$$

$$\varsigma_{k,n}\tau(\tilde{\Omega}_{k,n} + \rho_{k,n}) \leq \mathcal{E}_{k,n}, \tag{P1c}$$

$$0 \leq \tilde{\Omega}_{k,n} \leq \tilde{\Omega}_{\max}, \tag{P1d}$$

$$\frac{(1-\varsigma_{k,n})\tau\eta d^{-\gamma}\Omega_{j,n}}{Q} \left(\sum_{q=1}^{Q} |\tilde{h}_{j,k,q}| \right)^2 \leq \max\left\{\Omega_{j,n}, \mathcal{E}_{\max}\right\}, \tag{P1e}$$

$$\mathcal{E}_{k,n+1} = \min\left\{\frac{(1-\varsigma_{k,n})\tau\eta d^{-\gamma}\Omega_{j,n}}{Q}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{egc} - \tilde{\theta}_{egc} - \varsigma_{k,n}\tau(\tilde{\Omega}_{k,n}+\rho_{k,n}) + \mathcal{E}_{k,n}, \mathcal{E}_{\max}\right\}, \tag{P1f}$$

where constraint (P1a) enforces QoS by requiring the achievable rate of the SU to meet a minimum threshold φ in each time slot. Constraint (P1b) restricts the time-sharing factor $\varsigma_{k,n}$ to lie between 0 and 1, ensuring valid time allocation. Constraint (P1c) guarantees that the energy consumed for both transmission and circuit operations during the allocated transmission period does not exceed the available battery energy. Constraint (P1d) ensures that the SU's transmit power remains within its hardware-supported limits, promoting safe and feasible power control. Constraint (P1e) regulates the energy harvested from the PU's RF signals, ensuring it does not exceed the SU's maximum energy storage capacity or the PU's transmission limits, thereby preserving battery health and promoting energy conservation. Our analysis shows that (P1e) is binding in roughly 50% of the time steps per episode, indicating that SUs frequently operate near their energy limits. This frequent activation underscores the practical significance of this constraint, as it directly shapes the balance between energy-limited and channel-limited operation. Increasing the SU's storage capacity or the PU's transmit power would reduce the frequency with which (P1e) becomes active, potentially allowing SUs to store and utilize more energy, which could improve overall system performance and flexibility. Lastly, Constraint (P1f) models the SU's battery dynamics, ensuring that the updated energy level for the next time slot reflects net energy change while remaining within the battery's maximum storage limit.

We observe that Problem (P1) is non-convex due to the coupling between time-sharing and power allocation variables under stochastic energy dynamics. To address this, we decompose it into two subproblems. The process starts with defining $\bar{\mathcal{E}}_{k,n}$, the difference between the harvested and consumed energy.

$$\bar{\mathcal{E}}_{k,n} = \Bigg\{\underbrace{(1-\varsigma_{k,n})\tau\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2}_{\text{Harvested Energy}} - \underbrace{\theta_{egc} - \tilde{\theta}_{egc} - \varsigma_{k,n}\tau(\tilde{\Omega}_{k,n}+\rho_{k,n})}_{\text{Consumed Energy}}\Bigg\}. \tag{10}$$

In the first subproblem, we derive closed-form expressions for the optimal time-splitting factor $\varsigma_{k,n}(\hat{\mathcal{E}}_{k,n})$ and transmit power $\tilde{\Omega}_{k,n}(\hat{\mathcal{E}}_{k,n})$ using convex optimization based on the estimated harvested energy $\hat{\mathcal{E}}_{k,n}$. The second subproblem models the residual energy decision $\hat{\mathcal{E}}_{k,n}$ as a continuous action in a DRL framework. This action governs EH and transmission behavior over time to maximize cumulative efficiency. Accordingly, the first subproblem is formulated as:

$$\max_{\varsigma_{k,n},\,\tilde{\Omega}_{k,n}} \quad EE_{k,n}(\varsigma_{k,n},\tilde{\Omega}_{k,n}) \tag{P2}$$

$$\text{s.t.} \quad \hat{\mathcal{E}}_{k,n} = \left\{\frac{(1-\varsigma_{k,n})\tau\eta d^{-\gamma}\Omega_{j,n}}{Q}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_i - \tilde{\theta}_{\tilde{i}} - \varsigma_{k,n}\tau(\tilde{\Omega}_{k,n}+\rho_{k,n})\right\},$$

$$\text{(P1a), (P1b), (P1c), (P1d)} \tag{P2b}$$

for i ={EGC, MRC} and $\tilde{i}$={EGC, MRC, SC}, where θ_i and $\tilde{\theta}_{\tilde{i}}$ correspond to the combiner efficiency loss under the specified diversity combining scheme. The second subproblem is then expressed as

$$\max_{\hat{\mathcal{E}}_{k,n}} \quad \mathbb{E}\left\{\sum_{n=1}^{N}\frac{\psi^{n-1}\varsigma_{k,n}(\hat{\mathcal{E}}_{k,n})}{(\tilde{\Omega}_{k,n}(\hat{\mathcal{E}}_{k,n})+\rho_{k,n})\tau}\log_2\left(1+\frac{\tilde{\Omega}_{k,n}(\hat{\mathcal{E}}_{k,n})|\tilde{h}_{k,n}|^2}{1+\Omega_{j,n}|h_{j,k}|^2}\right)\right\} \tag{P3}$$

$$\text{s.t.} \quad \mathcal{E}_{k,n+1} = \min\left\{\mathcal{E}_{k,n} + \hat{\mathcal{E}}_{k,n}, \mathcal{E}_{\max}\right\}. \tag{P3a}$$

In this formulation, the closed-form mappings $\varsigma_{k,n}(\hat{\mathcal{E}}_{k,n})$ and $\tilde{\Omega}_{k,n}(\hat{\mathcal{E}}_{k,n})$ guide the decision process, enabling each agent to adaptively respond to its local environment and energy profile using DRL. Consequently, the closed-form solutions for EGC diversity are given as

$$\varsigma^*_{k,n^*_E}(\hat{\mathcal{E}}_{k,n}) = \min\left\{1, \mathcal{E}^*_E, \max\left\{\bar{\varsigma}_{k,n^*_E}, 0, A^*_E, B^*_E, C^*_E, D^*_E\right\}\right\}, \tag{11}$$

and

$$\tilde{\Omega}^*_{k,n^*_E}(\hat{\mathcal{E}}_{k,n}) = \frac{\left(1-\varsigma^*_{k,n^*_E}(\hat{\mathcal{E}}_{k,n})\right)\tau\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{egc}}{Q\varsigma^*_{k,n^*_E}(\hat{\mathcal{E}}_{k,n})\tau} - \frac{\tilde{\theta}_{egc} + \varsigma^*_{k,n^*_E}(\hat{\mathcal{E}}_{k,n})\tau\rho_{k,n} + \hat{\mathcal{E}}_{k,n}}{Q\varsigma^*_{k,n^*_E}(\hat{\mathcal{E}}_{k,n})\tau}, \tag{12}$$

where $\bar{\varsigma}_{k,n^*_E}$ represents the optimal time-sharing coefficient for EGC following the formulation in [43]. The expressions for the sub-variables in Eq. 11 are given by

$$A^*_E = 1 - \frac{Q\Omega_{j,n}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2},$$

$$B^*_E = 1 - \frac{Q\tilde{\Omega}_{\max}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2},$$

$$C^*_E = \frac{\tau\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{egc} - \tilde{\theta}_{egc} - \hat{\mathcal{E}}_{k,n}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 + \tau\rho_{k,n} - Q\tau\tilde{\Omega}_{\max}},$$

$$D^*_E = \frac{\tau\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{egc} - \tilde{\theta}_{egc} - \hat{\mathcal{E}}_{k,n} - Q\mathcal{E}_{k,n}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 + \tau\rho_{k,n}(1+Q)},$$

$$E^*_E = \frac{\tau\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{egc} - \tilde{\theta}_{egc} - \hat{\mathcal{E}}_{k,n}}{\tau\left(\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 + \rho_{k,n}\right)}.$$

Similarly, the closed-form solutions for MRC diversity are given as

$$\varsigma^*_{k,n^*_M}(\hat{\mathcal{E}}_{k,n}) = \min\left\{1, \mathcal{E}^*_M, \max\left\{\bar{\varsigma}_{k,n^*_M}, 0, A^*_M, B^*_M, C^*_M, D^*_M\right\}\right\}, \tag{13}$$

and

$$\begin{aligned}\tilde{\Omega}^*_{k,n^*_M}(\hat{\mathcal{E}}_{k,n}) = &\frac{\left(1-\varsigma^*_{k,n^*_M}(\hat{\mathcal{E}}_{k,n})\right)\tau\eta d^{-\gamma}\Omega_{j,n}\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|^2-\theta_{mrc}}{\varsigma^*_{k,n^*_M}(\hat{\mathcal{E}}_{k,n})\tau} \\ &- \frac{\tilde{\theta}_{mrc}+\varsigma^*_{k,n^*_M}(\hat{\mathcal{E}}_{k,n})\tau\rho_{k,n}+\hat{\mathcal{E}}_{k,n}}{\varsigma^*_{k,n^*_M}(\hat{\mathcal{E}}_{k,n})\tau},\end{aligned} \tag{14}$$

where $\bar{\varsigma}_{k,n^*_M}$ is the optimal value of $\varsigma_{k,n}$ for the case of MRC diversity [43]. The expressions for sub-variables in Eq. 13 are given by

$$A^*_M = 1 - \frac{\Omega_{j,n}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2},$$

$$B^*_M = 1 - \frac{\tilde{\Omega}_{\max}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2},$$

$$C^*_M = \frac{\tau\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{mrc} - \tilde{\theta}_{mrc} - \hat{\mathcal{E}}_{k,n}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 + \tau(\rho_{k,n}+\tilde{\Omega}_{\max})},$$

$$D^*_M = \frac{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{mrc} - \tilde{\theta}_{mrc} - \hat{\mathcal{E}}_{k,n} - \mathcal{E}_{k,n}}{\eta\tau d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 + \tau\rho_{k,n}},$$

$$E^*_M = \frac{\tau\eta d^{-\gamma}\Omega_{j,n}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{mrc} - \tilde{\theta}_{mrc} - \hat{\mathcal{E}}_{k,n}}{\tau\left(\eta d^{-\gamma}\Omega_{j,n}\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|^2+\rho_{k,n}\right)}.$$

Furthermore, the closed-form solutions for SC diversity are given as

$$\varsigma^*_{k,n^*_S}(\hat{\mathcal{E}}_{k,n}) = \min\left\{1, \mathcal{E}^*_S, \max\left\{\bar{\varsigma}_{k,n^*_S}, 0, A^*_S, B^*_S, C^*_S, D^*_S\right\}\right\}, \tag{15}$$

and

$$\begin{aligned}\tilde{\Omega}^*_{k,n^*_S}(\hat{\mathcal{E}}_{k,n}) = &\frac{\left(1-\varsigma^*_{k,n^*_S}(\hat{\mathcal{E}}_{k,n})\right)\max(\tau\eta d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2)}{\varsigma^*_{k,n^*_S}(\hat{\mathcal{E}}_{k,n})\tau} \\ &- \frac{\tilde{\theta}_{sc}+\varsigma^*_{k,n_{sc}}(\hat{\mathcal{E}}_{k,n})\tau\rho_{k,n}+\hat{\mathcal{E}}_{k,n}}{\varsigma^*_{k,n_{sc}}(\hat{\mathcal{E}}_{k,n})\tau},\end{aligned} \tag{16}$$

where $\bar{\varsigma}_{k,n^*_S}$ is the optimal value of $\varsigma_{k,n}$ for SC diversity [43]. The expressions for sub-variables in Eq. 15 are given by

$$A^*_S = 1 - \frac{\Omega_{j,n}}{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right)},$$

$$B^*_S = 1 - \frac{\tilde{\Omega}_{\max}}{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right)},$$

$$C^*_S = \frac{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right) - \tilde{\theta}_{sc} - \hat{\mathcal{E}}_{k,n}}{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right) + \tau(\rho_{k,n}+\tilde{\Omega}_{\max})},$$

$$D^*_S = \frac{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right) - \tilde{\theta}_{sc} - \hat{\mathcal{E}}_{k,n} - \mathcal{E}_{k,n}}{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right) - \tau\rho_{k,n}},$$

$$E^*_S = \frac{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right) - \tilde{\theta}_{sc} - \hat{\mathcal{E}}_{k,n}}{\max\left(\eta\tau d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2\right) + \tau\rho_{k,n}}.$$

IV. Fundamentals of the Implemented DRL and Non-DRL Algorithms

In this section, we present the foundational principles of the DRL and non-DRL algorithms employed in the optimization problem.

A. DRL Algorithms

1) DDPG: The DDPG algorithm is a widely adopted, model-free, off-policy DRL technique tailored for environments with continuous action spaces. It operates on an actor-critic framework, where the actor network maps states to deterministic actions, and the critic network evaluates the quality of these actions using a learned action-value function $Q(s, a)$. Unlike traditional value-based methods such as Q-learning or SARSA, which rely on discrete and tabulated representations of $Q(s, a)$, DDPG utilizes deep neural networks to approximate the action-value function over continuous domains [38].

DDPG maintains a replay buffer to store past transitions (s, a, r, s'), which are sampled uniformly during training to improve learning stability and sample efficiency. Moreover, it uses target networks, which are delayed copies of the actor and critic, to mitigate instability due to rapid parameter changes during learning. The core objective of the actor is to maximize the expected Q-value, leading to the optimal policy defined as $a^*(s) = \arg\max_a Q(s, a)$. Time complexity per update is dominated by actor and centralized critic network evaluations, scaling approximately as $O(K{\cdot}P{\cdot}Q{\cdot}R^2{+}P{\cdot}Q{\cdot}R^2)$, where P is the batch size, Q the number of layers, and R the number of neurons per layer. This algorithm forms the backbone of many advanced DRL variants used for real-time decision-making in complex, dynamic environments such as EH-enabled CR-NOMA-assisted IUA systems.

2) TD3: TD3 enhances the baseline DDPG algorithm by addressing its known vulnerability to overestimation bias in Q-value predictions. Like DDPG, TD3 uses an actor-critic architecture; however, it introduces three key improvements for enhanced training stability and performance [41].

1) Twin Q-networks: Two independent critic networks are trained, and the smaller of their Q-value estimates is used during target value computation. This conservative approach reduces overestimation and encourages more cautious policy updates.
2) Delayed Policy Updates: Unlike DDPG, which updates both actor and critic at each timestep, TD3 updates the actor and target networks less frequently than the critic. This decoupling stabilizes training by allowing the critic to converge before updating the actor.

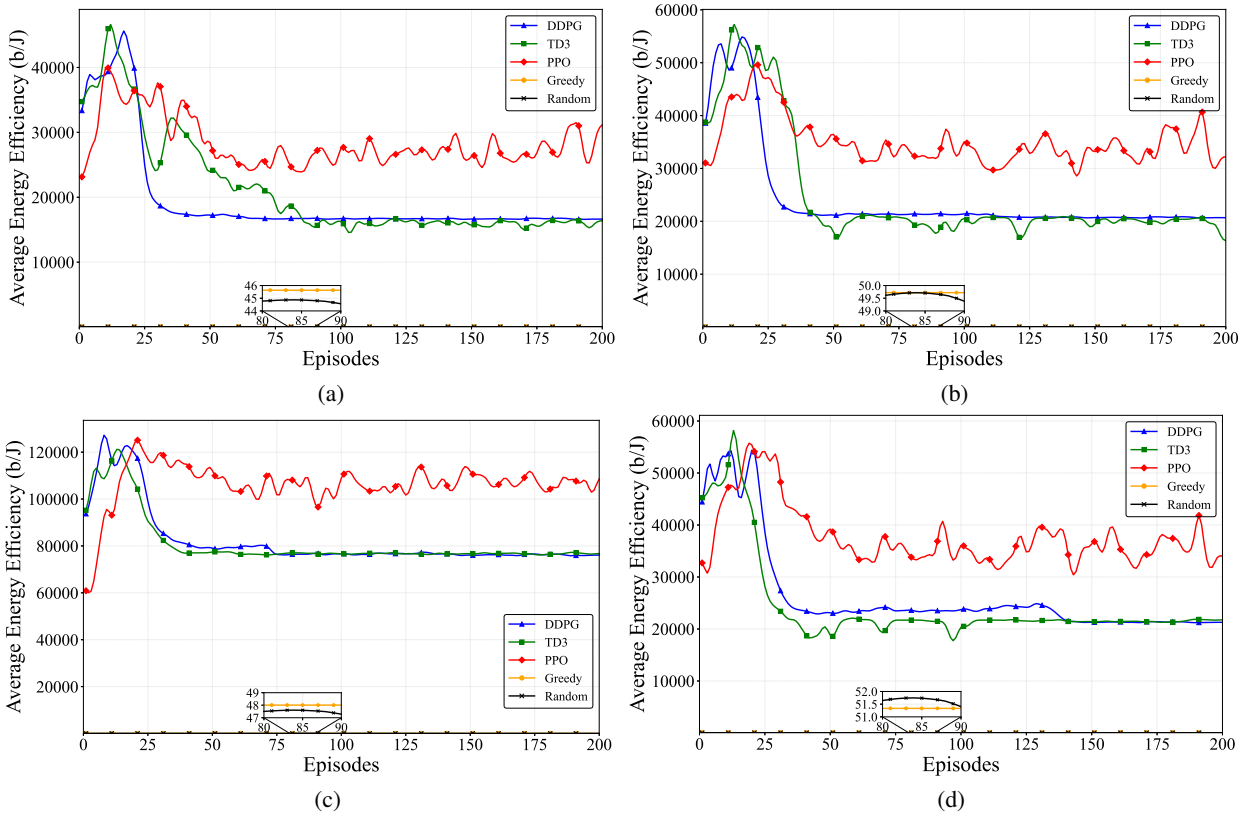


Fig. 2: Comparison of the average EE with single SU in different diversity environments: (a) no diversity, (b) MRC diversity, (c) EGC diversity, and (d) SC diversity.

3) Target Policy Smoothing: TD3 adds small clipped noise to the target action during Q-value updates, reducing the sensitivity of the critic to sharp policy changes.

Similar to DDPG, but with twin critics and delayed actor updates, the per-update complexity is roughly $O(K \cdot P \cdot Q \cdot R^2 + 2 \cdot P \cdot Q \cdot R^2)$, accounting for two critic networks per update. These innovations make TD3 particularly robust in environments with function approximation errors and continuous, high-dimensional action spaces, such as in joint energy and spectrum management scenarios.

3) PPO: The PPO is a state-of-the-art policy gradient method designed to improve training stability and efficiency over conventional policy optimization algorithms. It is applicable to both discrete and continuous action spaces, and it is especially valued for its robustness across a variety of environments [39].

PPO introduces a clipped surrogate objective function, which restricts the magnitude of policy updates to remain within a "trust region." This ensures that each policy update does not deviate too far from the previous policy, thereby avoiding performance collapse during training. Unlike methods requiring complex second-order derivatives or constrained optimization, PPO performs multiple epochs of stochastic gradient ascent on mini-batches of collected data using a first-order optimization approach. For PPO, each agent's actor update scales as $O(K \cdot P \cdot Q \cdot R^2)$ per mini-batch, while the centralized critic evaluation scales as $O(P \cdot Q \cdot R^2)$, giving a total per-update complexity of $O(K \cdot P \cdot Q \cdot R^2 + P \cdot Q \cdot R^2)$. Its effectiveness lies in striking a balance between exploration and exploitation, while its sample efficiency and ease of implementation make it a reliable choice in dynamic wireless networks with changing energy and QoS demands.

4) SAC: SAC is a recent advancement in off-policy DRL that extends the actor-critic architecture with an entropy-regularized objective. The key innovation in SAC lies in its goal to not only maximize the expected return but also the entropy of the policy. This promotes exploration by encouraging stochastic policies that assign non-zero probabilities to multiple actions, making SAC particularly effective in uncertain and dynamic environments [40].

SAC maintains two Q-functions (similar to TD3) to mitigate overestimation bias and employs a stochastic actor, typically parameterized as a Gaussian policy. During training, SAC optimizes a soft Q-function that incorporates an entropy term weighted by a temperature parameter α. This temperature can either be fixed or adjusted dynamically to balance exploitation and exploration. SAC's entropy-aware optimization leads to superior convergence and robustness, especially in continuous control tasks where exploration is crucial. The per-update complexity of SAC is $O(K \cdot P \cdot Q \cdot R^2 + 2 \cdot P \cdot Q \cdot R^2)$, similar to TD3 due to the use of twin critics and a stochastic actor, with the additional cost from the entropy term being negligible. Given the inherent uncertainties and variability in EH-enabled CR environments, SAC provides a compelling solution for policy learning under partial observability and noise.

B. Non-DRL Algorithms

1) Greedy Method: The greedy method represents a baseline, heuristic approach for decision-making in the considered

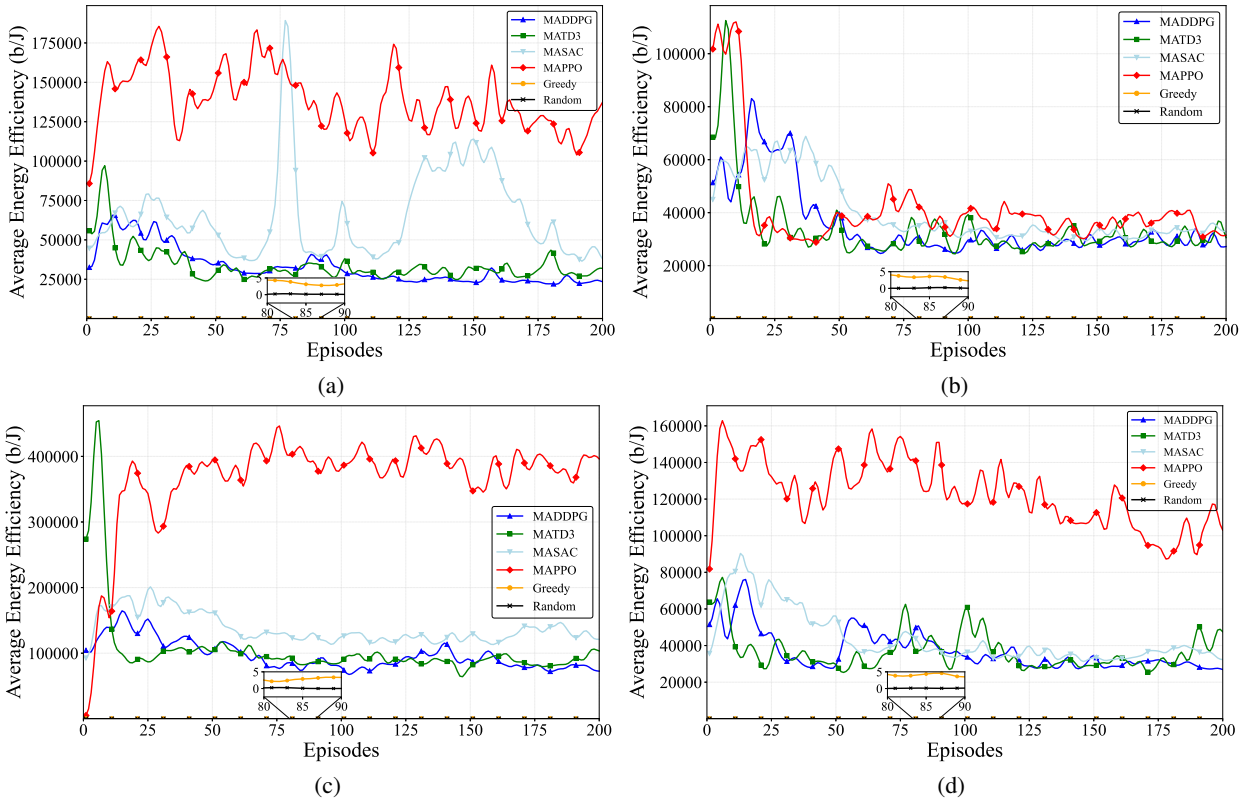


Fig. 3: Comparison of the average EE with multiple SUs in different diversity environments: (a) no diversity, (b) MRC diversity, (c) EGC diversity, and (d) SC diversity.

EH-enabled CR-NOMA system. This method does not rely on any form of learning or long-term planning; instead, it makes instantaneous decisions aimed at maximizing short-term performance. Specifically, in the context of transmission scheduling, the SU transmits with maximum allowable power $\Omega_{\max}$, thereby utilizing the entire available energy before initiating any EH process.

The energy-time allocation parameter ς_n is computed as

$$\varsigma_n = \min\left\{1, \frac{E_n}{T(\tilde{\Omega}_n + \Omega_n)}\right\},$$

where E_n represents the current energy level, T is the duration of the time slot, and $\tilde{\Omega}_k + \Omega_n$ accounts for total power-related costs. While this method is computationally efficient and easy to implement, it often leads to suboptimal performance in dynamic environments due to its lack of foresight and adaptivity.

2) Random Method: The random method is another non-learning baseline approach that introduces stochasticity into the decision-making process without relying on environmental feedback or optimization objectives. In this method, the SU again transmits at a fixed maximum power level $\Omega_{\max}$. However, unlike the greedy method, the value of ς_n is chosen randomly from a uniform distribution bounded by the feasible range

$$\varsigma_n \sim \mathcal{U}\left(0, \min\left\{1, \frac{E_n}{T(\tilde{\Omega}_n + \Omega_n)}\right\}\right).$$

This randomness allows the algorithm to explore a variety of actions but lacks any performance-driven guidance. As such, it may occasionally perform well by chance, but it does not exploit structure in the environment to improve decisions over time. The random method primarily serves as a baseline for comparison with more intelligent, learning-based algorithms.

C. Modeling EE Optimization with DRL

This section provides a generalized framework for modeling the EE maximization problem introduced in Section III using deep reinforcement learning. We define the corresponding state space, action space, and reward function for the considered setup.

1) State Space: The state space for the optimization problem is defined as

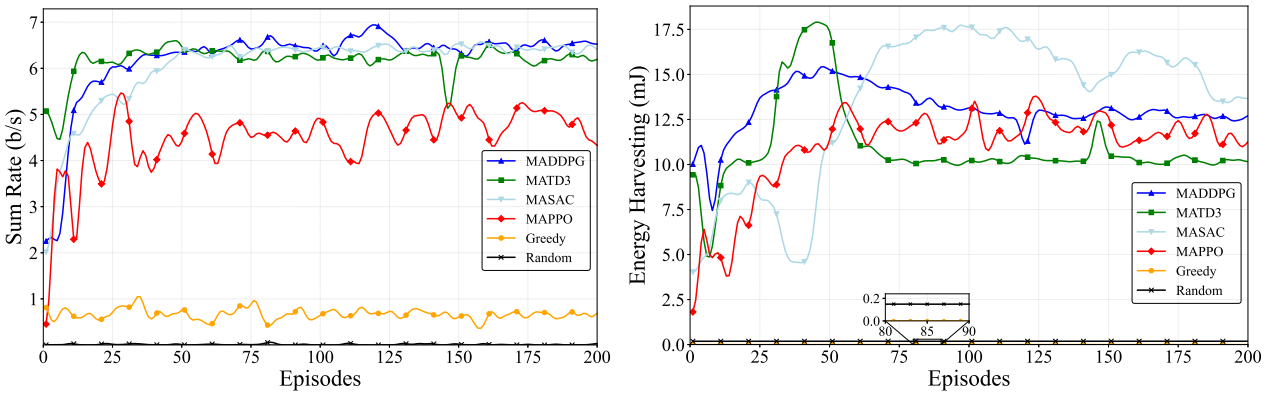
$$s_k = \left[\mathcal{E}_k,\ |\tilde{h}_k|^2,\ |h_{j,k}|^2\right]^T, \tag{17}$$

which contains the SU's battery level, its gain with the BS, and the currently selected j-th PD.

2) Action Space: The action parameter is defined as $\bar{\mathcal{E}}_k$ for any particular time slot n. The normalized ranges for the action parameter in the case of EGC, MRC, and SC are given by

$$\begin{aligned}\bar{\mathcal{E}}_{k_E^*} = \pi \min\Bigg\{\mathcal{E}_{\max} - \mathcal{E}_k, \frac{\tau\eta d^{-\gamma}\Omega_{j,n}}{Q}\left(\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|\right)^2 - \theta_{\text{egc}} \\ - \tilde{\theta}_{\text{egc}}\Bigg\} - (1-\pi)\min\Big\{\mathcal{E}_k, \tau(\tilde{\Omega}_{k,n} + \rho_k)\Big\},\end{aligned} \tag{18}$$

$$\begin{aligned}\bar{\mathcal{E}}_{k_M^*} = \pi \min\Bigg\{\mathcal{E}_{\max} - \mathcal{E}_k, \tau\eta d^{-\gamma}\Omega_{j,n}\sum_{q=1}^{Q}|\tilde{h}_{j,k,q}|^2 - \theta_{\text{mrc}} \\ - \tilde{\theta}_{\text{mrc}}\Bigg\} - (1-\pi)\min\Big\{\mathcal{E}_k, \tau(\tilde{\Omega}_{k,n} + \rho_k)\Big\},\end{aligned} \tag{19}$$



(a) (b)

Fig. 4: Additional performance metrics under the EGC environment: (a) achievable sum-rate, (b) harvested energy.

TABLE II: Simulation parameters.

Parameter Description	Symbol	Value
Path Loss Exponent	γ	3
Noise Power Spectral Density	$\mathcal{N}_0$	-170 dBm/Hz
Bandwidth	B	1 MHz
Circuit Power Consumption	ρ_k	1 μW
Exploration Factor	α	0.1
Battery Capacity	$\mathcal{E}_{\max}$	0.3 J
EH Efficiency	η	0.7
Timeslot Duration	τ	1 s

TABLE III: Summary of optimal hyperparameters.

Parameter	DDPG	PPO	SAC	TD3
Learning rate (Actor) λ_a	0.00018	0.00647	0.00889	0.00013
Learning rate (Critic) λ_c	0.00318	0.00546	0.00228	0.0088
Discount factor ∂	0.94	0.92	0.86	0.85
Soft update factor ϖ	0.00478	–	0.0128	0.0356
Batch size P	128	–	96	128
Buffer size $\mathcal{P}$	16,000	2,000	5,000	13,000

$$\bar{\mathcal{E}}_{k_{\substack{*\\S}}} = \pi \min\left\{\mathcal{E}_{\max} - \mathcal{E}_k, \max(\pi\eta d^{-\gamma}\Omega_{j,n}|\tilde{h}_{j,k,q}|^2) - \tilde{\theta}_{\text{sc}}\right\} - (1-\pi)\min\left\{\mathcal{E}_k, \tau(\tilde{\Omega}_{k,n} + \rho_k)\right\}, \tag{20}$$

respectively, where $\pi \in [0, 1]$.

3) Reward Parameter: The reward function is defined by the data rate achieved by the SU at the n-th time, which is given by

$$\tilde{R}_k\left(\bar{\varsigma}_k(\bar{\mathcal{E}}_k), \tilde{\Omega}_k(\bar{\mathcal{E}}_k)\right) = \bar{\varsigma}_k(\bar{\mathcal{E}}_k)\log_2\left(1 + \frac{\tilde{\Omega}_{k,n}(\bar{\mathcal{E}}_k)|\tilde{h}_{k,n}|^2}{1 + \Omega_{j,n}|h_{j,k}|^2}\right). \tag{21}$$

D. Simulation and Training Setup

1) Simulation Parameters: The simulation environment follows standard assumptions for wireless-powered communication networks. Table II summarizes the key physical layer and system parameters.

2) Hyperparameters: Table III reports the best values of common hyperparameters for all the DRL algorithms determined using Optuna. Algorithm-specific settings are mentioned separately. For PPO, additional hyperparameters included a clipping factor ($\epsilon_{\text{ppo}} = 0.5$), number of epochs per update ($W_{\text{ppo}} = 3$), and an initial action standard deviation ($SD_{\text{ppo}} = 0.2$). For TD3, noise-related hyperparameters were employed, namely policy noise ($\kappa_{\text{td3}} = 0.1$), noise clipping ($\epsilon_{\text{td3}} = 0.5$), and policy delay ($\mathcal{D}_{\text{td3}} = 5$).

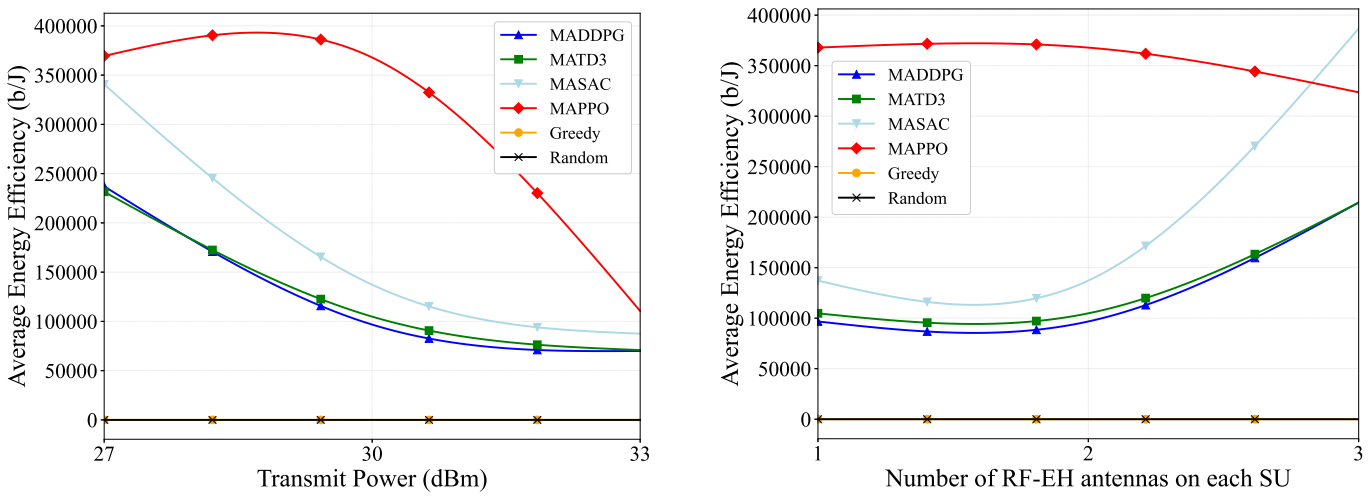
3) Actor and Critic Network Architectures: All algorithms employed feedforward neural networks for actor and critic functions. Each actor network consisted of two hidden layers with 64 units each and ReLU activation, followed by a Tanh output layer to ensure bounded actions in $[-1, 1]$. Both actor and critic networks used the Adam optimizer with algorithm-specific learning rates. For exploration, an initial Gaussian action variance was applied, which decayed over time. The centralized critic took as input the concatenated global state and joint actions of all agents, passing through two hidden layers of 64 ReLU units before outputting a scalar value estimate.

V. Simulation Results

. In this section, we evaluate the performance of the SU across different diversity combining techniques for EH while employing a combination of DRL and non-DRL algorithms to optimize each SU's data-rate in return the energy efficieny under dynamic network conditions. Specifically, we analyze the performance of DRL algorithms such as DDPG, TD3, PPO, and SAC, alongside non-DRL approaches including random schemes and greedy schemes, focusing on single-agent and multi-agent systems. The simulations are conducted under a Rayleigh fading channel model, and we adopt the path loss model presented in [44].

A. EE Analysis of a Single SU under different Diversity Combining Environments

In Fig. 2, We present the convergence plots for EE of both DRL and non-DRL algorithms in the simple environment and across all three RF-EH diversity-combining techniques and the simple environment. Across all sub-figures, PPO consistently achieves superior EE performance compared to other algorithms. It demonstrates rapid convergence, stabilizing around 50 episodes, regardless of the diversity combined environment it is trained in. In contrast, TD3 and DDPG also exhibit a similar convergence trend but achieve lower EE levels as compared to PPO because these algorithms have a relatively slower



(a) (b)

Fig. 5: Comparison of the average EE for SUs under EGC diversity with varying : (a) PD transmit power and (b) RF antennas.

convergence primarily during the initial convergence phase making them less sample efficient, and they struggle more with a delayed reward structure, which is the case in our scenario as we are trying to maximize our data-rate gradually. As expected, the non-DRL baselines, Greedy and Random, consistently yield the lowest EE outcomes across all diversity scenarios.

When comparing the diversity combining environments, PPO achieves the highest EE values within the EGC environment, outperforming both MRC and SC. Although MRC attains a higher data rate due to its weighted signal combining approach, EE considers the ratio between the achieved data rate and the power consumed. While MRC's maximization of SNR enhances data rate, the associated increase in circuit and signal processing power significantly reduces the resulting EE. In contrast, EGC relies on simple summation without complex weighting, leading to lower power consumption and a more favorable balance between data rate and energy usage. SC, on the other hand, selects only the signal with the highest SNR, offering limited diversity and adaptability, which results in comparatively lower EE.

B. EE Analysis of multiple SUs under different Diversity Combining Environments

Fig. 3 illustrates the performance of the implemented DRL and non-DRL algorithms in terms of EE under diverse RF-EH diversity techniques in a multi-agent setting with two SUs. Similar to the single-agent setting, MAPPO consistently demonstrates superior performance in terms of average EE across all diversity environments. This strong performance can be attributed to the ability of the algorithm to balance policy stability and exploration flexibility in non-stationary environments via a centralized training and decentralized execution approach. MADDPG uses deterministic policies, which are less suited for exploration in the continuous action spaces thus exhibiting poor results. However, MATD3 and MASAC are relatively stable due to the use of twin Q-networks and entropy-based exploration, respectively, but still fall short of MAPPO in overall performance.

Among the diversity combining techniques, EGC outperforms the others due to its simpler signal combining mechanism compared to the more complex approach used in MRC. SC performs the worst, primarily due to its limited flexibility and inability to fully exploit signal diversity. Additionally, it is important to note that the convergence plots in the multi-agent setting can be seen to have a higher variance and slower stabilization compared to the single-agent case due to concurrently learning agents. Each agent's policy updates influence the reward landscape for others, leading to oscillations and delayed convergence due to the centralized training, unlike the single-agent case, where the policy updates are limited to one SU only.

As further illustrated in Fig. 4, the EGC environment reveals additional insights into the performance behavior of the evaluated algorithms. MADDPG, MATD3, and MASAC exhibit comparable and consistently higher sum-rate performance than the MAPPO, Greedy, and Random baselines, reflecting the effectiveness of continuous control and value-based learning in optimizing throughput. Meanwhile, MASAC achieves the highest EH, attributed to its entropy-regularized policy that enhances exploration and enables more adaptive power allocation. These results reinforce the observed trade-off between throughput maximization and EE, emphasizing that different DRL algorithms inherently prioritize one objective over the other based on their exploration–exploitation balance.

C. Varying Transmit Power of the PU and Number of RF Receiving Antennas

Fig. 5a investigates the impact of transmit power on the average EE. The results, evaluated at transmit power levels of 27 dBm, 30 dBm, and 33 dBm, reveal an inverse relationship between transmit power and EE for the DRL algorithms. This trend arises because while increased transmit power can potentially lead to higher data rates and thus more bits transmitted

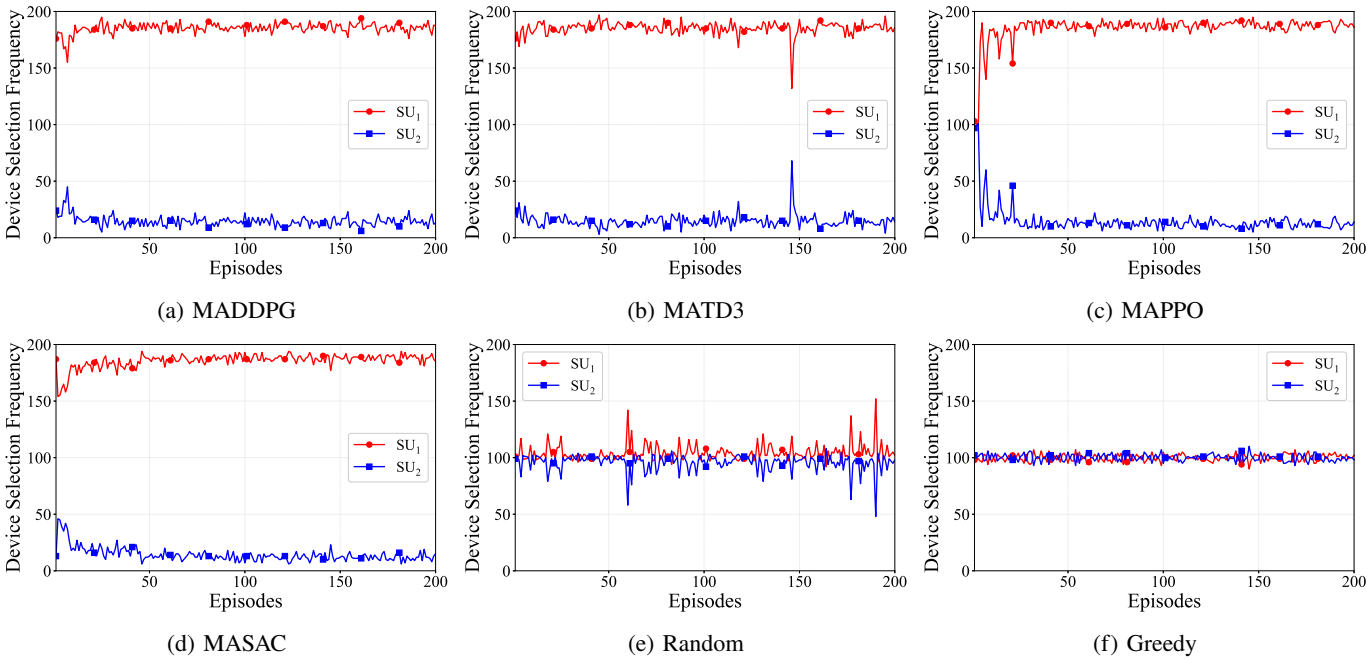


Fig. 6: Comparison of fairness values while training across all algorithms under EGC diversity.

per unit of time, the energy expenditure grows quadratically with the increase in power. Furthermore, the diminishing rate of decrease in EE at higher transmit power levels suggests that the marginal gain in data rate from increasing power eventually becomes less significant compared to the added energy cost, indicating a potential trade-off and a possible operating regime with better energy utilization at lower power levels.

Fig. 5b illustrates the impact of the number of RF-EH antennas on the average EE of each SU. Overall, the DRL-based algorithms benefit from more antennas, as additional RF-EH capability enables greater EH from ambient RF signals, which enhances the ratio of transmitted bits to consumed energy. Among the algorithms, MAPPO consistently achieves the highest EE across all configurations, maintaining a clear performance lead over the other approaches. MASAC and MATD3 deliver comparable results, both surpassing MADDPG, which attains the lowest EE among the MARL agents but still improves with additional antennas. The greedy algorithm exhibits only minor gains and remains significantly below the DRL-based methods, while the random strategy produces negligible EE, unaffected by antenna count. These findings emphasize that as harvesting opportunities increase with more antennas, the ability of DRL algorithms to effectively manage and exploit the harvested energy becomes increasingly critical.

D. Fairness among SUs

It is important to note that fairness was not explicitly considered as an optimization metric in our formulation. Instead, SU selection was driven by the criterion described in section II-C, which jointly accounts for the instantaneous channel gains and battery levels of users. Under this setup, SU_2 was positioned at a farther distance from the base station, leading to consistently weaker channel conditions compared to SU_1. As a result, SU_2

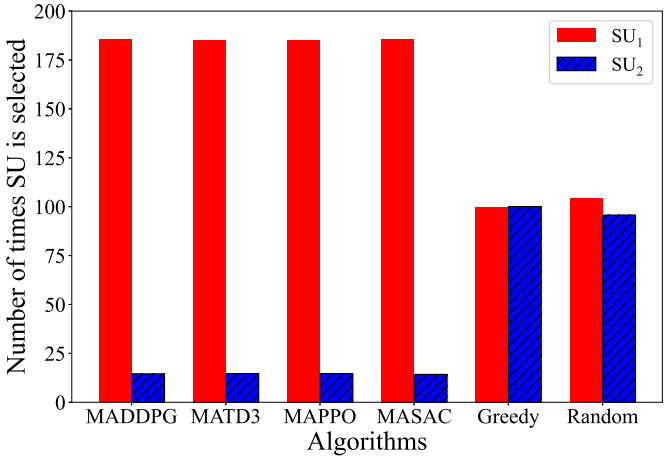


Fig. 7: Averaged fairness values for all SUs under EGC diversity.

experienced a significantly lower probability of being selected for transmission across most learning-based algorithms. This trend is clearly visible in Fig. 6, where MADDPG, MATD3, MAPPO, and MASAC overwhelmingly favor SU_1, while SU_2's selection frequency remains very low. On the other hand, the random and greedy baselines inherently balance the selection process, resulting in nearly equal counts for both users.

E. Performance Scaling with Additional SUs

While the presented results focus on two SUs for clarity, this framework can be extended to scenarios with more users. With an increasing number of SUs, competition for spectrum and EH opportunities intensifies, which may lead to reduced per-user performance but can still improve overall system throughput and EE due to multi-user diversity. However, as the number of agents

grows, training convergence typically slows, and coordination challenges become more pronounced.

VI. Conclusion

In this work, we presented a practical and energy-efficient CR-NOMA-based IoT framework tailored for the sustainable operation of multiple low-power SUs in IUA networks. Recognizing the dual constraints of energy scarcity and spectrum congestion, we formulated a joint transmit power control and RF-EH scheduling problem aimed at maximizing the long-term EE of SUs under strict QoS constraints. We employed a multi-agent DRL framework wherein each SU independently learns optimal strategies, facilitating scalable and distributed control without requiring centralized coordination. To further enhance system robustness and reception performance under realistic RF conditions, we incorporated RF-EH diversity combining techniques—EGC, MRC, and SC—while explicitly accounting for the additional power costs of RF circuitry and signal processing. A suite of advanced DRL algorithms (DDPG, PPO, SAC, and TD3), alongside baseline non-DRL techniques (random and greedy), was benchmarked to identify effective learning models under dynamic wireless conditions. Simulation results demonstrated the superiority of the proposed framework in achieving significant improvements in EE and sum-rate performance. In the future, this study could be extended to account for mobility-aware scenarios, where both SUs and PUs may move dynamically, influencing EH opportunities and interference patterns. Additionally, integrating intelligent reflecting surfaces (IRS) into the RF-EH and CR-NOMA model could offer new possibilities for improving link reliability and energy transfer efficiency. Moreover, while this work employed a simplified linear RF-EH model for tractability, future research can adopt more realistic nonlinear circuit models to better capture saturation and threshold effects, thereby aligning simulations more closely with practical RF-EH behavior. Finally, recent studies have also investigated hybrid strategies, including parameter sharing and partial cooperation among agents, to improve training stability. Extending our framework with such approaches constitutes another promising avenue for future research.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE communications surveys & tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] H. Yu, Z. Shen, and C. Leung, "From internet of things to internet of agents," in *2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing*. IEEE, 2013, pp. 1054–1057.

[3] "Cisco annual internet report," in *2018–2023 White Paper*. CA, USA, Mar. 2024.

[4] S. Aslam, W. Ejaz, and M. Ibnkahla, "Energy and spectral efficient cognitive radio sensor networks for internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3220–3233, 2018.

[5] L. Bariah, S. Muhaidat, and A. Al-Dweik, "Error performance of NOMA-based cognitive radio networks with partial relay selection and interference power constraints," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 765–777, 2019.

[6] S. Asad Ullah, S. Zeb, A. Mahmood, S. A. Hassan, and M. Gidlund, "Opportunistic CR-NOMA transmissions for zero-energy devices: A DRL-driven optimization strategy," *IEEE Wireless Communications Letters*, vol. 12, no. 5, pp. 893–897, 2023.

[7] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.

[8] L. Lv, J. Chen, Q. Ni, Z. Ding, and H. Jiang, "Cognitive non-orthogonal multiple access with cooperative relaying: A new wireless frontier for 5G spectrum sharing," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 188–195, 2018.

[9] N. Ansari and T. Han, *Green Mobile Networks: A Networking Perspective*. John Wiley & Sons, 2017.

[10] X. Zhang, J. Grajal, M. López-Vallejo, E. McVay, and T. Palacios, "Opportunities and challenges of ambient radio-frequency energy harvesting," *Joule*, vol. 4, no. 6, pp. 1148–1152, 2020.

[11] U. Muncuk, K. Alemdar, J. D. Sarode, and K. R. Chowdhury, "Multiband ambient RF energy harvesting circuit design for enabling batteryless sensors and IoT," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2700–2714, 2018.

[12] D. Elsheakh, "Microwave antennas for energy harvesting applications," *Microwave Systems and Applications*, 2017.

[13] F. S. Mohd Noor, Z. Zakaria, H. Lago, and M. A. Meor Said, "Dual-band aperture-coupled rectenna for radio frequency energy harvesting," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 29, no. 1, p. e21651, 2019.

[14] S. Palanisamy, B. Thangaraju, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, and F. Alassery, "A novel approach of design and analysis of a hexagonal fractal antenna array (HFAA) for next-generation wireless communication," *Energies*, vol. 14, no. 19, p. 6204, 2021.

[15] B. L. Pham and A.-V. Pham, "Triple bands antenna and high efficiency rectifier design for RF energy harvesting at 900, 1900 and 2400 MHz," in *2013 IEEE MTT-S International Microwave Symposium Digest (MTT)*. IEEE, 2013, pp. 1–3.

[16] K. Rapetswa and L. Cheng, "Convergence of mobile broadband and broadcast services: A cognitive radio sensing and sharing perspective," *Intelligent and Converged Networks*, vol. 1, no. 1, pp. 99–114, 2020.

[17] Z. Zhou, C. Zhang, J. Wang, B. Gu, S. Mumtaz, J. Rodriguez, and X. Zhao, "Energy-efficient resource allocation for energy harvesting-based cognitive machine-to-machine communications," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 595–607, 2019.

[18] D. Zhang, Z. Chen, M. K. Awad, N. Zhang, H. Zhou, and X. S. Shen, "Utility-optimal resource management and allocation algorithm for energy harvesting cognitive radio sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3552–3565, 2016.

[19] T.-V. Nguyen, V.-D. Nguyen, D. B. da Costa, and B. An, "Hybrid user pairing for spectral and energy efficiencies in multiuser MISO-NOMA networks with SWIPT," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4874–4890, 2020.

[20] J. Tang, J. Luo, M. Liu, D. K. So, E. Alsusa, G. Chen, K.-K. Wong, and J. A. Chambers, "Energy efficiency optimization for NOMA with SWIPT," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 452–466, 2019.

[21] F. Li, H. Jiang, R. Fan, and P. Tan, "Cognitive non-orthogonal multiple access with energy harvesting: An optimal resource allocation approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7080–7095, 2019.

[22] X. Wang, Z. Na, K.-Y. Lam, X. Liu, Z. Gao, F. Li, and L. Wang, "Energy efficiency optimization for NOMA-based cognitive radio with energy harvesting," *IEEE access*, vol. 7, pp. 139 172–139 180, 2019.

[23] H. Azarhava and J. M. Niya, "Energy efficient resource allocation in wireless energy harvesting sensor networks," *IEEE Wireless Communications Letters*, vol. 9, no. 7, pp. 1000–1003, 2020.

[24] S. Zhang, N. Zhang, S. Zhou, J. Gong, Z. Niu, and X. Shen, "Energy-aware traffic offloading for green heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1116–1129, 2016.

[25] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2015.

[26] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[27] P. Blasco, D. Gunduz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1872–1882, 2013.

[28] R. Wang, A. Yadav, E. A. Makled, O. A. Dobre, R. Zhao, and P. K. Varshney, "Optimal power allocation for full-duplex underwater relay

networks with energy harvesting: A reinforcement learning approach," *IEEE wireless communications letters*, vol. 9, no. 2, pp. 223–227, 2019.
[29] M. Chu, X. Liao, H. Li, and S. Cui, "Power control in energy harvesting multiple access system with reinforcement learning," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 9175–9186, 2019.
[30] Y. Zhao, J. Hu, K. Yang, and S. Cui, "Deep reinforcement learning aided intelligent access control in energy harvesting based WLAN," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 14 078–14 082, 2020.
[31] Z. Ding, R. Schober, and H. V. Poor, "No-pain no-gain: DRL assisted optimization in energy-constrained CR-NOMA networks," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5917–5932, 2021.
[32] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8577–8588, 2019.
[33] S. Asad Ullah, S. Zeb, A. Mahmood, S. A. Hassan, and M. Gidlund, "Opportunistic CR-NOMA transmissions for zero-energy devices: A DRL-driven optimization strategy," vol. 12, no. 5, 2023, pp. 893–897.
[34] M. Min, L. Xiao, Y. Chen, P. Cheng, D. Wu, and W. Zhuang, "Learning-based computation offloading for IoT devices with energy harvesting," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1930–1941, 2019.
[35] S. A. Ullah, S. Zeb, A. Mahmood, S. A. Hassan, and M. Gidlund, "Deep RL-assisted energy harvesting in CR-NOMA communications for next-G IoT networks," in *IEEE Globecom Wkshps*. IEEE, 2022, pp. 74–79.
[36] Z. Shi, X. Xie, H. Lu, H. Yang, J. Cai, and Z. Ding, "Deep reinforcement learning-based multidimensional resource management for energy harvesting cognitive NOMA communications," *IEEE Transactions on Communications*, vol. 70, no. 5, pp. 3110–3125, 2022.
[37] D. Altinel and G. K. Kurt, "Diversity combining for RF energy harvesting," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*. IEEE, 2017, pp. 1–5.
[38] D. Silver, G. Lever, N. Heess *et al.*, "Deterministic policy gradient algorithms," in *ICML*, 2014, pp. 387–395.
[39] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347
[40] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.
[41] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.
[42] A. Shahini, A. Kiani, and N. Ansari, "Energy efficient resource allocation in EH-enabled CR networks for IoT," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3186–3193, 2018.
[43] S. Asad Ullah, M. Abdullah Sohail, H. Jung, M. Omer Bin Saeed, and S. Ali Hassan, "Sum Rate Maximization in IoT Networks With Diversity-Enhanced Energy Harvesting: A DRL-Guided Approach," *IEEE Internet of Things Journal*, vol. 11, no. 18, pp. 30 309–30 322, 2024.
[44] S. Seidel and T. Rappaport, "914 MHz Path loss Prediction Models for Indoor Wireless Communications in Multifloored Buildings," *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 2, pp. 207–217, 1992.