Received 22 July 2024; accepted 16 August 2024. Date of publication 21 August 2024; date of current version 4 September 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3447152

DRL-Driven Optimization of a Wireless Powered Symbiotic Radio With Nonlinear EH Model

SYED ASAD ULLAH [1] (Student Member, IEEE), AAMIR MAHMOOD [2] (Senior Member, IEEE), ALI ARSHAD NASIR [3,4] (Member, IEEE), MIKAEL GIDLUND [2] (Senior Member, IEEE), AND SYED ALI HASSAN [1] (Senior Member, IEEE)

[1]School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad 44000, Pakistan

[2]Department of Computer and Electrical Engineering, Mid Sweden University, 85170 Sundsvall, Sweden

[3]Department of Electrical Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

[4]Center for Communication Systems and Sensing, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

CORRESPONDING AUTHOR: S. A. ULLAH (e-mail: sullah.phdee21seecs@seecs.edu.pk)

This work was supported in part by the Knowledge Foundation Sweden (KKS) Research Profile Next Generation Industrial IoT (NIIT).

ABSTRACT Given the rising demand for low-power sensing, integrating additional devices into an existing wireless infrastructure calls for innovative energy- and spectrum-efficient wireless connectivity strategies. In this respect, wireless-powered or energy-harvesting symbiotic radio (EHSR) is gaining attention for establishing the secondary relationship with the primary wireless systems in terms of RF EH and opportunistically sharing the spectrum or schedule. In this paper, assuming the commensalistic relationship with the primary system, we consider the energy-efficient optimization of such an EHSR by intelligently making EH and transmission decisions under the inherent nonlinearity of the EH circuitry and dynamics of pre-scheduled primary devices. We present a state-of-the-art deep reinforcement learning (DRL)-engineered, energy-efficient transmission strategy, which intelligently orchestrates EHSR's uplink transmissions, leveraging the cognitive radio-inspired non-orthogonal multiple access (CR-NOMA) scheme. We first formulate the energy efficiency (EE) optimization metric for EHSR considering the nonlinear EH model, and then we decompose the inherently complex, non-convex problem into two optimization layers. The strategy first derives the optimal transmit power and time-sharing coefficient parameters, using convex optimization. Subsequently, these inferred parameters are substituted in the subsequent layer, where the optimization problem with continuous action space is addressed via a DRL framework, named modified deep deterministic policy gradient (MDDPG). Simulation results reveal that, compared to the baseline DDPG algorithm, our proposed solution provides a 6% EE gain with the linear EH model and approximately a 7% EE gain with the non-linear EH model.

INDEX TERMS Symbiotic radio, RF EH, cognitive radio-inspired non-orthogonal multiple access (CR-NOMA), energy efficiency (EE), deep deterministic policy gradient (DDPG).

I. INTRODUCTION

With the escalating requirements for the meticulous sensing and monitoring of various physical environments, modern wireless networks are anticipated to cater not only to the standard or scheduled communication devices but also to facilitate opportunistic transmissions using symbiotic radios (SRs) [1]. For instance, SRs can serve as additional sensing nodes installed in remote, inaccessible, or perilous zones within an industrial setting. Still, given the impracticality of regular battery replacements in SRs due to cost and safety concerns, adopting zero-energy radios is vital. In this respect, radio-frequency (RF) energy harvesting (EH) is gaining momentum in modern wireless networks for its potential to create self-sustaining communication systems, particularly in the Internet-of-things (IoT) paradigm [2], [3] and 6G communications [4]. EH-capable systems reduce dependency on traditional power sources and increase device lifespans by converting RF signals from the environment into

usable electrical energy, thus promoting green communication [5], [6]. As a result, wireless-powered or EH symbiotic radios (EHSR) are gaining attention for establishing the secondary relationship with the primary wireless systems in terms of RF-EH and opportunistically sharing the spectrum or schedule. Hence, to effectively accommodate EHSRs in future IoT networks, it is imperative to formulate energy- and spectral-efficient communication strategies.

In practical scenarios, the conversion efficiency of the RF-DC circuit varies depending on the input power level. At lower input power levels, the efficiency typically tends to increase as input power increases. However, this increase in efficiency encounters diminishing returns and has upper bounds on the energy that can be harvested [7]. Traditional linear EH models fail to encapsulate this non-linear behavior of the RF-DC circuit. For certain input powers, the linear model overestimates the conversion efficiency compared to the actual RF-DC circuit's performance. Hence, the non-linearity introduced by RF circuits is equally significant, as many existing RF-EH models fail to accurately represent real-world complexities, leading to sub-optimal practical performance [8], [9].

The cognitive radio-inspired non-orthogonal multiple access (CR-NOMA) framework stands out as a revolutionary approach in wireless communication promising spectral efficiency, and capacity-enhancing solution. The fundamental principle of CR-NOMA is to serve the user n while ensuring the quality of service (QoS) of user m. In particular, the signal-to-interference-noise ratio (SINR) of the user m limits the transmit power allocated to the user n [10], [11]. Nevertheless, given the dynamic network conditions, i.e., channel conditions and energy supply conditions, it is a challenging task to implement the transmission plan for EHSRs. Fortunately, intelligent algorithms driven by deep reinforcement learning (DRL) can play a key role in EHSRs' capacity for intelligent decision-making, facilitating the creation of wireless networks of heterogeneous (industrial) IoT devices [12]. For example, supervised learning aids link adaptation and channel estimation in orthogonal frequency-division multiplexing (OFDM) systems [13], [14]. Unsupervised learning enhances wireless location accuracy [5] and reduces beamformer design complexity [6]. Reinforcement learning's applications encompass resource allocation, offloading data and computation, EH, and network security [15], [16]. Building on this, it is anticipated that next-generation wireless networks will be fueled in the future by machine learning, which stands out as a major enabling technique [17]. Its core principle is the use of data-driven learning to enhance decision-making [18], [19]. Due to its adaptability, it has been used to address a variety of wireless communication issues.

Given the aforementioned complexities and the challenges presented by the non-linearity of energy conversion in RF-EH circuits, this paper considers the application of DRL, specifically the combined experience replay deep deterministic policy gradient (CER-DDPG) algorithm, which we refer to as the modified DDPG (MDDPG). In MDDPG the latest tuple of experiences is assured to be incorporated, when selecting a random batch of experiences for training the agent [20], to optimize the energy efficiency (EE) of the EHSR. DRL, with its capability to navigate complex, dynamic environments and make intelligent decisions, is particularly tailored for this challenge. The proposed method divides the problem into two layers; first, it derives the optimal solutions for the optimal parameters using convex optimization, and then uses MDDPG to handle the one-dimensional continuous space optimization problem. This hierarchical approach, supported by DRL, ensures a more efficient sampling process and a quicker path to optimal solutions. Thus, the application of DRL not only addresses the inherent challenges of the research problem but also offers a robust pathway toward practical and effective real-world solutions.

A. PRIOR WORK

Wireless-powered communications and NOMA have been extensively researched for their potential to enhance EE and spectrum utilization in modern wireless networks. The study in [21] addressed the optimization of simultaneous wireless information and power transfer (SWIPT) under a non-linear EH model, but its reliance on semidefinite relaxation and variable substitutions limits its scalability and efficiency, particularly in dynamically changing environments. The work in [22] presented a resource allocation algorithm for multiuser SWIPT systems, yet its non-convex sum-of-ratios approach and iterative algorithm lack the robustness and comprehensive integration of RF circuit power and CR-NOMA as in our work. The study in [23] explored an intelligent reflecting surface (IRS)-aided wireless powered communication network (WPCN) with NOMA, but it adopted a simplified linear EH model and failed to ensure strict user quality-of-service requirements. Our research addresses these shortcomings by employing a non-linear EH model, integrating CR-NOMA, and utilizing an advanced MDDPG algorithm within a two-layer optimization framework, thereby providing a more robust and practical solution.

Similarly, the work in [24] optimized EH time and power allocation in a WPCN under linear and nonlinear EH models using time-division multiple access (TDMA) or orthogonal frequency-division multiple access (OFDMA). However, its approach is confined to simpler resource allocation schemes and does not address the integration of CR-NOMA or advanced DRL algorithms, thereby, limiting its applicability to complex scenarios. The authors in [25] focused on EE in WPCNs with NOMA and formulated resource allocation problems for half-duplex and asynchronous transmission; however, it lacked a realistic non-linear EH model and advanced DRL algorithms, making it less applicable to more intricate and dynamic environments.

To increase throughput in a peer-to-peer network, the study in [26] combines convex optimization with the DDPG to promote energy-efficient communication options. Using a linear EH model, this method creates optimal EH and power

allocation policies. Similarly, the authors in [27] proposed an optimal transmission policy for an energy-constrained device based on cognitive radio. A DRL-assisted approach is used to solve the long-term throughput maximization problem in a communication network that employs CR-NOMA. Similarly, the work presented in [28], addressed the EE maximization problem for an EH radio functioning amidst scheduled primary devices. By leveraging an RL approach, the authors proposed an optimal transmission policy that enhances spectral efficiency and maximizes the EE of the EH device through CR-NOMA. However, their analysis was predicated on a linear EH model. The authors in [29] addressed an optimization problem to maximize the EE of the EH device using the DDPG algorithm. The authors derived closed-form expressions that account for the impact of RF circuit power, yet the foundational EH model remained linear.

To the best of our knowledge, previous studies in wireless-powered communications and NOMA have limitations. The study in [21] optimizes SWIPT under a non-linear EH model but relies on semidefinite relaxation, limiting scalability. The resource allocation algorithm in [22] lacks robustness and RF circuit power integration. Research in [23] uses a simplified linear EH model, and [24] does not integrate CR-NOMA or advanced DRL algorithms. Studies such as [26] and [27] use linear EH models. Research in [28] and [29] also use linear EH models with RL and DDPG algorithms. Our contributions include the development of a non-linear EH model, accounting for RF circuit power consumption, and the implementation of CR-NOMA, all optimized through an advanced DRL algorithm, termed the MDDPG algorithm. These advancements set our work apart from existing literature and provide a comprehensive solution to the challenges faced in modern wireless communication systems.

In the subsequent section, we outline the key contributions of this work.

B. PAPER CONTRIBUTIONS

Our key contributions are listed as follows.

- We derive an explicit analytical expression for a non-linear EH model, offering practical insights into realistic EH scenarios. This model paves the way for more accurate and practical designs in energy-harvesting communications.
- To find the appropriate combination of transmission and EH parameters of the EHSR, we formulate the EE [30] metric of the EHSR as an optimization problem, integrating the non-linear EH model. Our approach uniquely accounts for RF circuit power and signal processing power consumption, ensuring a holistic and practical optimization framework.
- We decompose the inherently complex and non-convex optimization problem into a two-layer structure. The first layer leverages convex optimization to derive closed-form expressions for transmit power and time-sharing coefficients. These solutions seamlessly integrate into the second layer, where a DRL framework, specifically, the MDDPG algorithm efficiently addresses the one-dimensional continuous action space optimization problem of EE maximization.
- We thoroughly explore various influencing factors, such as the number of primary devices in the network, RF circuit, and signal processing power consumption, and transmit power of primary devices. Our comprehensive comparative analysis highlights the superior performance of our proposed MDDPG algorithm over baseline methods, including DDPG, random, and greedy algorithms, under both linear and non-linear EH models.

C. PAPER ORGANIZATION

The rest of the paper is organized as follows. In Section II, we present the system model, followed by Section III, which articulates the problem formulation in depth. In Section IV, we discuss the core principles of the proposed DRL framework and structure our optimization problem accordingly. Extensive simulation results are presented in Section V, and finally, Section VI concludes the paper.

II. SYSTEM MODEL

We examine a wireless IoT network configuration consisting of a central base station (BS) and J prescheduled primary IoT devices, represented by M_j, where $1 \leq j \leq J$. The primary devices transmit using a time division multiple access (TDMA) mechanism, whereby each device is allocated a designated time slot lasting T seconds within an overarching frame duration of NT, given that $N \geq J$. The scheduling for these devices is as follows: During the k-th time slot, labeled, t_k and satisfying $1 \leq k \leq N$, the j-th primary device is engaged. Here j is determined by the relation $((k-1) \oplus J) + 1$, with $\oplus$ as the modulo operation.

In this setup, an energy-constrained device, termed EHSR, transmits sensor readings to the BS within the time slots reserved for primary devices. EHSR achieves this using the CR-NOMA method. The CR-NOMA methodology permits the transmissions of EHSR to be integrated seamlessly within the primary devices' designated time slots while maintaining their stipulated QoS requirements. The assurance of this QoS is realized through a unique QoS-driven successive interference cancelation (SIC) decoding sequence. This ensures that the signal of the EHSR undergoes decoding in the initial phase of the SIC decoding process [10].

A standout feature of the EHSR is its ability to derive energy from the uplink RF transmissions initiated by the primary devices. Given the k-th time slot as an example, where t_k is equal for all k, the EHSR allocates the first $\gamma_k T$ seconds to data transfer while designating the subsequent $(1-\gamma_k)T$ seconds to energy accumulation, where $\gamma_k \in [0, 1]$, is the time-sharing coefficient. To enhance notation readability, let the primary device scheduled at time t_k be denoted by M_k, wherein $M_k = M_j$, and $j = ((k-1) \oplus J)+1$. Hereafter, the symbol k will denote both the time slot and

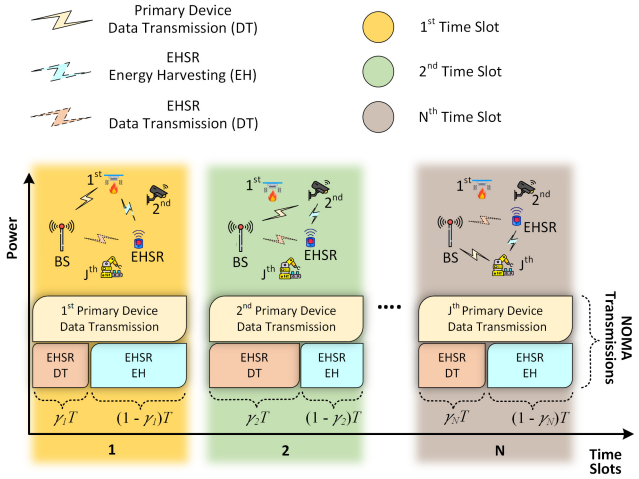

FIGURE 1. Depiction of the system model for EH symbiotic radio (EHSR).

the primary device. The channel gain between EHSR and the BS at the k-th time slot is symbolized by $\tilde{h}_k$. Moreover, for any k-th primary device, its channel gains to the BS and EHSR at the k-th time slot are denoted by h_k, and $h_{k,o}$, respectively. The detailed system model, including the EH and CR-NOMA transmission scheme, is shown in Fig. 1.

We assume that the EHSR has full battery power at the start of communication and that it is aware of the channel condition information [1] of each primary device broadcasting at time t_k before the communication starts. Since we define Γ_k as the amount of energy stored in the battery of EHSR at time t_k, therefore, the total transmission energy of the EHSR is upper bounded on Γ_k, which is expressed as

$$\gamma_k T\left(\tilde{\Psi}_k + \lambda\right) \leq \Gamma_k, \tag{1}$$

where λ denotes the constant RF circuit power and RF signal processing power of the EHSR, which characterizes the fixed power consumption associated with RF circuitry and signal processing tasks and $\tilde{\Psi}_k$ represents the transmit power of EHSR at time t_k. Consequently, the total energy available in the battery of the EHSR at time t_{k+1} is represented as

$$\Gamma_{k+1} = \min\Big\{\Gamma_k + (1-\gamma_k)T[\Phi_{\text{Prac}}(\Psi_k)]|h_{k,o}|^2 - \gamma_k T\left(\tilde{\Psi}_k + \lambda\right), \Gamma_{\max}\Big\}, \tag{2}$$

where $\Phi_{\text{Prac}}(\Psi_k)$ is the practical non-linear EH model of the EHSR, and Ψ_k represents the transmit power of k-th primary device. In Eq. (2), $\Gamma_{\max}$ represents the maximum battery capacity of the EHSR.

[1]The required CSI can be obtained as follows. Before transmission, the k-th primary device, M_k, can broadcast a pilot signal. This signal allows both the BS and the EHSR to simultaneously estimate h_k and $h_{k,0}$, respectively. Through a reliable feedback channel, the BS can communicate its knowledge of h_k to the EHSR. Additionally, the BS broadcasts another pilot signal to enable the EHSR to estimate $\tilde{h}_k$. For applications with low mobility, such as static sensors in the IoT, the system overhead from channel estimation is moderate because pilot signals can be transmitted infrequently.

A. NON-LINEAR EH MODEL

The expression for the practical non-linear EH model for the EHSR is given by [7]

$$\Phi_{\text{Prac}}(\Psi_k) = \frac{\bar{P} - \alpha_3 S}{1 - S}, \tag{3}$$

where,

$$\bar{P} = \frac{\alpha_3}{1 + e^{(-\alpha_1(\Psi_k - \alpha_2))}},$$

and

$$S = \frac{1}{1 + e^{(\alpha_1 \alpha_2)}}.$$

Here, $\bar{P}$ represents a traditional logistic function with respect to the power of the k-th primary device, i.e., input received power, the constant S is used to ensure zero-input/zero-output response for the EH. Constants α_1 and α_2 represent the EH circuit specifications, for instance, capacitance, diode turn-on voltage, and resistance. Whereas, α_3 denotes the maximum harvested energy when the EH circuit is saturated. Practically, the EH circuit of each EHSR is fixed and the constant parameters, i.e., α_1, α_2, α_3, and S can be determined by employing a standard curve fitting tool.

Lemma-1: An explicit expression for (3) can be expressed as:

$$\Phi_{\text{Prac}}(\Psi_k) = \frac{\alpha_3\left[e^{\alpha_1 \Psi_k} - 1\right]}{e^{\alpha_1 \Psi_k} + e^{\alpha_1 \alpha_2}}. \tag{4}$$

Proof: See the Appendix.

In the sequel, we adopt the practical non-linear EH model given in (4), hence, (2) is reformulated as follows

$$\Gamma_{k+1} = \min\Bigg\{\underbrace{(1-\gamma_k)T\left[\frac{\alpha_3\left[e^{\alpha_1 \Psi_k} - 1\right]}{e^{\alpha_1 \Psi_k} + e^{\alpha_1 \alpha_2}}\right]|h_{k,o}|^2}_{\text{Harvested Energy}} - \underbrace{\gamma_k T\left(\tilde{\Psi}_k + \lambda\right)}_{\text{Consumed Energy}} + \Gamma_k, \Gamma_{\max}\Bigg\}, \tag{5}$$

which is always positive under the condition set in (1).

III. PROBLEM FORMULATION

This section provides the mathematical modeling of the EE maximization problem and its formulation into a DRL framework.

We define the data rate achieved by the EHSR at time t_k by

$$\tilde{R}_k = \gamma_k \log_2\left(1 + \frac{\tilde{\Psi}_k |\tilde{h}_k|^2}{1 + \Psi_k |h_k|^2}\right). \tag{6}$$

The order of SIC decoding is specified by (6), where the signal of EHSR is decoded in the first cycle of SIC decoding, and that of the scheduled primary device is decoded in the later phase of the SIC decoding cycle, thereby, ensuring the

QoS requirements of the primary devices. As our goal is to maximize the EE of the EHSR, hence, based on (6), the EE of the EHSR, at k-th time slot, is given by

$$\Pi_k\left(\gamma_k, \tilde{\Psi}_k\right) = \frac{\gamma_k \log_2\left(1 + \frac{\tilde{\Psi}_k|\tilde{h}_k|^2}{1+\Psi_k|h_k|^2}\right)}{\bar{\lambda} + \lambda}, \tag{7}$$

which is the ratio of the instantaneous data rate the EHSR achieves to the total power used by the EHSR. In (7), $\bar{\lambda}$ denotes the total transmit power and reflects the average power available for transmission over a specified period. We can observe that (7) satisfies the QoS of the scheduled primary device by first decoding the signal of the EHSR and decoding the signal of the scheduled primary device, without any interference from the EHSR. In EE expression, the parameters to be optimized are γ_k and , $\tilde{\Psi}_k$. Accordingly, the maximization problem is formulated as

$$\underset{\gamma_k,\ \tilde{\Psi}_k}{\text{maximize}} \quad \mathbb{E}\left\{\sum_{k=1}^{N} \beta^{k-1} \Pi_k\left(\gamma_k, \tilde{\Psi}_k\right)\right\} \tag{P1}$$

$$\text{s.t.} \quad \Gamma_{k+1} = \min\left\{(1-\gamma_k)T\left[\frac{\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]}{e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}}\right]|h_{k,o}|^2 - \gamma_k T\left(\tilde{\Psi}_k + \lambda\right) + \Gamma_k, \Gamma_{\max}\right\}, \tag{P1a}$$

$$R_k \geq \tau_k, \tag{P1b}$$

$$\gamma_k T\left(\tilde{\Psi}_k + \lambda\right) \leq \Gamma_k, \tag{P1c}$$

$$0 \leq \tilde{\Psi}_k \leq \Psi_{\max}, \tag{P1d}$$

$$0 \leq \gamma_k \leq 1. \tag{P1e}$$

In Problem (P1), $\mathbb{E}\{\cdot\}$ is the expected sum of discounted energy efficiencies of the EHSR, with β denoting the discount factor, preferring to yield a long-term gain, $R_k = \log(1 + \Psi_k|h_k|^2)$, presents the data rate of the k-th primary device, and τ_k denotes the minimum achievable rate by the k-th primary device. Constraint (P1a) gives the total amount of energy available in the battery of the EHSR at time t_{k+1}, whereas, Constraint (P1b) ensures the QoS requirements of the primary devices. Constraint (P1c) caps the total energy consumed by the EHSR by the total energy available in the battery of the EHSR at time t_k. Constraints (P1d) and (P1e) limit the values of transmit power and time sharing coefficient of EHSR, respectively, where $\Psi_{\max}$ is the maximum transmit power of the EHSR.

As it can be observed, Problem (P1) is non-convex because i) the objective function is a non convex long-term EE function, ii) Constraint (P1a) is not an affine function, and iii) Constraint (P1c) involves both optimization variables being multiplied making it non convex. However, in the considered optimization problem, timely decisions about EH and transmissions must be made according to the resources available at the EHSR. These decisions motivate the application of reinforcement learning to address this optimization problem, and the continuous nature of the optimization variables makes Problem (P1) well-suited to be approached using the MDDPG algorithm. However, (P1d) and (P1e) assume different ranges of values, which makes it challenging to directly apply the MDDPG algorithm. Therefore, we take advantage of the primal decomposition method and deal with Problem (P1) as a two-layer optimization problem to facilitate the application of the MDDPG algorithm.

Similar to [26], we first introduce an energy fluctuation parameter $\bar{\Gamma}_k$, which is the difference between harvested and consumed energy, and is given by

$$\bar{\Gamma}_k = \underbrace{(1-\gamma_k)T\left[\frac{\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]}{e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}}\right]|h_{k,o}|^2}_{\text{Harvested Energy}} - \underbrace{\gamma_k T\left(\tilde{\Psi}_k + \lambda\right)}_{\text{Consumed Energy}}. \tag{8}$$

Note that $\bar{\Gamma}_k$ can be interpreted as energy deficit if $\bar{\Gamma}_k < 0$ or energy surplus if $\bar{\Gamma}_k > 0$ at time t_k. Accordingly, the first layer problem can be formulated as follows

$$\underset{\gamma_k,\ \tilde{\Psi}_k}{\text{maximize}} \quad \Pi_k\left(\gamma_k, \tilde{\Psi}_k\right) \tag{P2}$$

$$\text{s.t.} \quad \bar{\Gamma}_k = (1-\gamma_k)T\left[\frac{\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]}{e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}}\right]|h_{k,o}|^2 - \gamma_k T\left(\tilde{\Psi}_k + \lambda\right), \tag{P2a}$$

$$\text{(P1b), (P1c), (P1d), (P1e)}, \tag{P2b}$$

and the second layer optimization problem is given by

$$\underset{\bar{\Gamma}_k}{\text{maximize}} \quad \mathbb{E}\left\{\sum_{k=1}^{N} \beta^{k-1} \Pi_k\left(\gamma_k, \tilde{\Psi}_k\right)\right\} \tag{P3}$$

$$\text{s.t.} \quad \Gamma_{k+1} = \min\left\{\Gamma_{\max}, \bar{\Gamma}_k + \Gamma_k\right\}, \tag{P3a}$$

which has been formulated following [31].

A. SOLUTION APPROACH

The solution approach to this problem is divided into two phases; in the first phase, we use convex optimization to derive closed-form expressions for the optimization variables for a given $\bar{\Gamma}_k$ in Problem (P2). Hence, the optimal solution is expressed as functions of $\bar{\Gamma}_k$ in Problem (P2), i.e., $\gamma_k^*(\bar{\Gamma}_k)$ and $\Psi_k^*(\bar{\Gamma}_k)$. In the second phase, we employ MDDPG algorithm to solve Problem (P3) while using the derived optimal solutions in the first phase. Consequently using the closed-form expressions, Problem (P3) could be reformulated as follows

$$\underset{\bar{\Gamma}_k}{\text{maximize}} \quad \mathbb{E}\left\{\sum_{k=1}^{N} \beta^{k-1} \Pi_k\left(\gamma_k^*(\bar{\Gamma}_k), \tilde{\Psi}_k^*(\bar{\Gamma}_k)\right)\right\} \tag{P4}$$

$$\text{s.t.} \quad \Gamma_{k+1} = \min\left\{\Gamma_{\max}, \bar{\Gamma}_k + \Gamma_k\right\}, \tag{P4a}$$

which highlights the fact that the action of the EHSR is to choose $\bar{\Gamma}_k$. We can observe that Problem (P4) is a single

variable function and the continuous nature of the parameter of interest, i.e., $\bar{\Gamma}_k$, makes this one-dimensional, continuous action space optimization problem best fitted to be addressed by a DRL algorithm, which in our case is the MDDPG algorithm.

B. CLOSED-FORM EXPRESSIONS FOR $\gamma_K^*(\bar{\Gamma}_K)$ AND $\Psi_K^*(\bar{\Gamma}_K)$

To find closed-form expressions, we rewrite Problem (P2) as follows

$$\underset{\gamma_k,\ \tilde{\Psi}_k}{\text{maximize}} \quad \Pi_k\left(\gamma_k, \tilde{\Psi}_k\right) \tag{P5}$$

$$\text{s.t.} \quad (1-\gamma_k)T\left[\frac{\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]}{e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}}\right]|h_{k,o}|^2 - \gamma_k T\left(\tilde{\Psi}_k+\lambda\right) - \bar{\Gamma}_k = 0, \tag{P5a}$$

$$(P1b), (P1c), (P1d), (P1e). \tag{P5b}$$

Problem (P5) is clearly a non-convex problem due to Constraint (P5a) not being affine and also the fact that optimization variables appear in multiplication in (P1c). Following [31], an equivalent form of (P5) is given by

$$\underset{\gamma_k}{\text{maximize}} \quad \hat{f}_o(\gamma_k) \tag{P6}$$

$$\text{s.t.} \ 0 \le \gamma_k \le 1, \tag{P6a}$$

where, $\hat{f}_o(\gamma_k) = \sup\{\Pi_k(\gamma_k, \tilde{\Psi}_k) |\ (P5a), (P1b), (P1c), (P1d)\}$. Using (P5a), we develop an expression for $\tilde{\Psi}_k$ as follows

$$\tilde{\Psi}_k = \left[\frac{(1-\gamma_k)T\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]|h_{k,o}|^2}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T}\right] - \frac{\bar{\Gamma}_k}{\gamma_k T} - \lambda. \tag{9}$$

Similarly, using (P1c), we develop an inequality expression for $\tilde{\Psi}_k$ as follows

$$\tilde{\Psi}_k \le \frac{\Gamma_k}{\gamma_k T} - \lambda. \tag{10}$$

Using (9) and (10), Problem (P5) is reformulated as a function of $\tilde{\Psi}_k$ as follows

$$\underset{\tilde{\Psi}_k}{\text{maximize}} \quad \Pi_k\left(\gamma_k, \tilde{\Psi}_k\right) \tag{P7}$$

$$\text{s.t.} \quad \tilde{\Psi}_k = \left[\frac{(1-\gamma_k)T\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]|h_{k,o}|^2}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T}\right] - \frac{\bar{\Gamma}_k}{\gamma_k T} - \lambda, \tag{P7a}$$

$$\tilde{\Psi}_k \le \frac{\Gamma_k}{\gamma_k T} - \lambda, \tag{P7b}$$

$$(P1b), (P1d). \tag{P7c}$$

We can observe that Problem (P7) is a function of $\tilde{\Psi}_k$ only, where Γ_k is fixed. Using (P7a), an optimal solution for $\hat{f}_o(\gamma_k)$, i.e., $\hat{f}_o^*(\gamma_k)$, can be computed as follows

$$\hat{f}_o^*(\gamma_k) = \frac{\gamma_k \log_2\left(1+\frac{|\tilde{h}_k|^2\left((1-\gamma_k)T\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]|h_{k,o}|^2 - A\bar{\Gamma}_k - A\gamma_k T\lambda\right)}{A\gamma_k T\left(1+\Psi_k|h_k|^2\right)}\right)}{\bar{\lambda}+\lambda}, \tag{11}$$

where $A = e^{\alpha_1\Psi_k} + e^{\alpha_1\alpha_2}$. Constraints (P7b) and (P1d) in Problem (P7) are guaranteed by the domain of $\hat{f}_o(\gamma_k)$, which can be expressed as follows

$$D = \left\{\gamma_k \middle| 0 \le \left[\frac{(1-\gamma_k)T\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]|h_{k,o}|^2}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T}\right] - \frac{\bar{\Gamma}_k}{\gamma_k T} - \lambda \le \min\left\{\Psi_{\max}, \frac{\Gamma_k}{\gamma_k T} - \lambda\right\}\right\}. \tag{12}$$

Using the optimal solution, $\hat{f}_o^*(\gamma_k)$, and using constraints in the domain of $\hat{f}_o(\gamma_k)$, we can reformulate Problem (P6) as in the following lemma.

Lemma-2: Reformulation of Problem (P6) can be expressed as:

$$\underset{\gamma_k}{\text{maximize}} \quad \hat{f}_o^*(\gamma_k) \tag{P8}$$

$$\text{s.t.} \quad \gamma_k \ge \frac{B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k}{B + T\left[\lambda+\Psi_{\max}\right]\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)}, \tag{P8a}$$

$$\gamma_k \ge 1 - \frac{\left[\bar{\Gamma}_k+\Gamma_k\right]\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)}{B}, \tag{P8b}$$

$$\gamma_k \le \frac{B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k}{B + \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)T\lambda}, \tag{P8c}$$

$$(P1b), (P1e). \tag{P8d}$$

In Problem (P8), $B = T\alpha_3[e^{\alpha_1\Psi_k}-1]|h_{k,o}|^2$. By exploiting the properties of constraints in the Problem (P8) a closed-form optimal solution can be derived as follows. First, to simplify, we recast Problem (P8) and get

$$\underset{\gamma_k}{\text{maximize}} \quad \frac{\gamma_k \log_2\left(1+\frac{\left((1-\gamma_k)p_1-p_2\right)}{\gamma_k}-p_3\right)}{p_4}, \tag{P9}$$

$$\text{s.t.} \quad \gamma_k \ge \max\left\{\frac{B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k}{B + T\left[\lambda+\Psi_{\max}\right]\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)}, 0, 1 - \frac{\left[\bar{\Gamma}_k+\gamma_k\right]\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)}{B}\right\}, \tag{P9a}$$

$$\gamma_k \le \min\left\{1 - \frac{\left[\bar{\Gamma}_k+\Gamma_k\right]\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)}{B}, 1\right\}, \tag{P9b}$$

$$(P1b), \tag{P9c}$$

where, $p_1 = \frac{|\tilde{h}_k|^2 T\alpha_3[e^{\alpha_1\Psi_k}-1]|h_{k,o}|^2}{AT(1+\Psi_k|h_k|^2)}$, $p_2 = \frac{\bar{\Gamma}_k|\tilde{h}_k|^2}{T(1+\Psi_k|h_k|^2)}$, $p_3 = \frac{\lambda|\tilde{h}_k|^2}{(1+\Psi_k|h_k|^2)}$, and $p_4 = \bar{\lambda}+\lambda$. Let $\hat{\gamma}_k$ be the optimal solution for (11), after some algebraic manipulations we get

$$\hat{f}_o^*(\hat{\gamma}_k) = \frac{\hat{\gamma}_k \log_2\left(1 - p_1 - p_3 + \frac{p_1 - p_2}{\hat{\gamma}_k}\right)}{p_4}, \tag{13}$$

where the argument of the log function, i.e., $(1 - p_1 - p_3 + \frac{p_1-p_2}{\hat{\gamma}_k})$ is strictly positive once all the constraints in Problem (P8) are met. The first and second derivatives of (13) w.r.t $\hat{\gamma}_k$ are, respectively, given as

$$= \frac{\log_2\left(1 - p_1 - p_3 + \frac{p_1-p_2}{\hat{\gamma}_k}\right) - \left(\frac{p_1-p_2}{\hat{\gamma}_k - p_1\hat{\gamma}_k - p_3\hat{\gamma}_k + p_1 - p_2}\right)}{p_4}, \tag{14}$$

$$= -\frac{\left(p_1^2 - p_2^2\right)^2}{p_4\hat{\gamma}_k\left(\hat{\gamma}_k - p_1\hat{\gamma}_k - p_3\hat{\gamma}_k + p_1 - p_2\right)^2}, \tag{15}$$

which is less than 0, and thus implies that the function $\hat{f}_o^*(\gamma_k)$ is a concave function for $\gamma_k \geq 0$, hence there exists a single solution (maximum) $\hat{\gamma}_k$. Following the concave nature of (13), it becomes the root of the following equation

$$\frac{\log_2\left(1 - p_1 - p_3 + \frac{p_1-p_2}{\hat{\gamma}_k}\right) - \left(\frac{p_1-p_2}{\hat{\gamma}_k - p_1\hat{\gamma}_k - p_3\hat{\gamma}_k + p_1 - p_2}\right)}{p_4} = 0. \tag{16}$$

Simplifying (16) further we get

$$\frac{z^*\log_2(z^*) - (z^* - 1 + p_1 + p_3)}{z^* p_4} = 0, \tag{17}$$

where, $z^* = (1 - p_1 - p_3 + \frac{p_1-p_2}{\hat{\gamma}_k})$. Further algebraic manipulation reduces (17) to

$$z^*\log_2(z^*) - z^* = 1 - p_1 - p_3. \tag{18}$$

Note that (18) could easily be transformed into the form of Lambert W function. Let $z^* = e^{q^*}$, Eq. (18) is transformed to

$$e^{q^*}(q^* - 1) = 1 - p_1 - p_3. \tag{19}$$

To make the application of Lambert W function to (19) possible, we multiply both sides of (19) with e^{-1}, consequently, we get

$$(q^* - 1)\, e^{q^* - 1} = e^{-1}(1 - p_1 - p_3), \tag{20}$$

which is in the form of the Lambert W function, i.e., $Ue^U = Y$. Since the value of $(1 - p_1 - p_3)$ can be positive or negative, two real-valued solutions of q^* are given by

$$q^* = U_0\left(e^{-1}(1 - p_1 - p_3)\right) + 1, \tag{21}$$

and

$$q^* = U_{-1}\left(e^{-1}(1 - p_1 - p_3)\right) + 1. \tag{22}$$

Fortunately, the solution corresponding to the principal branch of Lambert W function given by (22), could be discarded, as it is not a feasible solution. Consequently, using (21), the optimal solution for z^* is given by

$$z^* = e^{q^*} = e^{U_0\left(e^{-1}(1-p_1-p_3)\right)+1}. \tag{23}$$

Using the relation $z^* = (1 - p_1 - p_3 + \frac{p_1-p_2}{\hat{\gamma}_k})$ an explicit equation for $\hat{\gamma}_k$ is given by

$$\hat{\gamma}_k = \frac{p_1 - p_2}{z^* + p_1 + p_3 - 1}. \tag{24}$$

Based on (23), (24) becomes

$$\hat{\gamma}_k = \frac{p_1 - p_2}{e^{U_0\left(e^{-1}(1-p_1-p_3)\right)+1} + p_1 + p_3 - 1}. \tag{25}$$

Although $\hat{\gamma}_k$ could be an optimal solution for (11), it may not necessarily be an optimal solution for (P8), because $\hat{\gamma}_k$ is obtained by discarding the constraints, (P8a) and (P8b), and (P8c), thereby, it may violate any of the inequality constraints. Therefore, an optimal solution for problem (P8) is computed as follows:

Firs we define the feasibility set for γ_k as

$$\max\{0, \Xi\} \leq \gamma_k \leq \min\{1, \tilde{\Xi}\}, \tag{26}$$

where, $\Xi = \max\{\frac{B-(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2})\bar{\Gamma}_k}{B+T[\lambda+\Psi_{\max}](e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2})}, 1 - \frac{[\bar{\Gamma}_k+\Gamma_k](e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2})}{B}\}$, and $\tilde{\Xi} = \min\{\frac{B-(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2})\bar{\Gamma}_k}{B+(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2})T\lambda}\}$. By using the concavity of the objective function, i.e., $\hat{f}_o^*(\gamma_k)$, (given by Eq. (16)), we have

$$0 \leq \hat{\gamma}_k \leq \tilde{\Xi}. \tag{27}$$

Based on (28), it is evident that the relationship among 0, $\tilde{\Xi}$, and $\hat{\gamma}_k$ is predetermined. However, the relationship involving $\hat{\gamma}_k$, 1, and Ξ is not fixed, leading to three distinct cases resulting in three different solutions:

- Case 1: When $\hat{\gamma}_k \geq \max\{0, \Xi\}$ and $\hat{\gamma}_k \leq \min\{1, \tilde{\Xi}\}$, $\hat{\gamma}_k$ is inside of the feasible set of problem (P8) as shown in Fig. 2 (a). Hence, $\hat{\gamma}_k$ is the optimal solution to the problem.
- Case 2: When $\hat{\gamma}_k \leq \Xi$, the feasible set of γ_k is $[\Xi, \min\{1, \tilde{\Xi}\}]$ and $\hat{\gamma}_k$ is at the left-hand side of the feasible set as shown in Fig. 2 (b). The function is monotonically decreasing over the feasible set, and hence Ξ is the optimal solution to the problem.

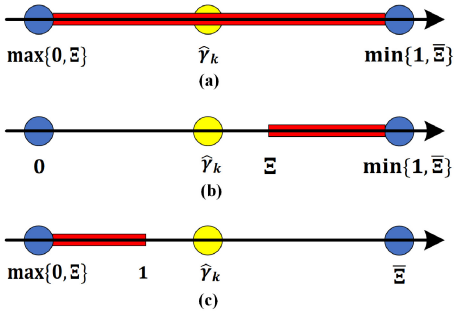

FIGURE 2. Illustration for possible scenarios for $\hat{\gamma}_k$.

- Case 3: When $\hat{\gamma}_k \geq 1$, the feasible set of γ_k becomes $[\max\{0, \Xi\}, 1]$ and $\hat{\gamma}_k$ is at the right-hand side of the feasible set as shown in Figure 2(c). The function is monotonically increasing over the feasible set, and hence 1 is the optimal solution to the problem.

Consequently, the optimal solution for Problem (P8) is given by

$$\gamma_k^*(\bar{\Gamma}_k) = \begin{cases} \min\{\Omega, 1\} & \text{if } R_k \geq \tau_k \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

where $\Omega = \max\{\hat{\gamma}_k, \Xi\}$, and it can be observed that (28) is a function of $\bar{\Gamma}_k$. Finally, using (28), the optimal solution $\Psi_k^*(\bar{\Gamma}_k)$ is given by (from (9))

$$\Psi_k^*(\bar{\Gamma}_k) = \left[\frac{(1-\gamma_k^*(\bar{\Gamma}_k))T\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]|h_{k,o}|^2}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k^*(\bar{\Gamma}_k)T}\right] - \frac{\bar{\Gamma}_k}{\gamma_k^*(\bar{\Gamma}_k)T} - \lambda. \tag{29}$$

Consequently, we employ the optimal values obtained from (28) and (29) to address Problem (P4) through the application of the MDDPG algorithm. Specifically, the MDDPG algorithm facilitates the determination of an optimal action, denoted as $\bar{\Gamma}_k$. This action is used in calculating the values in (28) and (29), which are subsequently reincorporated into the objective function of Problem (P4), thereby enhancing the EE of the EHSR.

IV. PROPOSED DRL FRAMEWORK

In this section, we first discuss the basic principles of the DDPG algorithm. Following that, we briefly explain how the proposed MDDPG algorithm differs from the DDPG algorithm. Finally, we give insight into how to structure our optimization Problem (P4) within the framework of the MDDPG algorithm.

A. UNDERLYING FRAMEWORK OF MDDPG

The DDPG algorithm generally operates as an actor-critic DRL method, tailored ideally for DRL challenges with continuous action dimensions. It finds its roots in the Deep Q-Network (DQN) and Deterministic Policy Gradient (DPG) [32]. In the DDPG algorithm, a pool of past experiences is stored in the replay buffer, and a batch of experiences is chosen at random from this pool to train the agent and enhance its learning process. The size of the pool and the batch of experiences can significantly affect the performance of the DDPG algorithm. For instance, a random selection of a batch from a small-sized pool can degrade the episodic reward performance, as it may not provide a diverse and representative sample of experiences for effective learning. Consequently, a larger pool size is often selected in the DDPG algorithm to ensure a richer variety of experiences for training.

To further enhance performance, the MDDPG algorithm employs the concept of combined experiences, integrating efficient sampling of experience tuples. This process ensures that the latest experience tuples encountered by the agent are included in the batch selection procedure. By incorporating the most recent experiences, the algorithm benefits from up-to-date information about the environment, which helps improve learning efficiency and stability. This approach reduces the dependency on a large experience pool size while improving performance in terms of episodic rewards. The inclusion of recent experiences ensures that the agent adapts more quickly to changes in the environment, thereby achieving better overall performance compared to the traditional DDPG algorithm.

B. EXECUTION OF MDDPG FRAMEWORK

To adapt the MDDPG to our problem, we define the necessary state space, action space, and reward parameter as follows.

1) STATE SPACE

The available energy from the EHSR and all of the channel gains are given as a tuple in the state space as

$$\bar{s}_k = \left[\Gamma_k,\ |h_k|^2,\ |h_{k,o}|^2,\ |\tilde{h}_k|^2\right]^T. \tag{30}$$

2) ACTION SPACE

The action space contains the action taken by the EHSR. In our case, the action space is $\bar{\Gamma}_k$. The upper and lower bounds on the value of $\bar{\Gamma}_k$ is given by

$$-\min\left\{T(\Psi_{\max}+\lambda), \Gamma_k\right\} \leq \bar{\Gamma}_k \leq \min\left\{\Gamma_{\max}-\Gamma_k, \frac{B}{A}\right\}. \tag{31}$$

The upper limit is valid when the EHSR refrains from transmission, specifically when $\gamma_k = 0$, and solely harvests energy. Conversely, the lower limit applies when the EHSR exclusively transmits without harvesting energy; that is, when $\gamma_k = 1$. Given its capacity to adopt considerably smaller or substantially larger values, (31) can be normalized as follows

$$\bar{\Gamma}_k = \zeta_k \min\left\{\Gamma_{\max}-\Gamma_k, \frac{B}{A}\right\} - (1-\zeta_k)\min\left\{T(\Psi_{\max}+\lambda), \Gamma_k\right\}. \tag{32}$$

Hence, the suitable action parameter for the suggested algorithm is $\zeta_k \in [0, 1]$.

3) REWARD PARAMETER

The reward metric is defined by the EE of the EHSR, as represented by

$$\Pi_k\left(\gamma_k^*(\bar{\Gamma}_k), \tilde{\Psi}_k^*(\bar{\Gamma}_k)\right) = \frac{\gamma_k^*(\bar{\Gamma}_k)\log_2\left(1+\frac{\tilde{\Psi}_k^*(\bar{\Gamma}_k)|\tilde{h}_k|^2}{1+\Psi_k|h_k|^2}\right)}{\bar{\lambda}+\lambda}. \tag{33}$$

Using the specified state space, action space, and reward allows for the direct application of MDDPG to the Problem (P4).

Algorithm 1 MDDPG Algorithm Steps for EE Maximization

Initialization

1: Initialize the actor network, $\mathcal{A}(\bar{s}|\theta^{\mathcal{A}})$ with random weights, $\theta^{\mathcal{A}}$.
2: Initialize the critic network, $\mathcal{C}(\bar{s}, a|\theta^{\mathcal{C}})$, randomly with random weights, $\theta^{\mathcal{C}}$;
3: Initialize the actor target network, $\mathcal{A}_k(\bar{s}_k|\theta^{\mathcal{A}_k})$ and the critic target network $\mathcal{C}_k(\bar{s}_k, a_k|\theta^{\mathcal{C}_k})$ with weights $\theta^{\mathcal{A}_k}$ and $\theta^{\mathcal{C}_k}$, respectively;
4: Create a replay buffer with a capacity of $\mathcal{B}$;

Main Loop

5: **for** `Episode` $i = 1$ `to` $\mathcal{Q}$ **do**
6: Produce a random noise process to facilitate action exploration;
7: Capture the initial state $\bar{s}_k$ from the environment;
8: **for** `Episode Steps` $j = 1$ `to` $\bar{\mathcal{Q}}$ **do**
9: Select an action a_j using the current policy and explor-action noise and obtain $\bar{\Gamma}_k$ from (32);
10: Execute the action a_k, obtain reward r_k and retrieve the subsequent state $\bar{s}_{k+1}$ from the environment
11: Save the observed experience tuple, $(\bar{s}_k, a_k, r_k, \bar{s}_{k+1})$, into the replay buffer;
12: Draw a random batch of $\mathcal{S} - 1$ experiences from the replay buffer;
13: Incorporate the most recent experience tuple into the batch of experiences chosen in step 12;
14: Define $y_k = r_k + \gamma \mathcal{C}_k(\bar{s}_{k+1}, \mathcal{A}_k(\bar{s}_{k+1}|\theta^{\mathcal{A}_k})|\theta^{\mathcal{C}_k})$;
15: Minimize the loss function $L(\theta^{\mathcal{C}}) = \sum_k |y_k - \mathcal{C}(\bar{s}_k, a_k|\theta^{\mathcal{C}})|^2$ and subsequently update the critic network;
16: Update the actor network by applying the gradient: $\nabla_{\theta^{\mathcal{A}}} J(\theta^{\mathcal{A}}) = \nabla_{a_k} Q(\bar{s}_k, a_k|\theta^{\mathcal{C}}) \nabla_{\theta^{\mathcal{A}}} \mathcal{A}(\bar{s}_k|\theta^{\mathcal{A}})$;
17: Update both the actor and critic target networks: $\theta^{\mathcal{A}_k} \rightarrow \xi\theta^{\mathcal{A}} + (1-\xi)\theta^{\mathcal{A}_k}$ and $\theta^{\mathcal{C}_k} \rightarrow \xi\theta^{\mathcal{C}} + (1-\xi)\theta^{\mathcal{C}_k}$
18: **end for**
19: **end for**

4) STEPS INVOLVED IN EXECUTING MDDPG

Algorithm 1 outlines the execution steps of the MDDPG algorithm. Initially, the actor and critic networks, together with their respective target networks, have their network weights randomly initialized in the first stage (lines 1-3, in Alg. 1). Subsequently, a buffer replay of capacity $\mathcal{B}$ is created (line 4, in Alg. 1) to store and sample the past experiences. As the algorithm initiates, in each episode, random noise is generated to help the agent explore the environment first and concurrently, observe initial states from the environment (lines 6-7, in Alg. 1). Progressing sequentially, in each episodic step, the agent selects the action using the current policy, evaluates the reward using the predicted action and then transitions to the next state (lines 9-10, in Alg.1). Within this iterative framework, the newly observed state along with the preceding state, action, and reward is stored in the buffer replay (line 11, in Alg.1). An experiences extraction procedure is then executed, wherein a random batch of experiences, denoted as $\mathcal{S} - 1$, selected from replay buffer, to which the most recent tuple of experiences is appended (lines 12-13, in Alg.1). The value network is updated next, using the Bellman equation (line 14, Alg. 1). The critic network is updated using the loss function given in line 15, in Alg. 1. The policy of the actor network is updated by applying the gradient to the objective function with respect to the policy parameter, $\theta^{\mathcal{A}}$ (line 16, Algorithm 1). To monitor the learning progress of the networks, a copy of the target networks parameters are used (line 17, in Alg. 1). The cycle is repeated $\bar{\mathcal{Q}}$ times in each episodes and a total of $\mathcal{Q}$ episodes are executed for training and learning of the agent.

V. SIMULATION RESULTS

In this section, we analyze the EHSR's performance using the key performance metrics, i.e., EE, sum rate, and EH through extensive simulations. These metrics are assessed for both the proposed MDDPG and the baseline DDPG algorithms, while simultaneously comparing their performance for the nonlinear and baseline linear EH models. We then describe the outcomes of the simulations, provide an in-depth discussion, and highlight critical observations from the simulation results.

To optimize the EE of the EHSR, we train an RL agent consisting of an actor-critic network pair, where neural network architectures of both the actor and critic networks employ two hidden layers, each employing $M_k = 64$ nodes. The hidden layers use rectified linear activation (ReLU) to avoid vanishing gradient concerns, while the output layer exploits the capabilities of the hyperbolic tangent function. The actor and critic networks' learning rates are, respectively, set to 0.002 and 0.004. We incorporate a reward discount coefficient, $\beta = 0.9$, and a network update factor, $\xi = 0.01$. A batch size of $\mathcal{S} = 64$ is designated for replay experience.

As for network setup according to the system model (c.f., Section II), BS is deployed at the intersection of the $x - y$ plane, and the EHSR is placed at a point (1 m, 1 m). Fig. 3 shows the overall network setup, including the different deployment configurations of the primary devices. The system, having a bandwidth of 1 MHz, operates at a center frequency of 914 MHz. We assume a Rayleigh fading environment for our simulations, and we have integrated dynamic channels to more accurately simulate real-world conditions. The noise profile involves an additive white Gaussian noise with a power spectral density of -170 dBm/Hz. Lastly, we consider the path loss model from [33], with a path loss exponent of 3. We assume nonlinear EH model parameters as, $\alpha_1 = 150$, $\alpha_2 = 0.014$, and $\alpha_3 = 13.8$ dBm [34], [35]. For the linear EH model, we set an efficiency rate, $\eta = 0.8$. The initial energy reserve within the battery is fixed at $\Gamma_{\max} = 0.2$ J. We assume a time slot of length, $T = 1$s, and the maximum transmit power of the EHSR is set to $\Psi_{\max} = 0.1$ W.

A. COMPLEXITY ANALYSIS NOTE

Using DRL algorithms results in higher computational complexity compared to non-DRL methods. For instance, the greedy scheme, which uses all available energy for data transmission, is computationally efficient. In contrast,

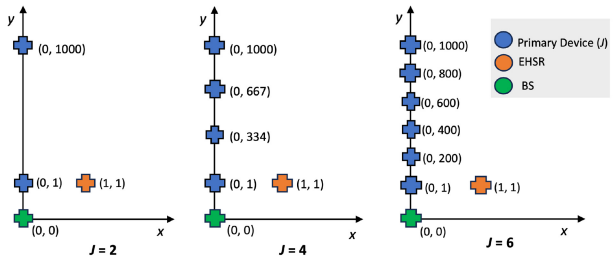

FIGURE 3. Network setup used for performance analysis, with different deployment scenarios of the primary devices.

assuming a fully trained actor network, DRL algorithms have a complexity of $O(M_k^2)$, with $O(\cdot)$ presenting the computational complexity operator. The training phase's complexity is difficult to analyze due to dependencies on gradient iterations and the structure of the critic network. Detailed complexity analysis and mitigation strategies are important future research directions. Leveraging mobile edge computing (MEC) for offloading computations to the base station is a promising approach to address the high complexity of DRL algorithm training.

B. STABILITY AND CONVERGENCE OF THE MDDPG

To ensure the stability and convergence of the MDDPG framework in dynamic environments, we employ several strategies: (i) Experience replay and mini-batch training, which break correlations between experiences and reduce gradient variance; (ii) target networks for both the actor and critic, providing consistent updates by smoothing abrupt changes; (iii) parameter noise, enhancing exploration and avoiding local optima; and (iv) careful tuning of hyperparameters and robust training, including early stopping and periodic evaluation to prevent overfitting. These measures collectively enhance the stability, reliability, and effectiveness of our MDDPG framework.

C. LEARNING EFFICIENCY ANALYSIS OF DRL MODELS

In the learning efficiency analysis, we compare the performance of the proposed MDDPG algorithm with the baseline DDPG algorithm along with non-DRL algorithms including the random method and greedy method in terms of episodic reward (or EE) and sum rate. In the greedy algorithm, the EHSR uses all available energy for data transmission before it starts EH. Specifically, it sets its transmission power at $\Psi_{\max}$ and calculates γ_k as $\gamma_k = \min\{1, \frac{\Gamma_k}{T\Psi_{\max}}\}$. In the random scheme, the transmit power of the EHSR is set to $\Gamma_{\max}$, whereas the value of γ_k is chosen randomly from a uniform distribution ranging between 0 and $\min\{1, \frac{\Gamma_k}{T\Psi_{\max}}\}$. The analysis focuses on evaluating the convergence performance of the algorithms when exposed to non-linear and linear EH dynamics. In Fig. 4, when comparing the EHSR's episodic reward for the proposed MDDPG with the baseline schemes, we can make the following observations.

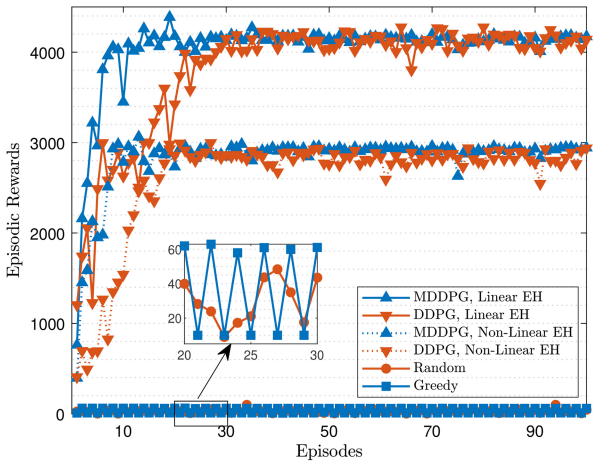

FIGURE 4. Comparison of the episodic rewards of EHSR for the proposed MDDPG and baseline algorithms under linear and non-linear EH models, given $J = 2$, $\lambda = 0$ dBm, and $\Psi_k = 30$ dBm.

TABLE 1. Statistical comparison of the EE performance between the proposed MDDPG algorithm and the baseline algorithms for both linear and non-linear cases.

Algorithm	EH Model	Average Reward	Standard Deviation
MDDPG	Linear	≈ 4037	≈ 449.0125
MDDPG	Non Linear	≈ 2824	≈ 353.14
DDPG	Linear	≈ 3810	≈ 654.39
DDPG	Non Linear	≈ 2615	≈ 592.82
Random	Linear/Non Linear	≈ 30.73	≈ 16.64
Greedy	Linear/Non Linear	≈ 35.32	≈ 25.29

- Episodic reward reduces by approximately 30% for the non-linear EH model, irrespective of the learning approach, demonstrating the complexity of learning in practical EH settings with inherent non-linearities.
- Still, the proposed MDDPG outperforms the DDPG and the non-DRL methods for the non-linear EH model by converging to higher episodic reward.
- Meanwhile, the proposed MDDPG converges faster than the DDPG, especially for linear EH.

Table 1 presents a statistical comparison of the episodic reward performance between the proposed MDDPG algorithm and baseline algorithms under both linear and non-linear EH models. The MDDPG algorithm outperforms the DDPG and other baseline algorithms in terms of average reward for both linear and non-linear EH models, showing more consistency (lower standard deviation) in the non-linear case compared to DDPG. The random and greedy methods perform significantly worse in both models. In the following analysis, we will omit discussions on non-DRL methods due to their significantly inferior performance compared to DRL algorithms. Therefore, our focus will be on comparing the proposed MDDPG algorithm with the baseline DDPG algorithm.

Fig. 5 compares the EHSR's sum rate for the proposed MDDPG and the baseline algorithms, with two different locations of the primary devices. In both scenarios, the MDDPG algorithm exhibits superior performance in terms of

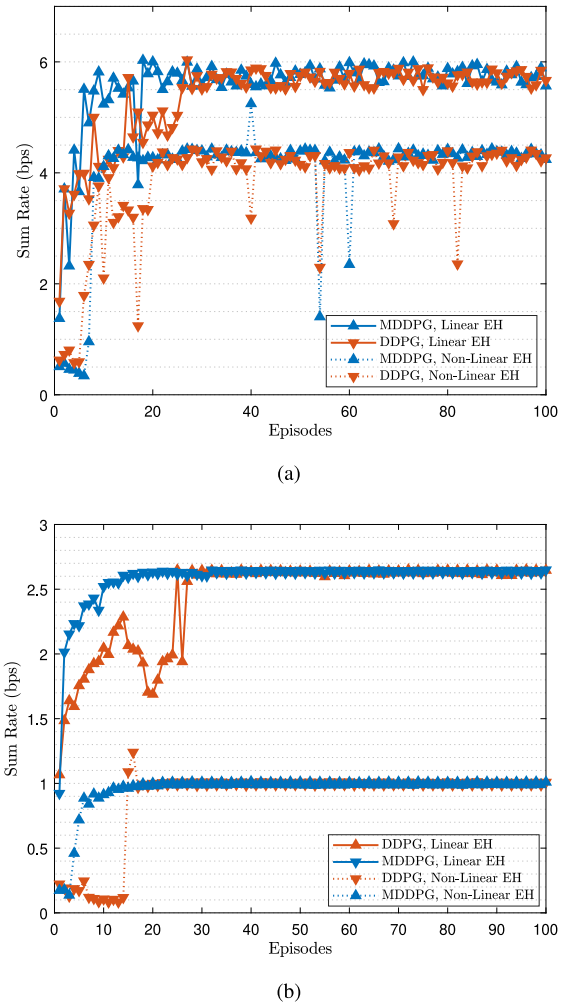

FIGURE 5. Comparison of the sum rate of EHSR for the proposed MDDPG and baseline algorithms under linear and non-linear EH models, given $J = 2$, $\lambda = 0$ dBm, and $\Psi_k = 30$ dBm, (a) for M_1 located at (0 m, 1 m), and M_2 located at (0 m, 1000 m), (b) for M_1 located at (0 m, 1 m), and M_2 located at (0 m, 100 m).

convergence rate. This pattern occurs because the MDDPG algorithm assembles an optimal collection of experience tuples, facilitating the agent to learn more efficiently and achieve convergence before the DDPG algorithm. Moreover, the reduction in the sum rate when the EHSR operates under the influence of a non-linear EH model is evident. With linear EH models offering an optimistic perspective, the proposed MDDPG offers an efficient solution for the transmission and harvesting cycle optimization of an EHSR in practical settings. It can be observed that, In scenarios where primary users are in closer proximity to the EHSR, it becomes apparent that the EHSR experiences a diminished sum rate. This effect arises from the increased interference generated by both primary devices.

D. ENERGY EFFICIENCY (EE) ANALYSIS

After showing the suitability of the proposed MDDPG for efficient learning in the presence of EH non-linearities, we analyze the EE of EHSR with respect to various parameters,

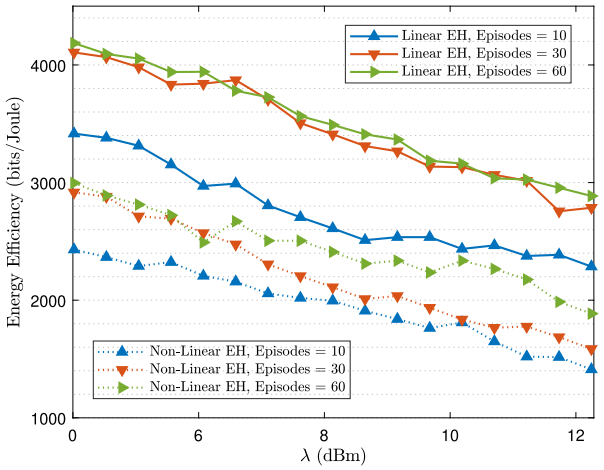

FIGURE 6. EE evaluation of the EHSR in relation to total RF power under linear and non-linear EH models, with MDDPG algorithm and parameters $J = 2$, and $\Psi_k = 30$ dBm.

including RF circuit power (λ), primary devices (J), and the power of primary devices (Ψ_k). This analysis considers both the linear and non-linear EH models.

1) VARYING RF POWER

Fig. 6 examines the EE of EHSR with respect to total RF power for selected episodes, where the EE tends to decrease with the increase in total RF power. Importantly, EE for the non-linear EH model is approximately 30% lower than the linear model, with the trend maintaining for increasing RF power. Further, the EE of the RL agent converges after 30 episodes for the linear model, whereas the non-linear model exhibits dynamic variations even at 60 episodes, highlighting the intricate complexities and nuances of learning in practical EH conditions.

2) DIFFERENT NUMBERS OF PRIMARY DEVICES

In Fig. 7, we present the EHSR's EE for different numbers of primary devices (J) in the environment. It can be observed that by increasing the number of primary devices in the network, the EE of the EHSR is reduced. Further, the algorithm convergence for the linear EH model is faster than the non-linear model. This arises due to two primary factors. Firstly, the state space that the DRL agent must navigate becomes exponentially complex as the number of devices increases. Secondly, the primary users tend to be situated farther away from the BS, consequently impacting the overall EE. As a result, the learning process of the DRL agent struggles to optimize its policy over a varied and dynamic environment, consequently reducing the EE of the EHSR. Still, capturing this interplay is vital for determining the optimal operational state and energy performance of the EHSR.

3) IMPACT OF PRIMARY DEVICE'S POWER

Fig. 8 presents the EE performance of the EHSR for the MDPPG algorithm when subjected to different power levels

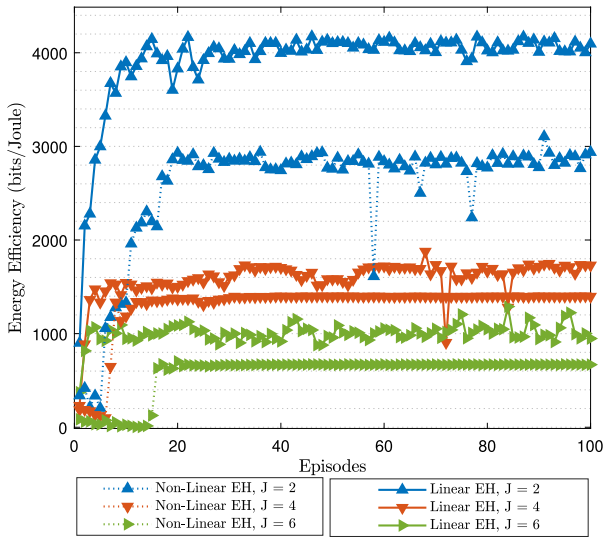

FIGURE 7. EE evaluation of the EHSR under linear and non-linear EH models for different number of primary devices (J), with MDDPG algorithm and parameters $\lambda = 0$ dBm and $\Psi_k = 30$ dBm.

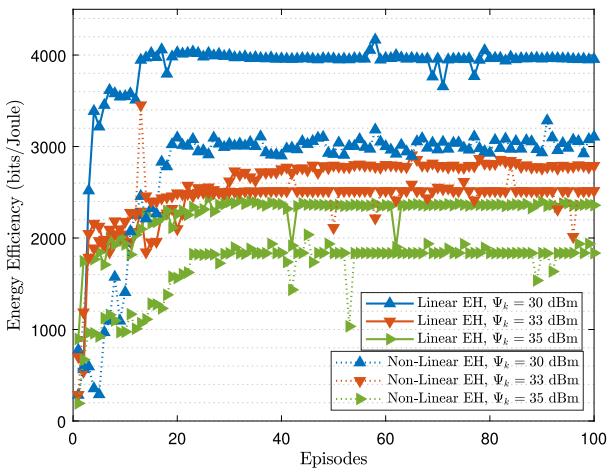

FIGURE 8. EE performance of the EHSR under linear and non-linear EH models for different power settings of the primary device, with MDDPG algorithm and parameters $J = 2$, and $\lambda = 0$ dBm.

of the primary device (Ψ_k) with parameters $J = 2$ and $\lambda = 0$ dBm. The EHSR's EE performance is observed to be adversely affected by the increase in Ψ_k because the EHSR experiences increased interference from the primary device (see (7) for the EE of EHSR). For instance, under any EH model, the EE is reduced by 41% when the Ψ_k increases from 30 dBm to 35 dBm. It implies that although the amount of energy harvested is expected to increase with the increase in primary devices' power level (as shown in the next section), the EHSR needs more transmit power for CR-NOMA transmissions. Meanwhile, the non-linear EH model leads to a significant reduction in EE compared to the linear model, while the DRL agents' convergence rate is also affected by the non-linear dynamics of the EH. This analysis highlights the significance of the appropriate EH model for

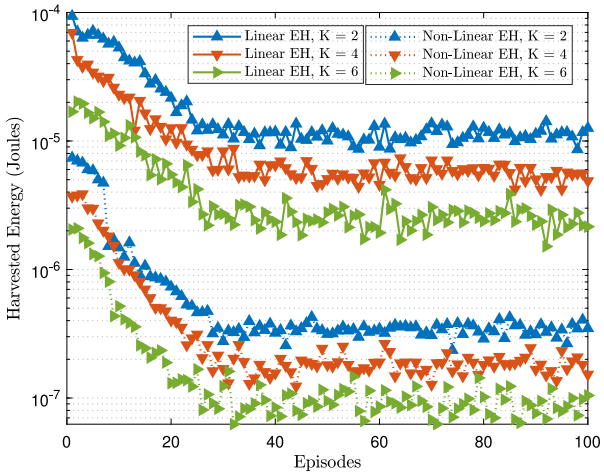

FIGURE 9. Analysis of the energy harvested by the EHSR for linear and non-linear EH models for different number of primary devices, with the MDDPG algorithm and parameters $\lambda = 0$ dBm and $\Psi_k = 30$ dBm.

optimizing the performance of EHSR when exposed to the different dynamics of the primary system.

E. ENERGY HARVESTING (EH) ANALYSIS

In the following, we turn to the EH analysis of EHSR with respect to the behavior of the primary network/devices.

1) DIFFERENT NUMBER OF PRIMARY DEVICES

We evaluate the EHSR's EH performance under the MDDPG algorithm while exposing it to different numbers of primary devices in Fig. 9. It can be observed that the agent's EH performance at initial episodes is higher due to policy exploration before converging at around 30 episodes. After convergence, there is a significant performance gap between the ideal and practical EH models, where the non-linear model captures the diverse dynamics of practical EH mechanisms. Further, the harvested energy appears to be influenced by the density and placements, emphasizing the primary devices' role in influencing the overall EH efficiency of EHSR.

2) IMPACT OF PRIMARY DEVICE'S POWER

Fig. 10 shows the energy captured by the EHSR with respect to the transmit power of the primary devices, with $J = 2$ and $\lambda = 0$ dBm. It is evident that the power level of the primary devices significantly impacts the EH efficiency. By increasing the transmit power of the primary device, the agent is able to harvest more energy from the RF transmission; however, the impact is prominent for the linear EH model compared to the non-linear EH model. The relationship between primary device power and harvested energy, especially under the non-linear model, provides insight into potential saturation points and the practical limitations of EH. In conclusion, Fig. 10 illustrates the significance of comprehending the intricacies of EH behaviors, especially in relation to primary device power levels, and highlights the trade-offs and considerations

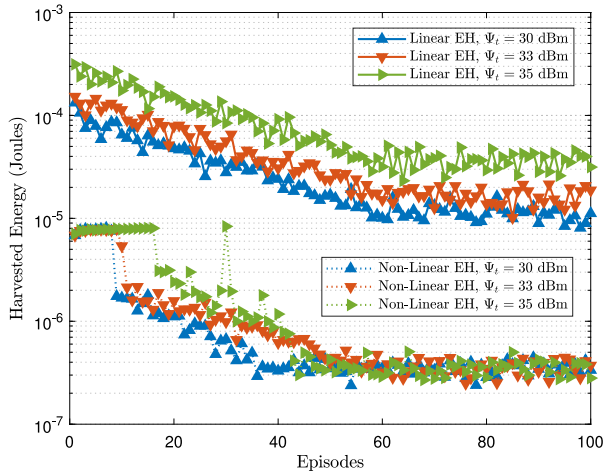

FIGURE 10. Analysis of the energy captured by the EHSR using both linear and non-linear EH paradigms for different power levels of primary devices, with the MDDPG algorithm and parameters $J = 2$ and $\lambda = 0$ dBm.

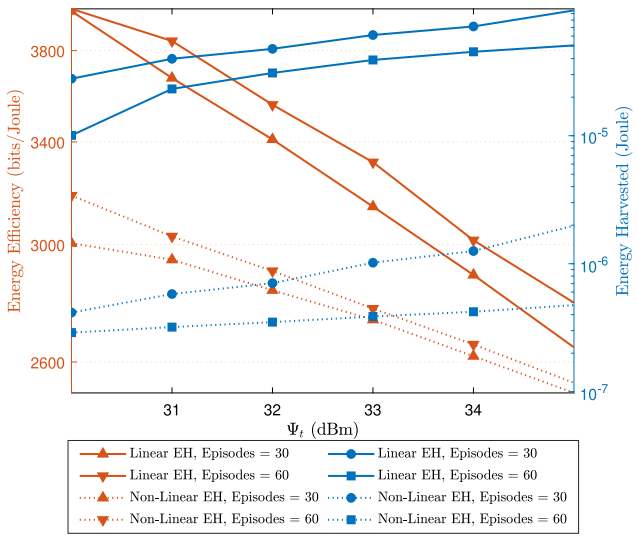

FIGURE 11. Joint influence of transmit power of primary devices (Ψ_k) on both EE and EH of the EHSR, evaluated under linear and non-linear EH models. The simulation results consider two different episode counts of the MDDPG algorithm, and $J = 2$.

essential for optimizing the transmission and harvesting parameters of a symbiotic radio under practical EH models.

F. JOINT EE AND EH ANALYSIS

To provide a cohesive understanding/visualization of EE and EH metrics of EHSR, Fig. 11 shows a y-y plot that simultaneously contrasts the impact of the transmit power of primary devices (Ψ_k) on the two metrics. Evidently, the Ψ_k impacts both EE and EH; EE decreases while EH increases linearly with the increase in Ψ_k. This result can be interpreted as the agent receiving more interference (power) by increasing Ψ_k; hence it can harvest more but needs to spend more energy in CR-NOMA. In conclusion, Fig. 11 symbolizes the complex influences of transmit power on the operational effectiveness and energy-harvesting capacities of the EHSR. These two metrics, when compared for different EH models, highlight challenges associated with energy management and transmission optimization in realistic settings.

VI. CONCLUSION

To integrate RF energy-harvesting symbiotic radios (EHSRs) with scheduled primary devices in the network, this article employed the cognitive radio-inspired non-orthogonal multiple access (CR-NOMA) framework. We designed an energy efficiency (EE) metric while including a non-linear practical EH model and also considered the effect of RF circuitry and RF signal processing power consumption inherent to the EHSR. To maximize the EE of the EHSR, the study proposed a novel deep reinforcement learning (DRL)–engineered transmission approach that coordinates the uplink transmissions of EHSR. The two-layer optimization problem is solved by adopting convex optimization in the first layer, where closed-form expressions are derived for the optimal parameters, and using the derived parameters, DRL, in specific, modified deep deterministic policy gradient (MDDPG) is deployed in the subsequent layer to solve a high dimensional continuous space optimization problem. Simulation results showed that the EE performance under a non-linear model is mostly lower than the ideal linear EH model, which overestimates the potential of EH. Nevertheless, the EE optimization strategy proposed in this study is more practical, encompassing the complexities and non-linearities induced by the non-linear model, thereby leading to the realistic design of transmission and harvesting policy. In the future, the proposed model can be extended to accommodate multiple EHSRs in the network, which could be addressed as a multiagent DRL problem. Additionally, backscatter communication can be considered a promising research direction for ultra-low power communication scenarios.

APPENDIX
PROOF FOR LEMMA-1

Recall that the practical non-linear EH model is defined as

$$\Phi_{\text{Prac}}(\Psi_k) = \frac{\bar{P} - \alpha_3 S}{1 - S}, \tag{34}$$

where,

$$\bar{P} = \frac{\alpha_3}{1 + e^{(-\alpha_1(\Psi_k - \alpha_2))}},$$

and

$$S = \frac{1}{1 + e^{(\alpha_1 \alpha_2)}}.$$

Putting the values of $\bar{P}$ and S in (34) gives

$$\Phi_{\text{Prac}}(\Psi_k) = \frac{\frac{\alpha_3}{1+e^{-\alpha_1(\Psi_k - \alpha_2)}} - \alpha_3 \frac{1}{1+e^{\alpha_1 \alpha_2}}}{1 - \frac{1}{1+e^{\alpha_1 \alpha_2}}}. \tag{35}$$

Factoring out common terms and taking LCM in the denominator of (35), we get

$$\Phi_{\text{Prac}}(\Psi_k) = \frac{\alpha_3\left[\frac{1}{1+e^{-\alpha_1(\Psi_k-\alpha_2)}} - \frac{1}{1+e^{\alpha_1\alpha_2}}\right]}{\frac{1+e^{\alpha_1\alpha_2}-1}{1+e^{\alpha_1\alpha_2}}}. \tag{36}$$

After performing some algebraic manipulation, we get

$$\begin{aligned}\Phi_{\text{Prac}}(\Psi_k) &= \alpha_3\left[\frac{1}{1+e^{-\alpha_1(\Psi_k-\alpha_2)}} - \frac{1}{1+e^{\alpha_1\alpha_2}}\right]e^{-\alpha_1\alpha_2}\left(e^{\alpha_1\alpha_2}+1\right)\\ &= \alpha_3 - \frac{\left[e^{-\alpha_1\alpha_2}(e^{\alpha_1\alpha_2}+1)\right]\alpha_3}{e^{\alpha_1(\Psi_k-\alpha_2)}+1}.\end{aligned} \tag{37}$$

Further simplifying (37) completes the proof.

PROOF FOR LEMMA-2

In this lemma, we show the proof of constraints (P8a), (P8b), and (P8c).

PROOF FOR (p8a)

Recall from (9) that

$$\tilde{\Psi}_k = \left[\frac{(1-\gamma_k)T\alpha_3\left[e^{\alpha_1\Psi_k}-1\right]|h_{k,o}|^2}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T}\right] - \frac{\bar{\Gamma}_k}{\gamma_k T} - \lambda, \tag{38}$$

While from (P1d), we have

$$\tilde{\Psi}_k \geq 0. \tag{39}$$

To simplify, let $T\alpha_3[e^{\alpha_1\Psi_k}-1]|h_{k,o}|^2 = B$, and we recast (38) and (39) as follows

$$\frac{(1-\gamma_k)B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\lambda\gamma_k T}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T} \geq 0, \tag{40}$$

$$B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k \geq B\gamma_k + \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\lambda\gamma_k T, \tag{41}$$

After performing some algebraic manipulations and simplifications, (41) is transformed to

$$\gamma_k \leq \frac{B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k}{\left(B + (e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2})T\lambda\right)}, \tag{42}$$

which completes the proof.

PROOF FOR (p8b)

From (10), we know that

$$\tilde{\Psi}_k \leq \frac{\Gamma_k}{\gamma_k T} - \lambda. \tag{43}$$

Using (40), we have

$$\begin{aligned}&\frac{(1-\gamma_k)B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\lambda\gamma_k T}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T}\\ &\leq \frac{\Gamma_k - \lambda\gamma_k T}{\gamma_k T}.\end{aligned} \tag{44}$$

After performing some algebraic manipulations to (44) we get

$$(1-\gamma_k)B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k \leq \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k, \tag{45}$$

Solving (45) for γ_k gives

$$\gamma_k \geq 1 - \frac{\left[\bar{\Gamma}_k + \Gamma_k\right]\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)}{B}, \tag{46}$$

which completes the proof.

PROOF FOR (p8c)

From (P1d), we have

$$\tilde{\Psi}_k \leq \Psi_{\max} \tag{47}$$

Using (44), we have

$$\begin{aligned}&\frac{(1-\gamma_k)B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\lambda\gamma_k T}{\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T}\\ &\leq \Psi_{\max}.\end{aligned} \tag{48}$$

Performing some algebraic manipulations on (48), we get

$$\begin{aligned}B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k \leq \Psi_{\max}&\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\gamma_k T\\ &+ B\gamma_k + \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\lambda\gamma_k T.\end{aligned} \tag{49}$$

Simplifying (49) further, we get

$$\gamma_k \geq \frac{B - \left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)\bar{\Gamma}_k}{B + T\left[\lambda + \Psi_{\max}\right]\left(e^{\alpha_1\Psi_k}+e^{\alpha_1\alpha_2}\right)}, \tag{50}$$

which concludes the proof.

REFERENCES

[1] A. Mahmood et al., "Industrial IoT in 5G-and-beyond networks: Vision, architecture, and design trends," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4122–4137, Jun. 2022.

[2] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with RF energy harvesting: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 757–789, 2nd Quart., 2014.

[3] A. M. Seid, J. Lu, H. N. Abishu, and T. A. Ayall, "Blockchain-enabled task offloading with energy harvesting in multi-UAV-assisted IoT networks: A multi-agent DRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 12, pp. 3517–3532, Dec. 2022.

[4] C. Psomas et al., "Wireless information and energy transfer in the era of 6G communications," *Proc. IEEE*, early access, May 20, 2024, doi: 10.1109/JPROC.2024.3395178.

[5] S. Guo, Y. Shi, Y. Yang, and B. Xiao, "Energy efficiency maximization in mobile wireless energy harvesting sensor networks," *IEEE Trans. Mob. Comput.*, vol. 17, no. 7, pp. 1524–1537, Jul. 2018.

[6] D. Pradhan and K. Priyanka, "RF-energy harvesting (RF-EH) for sustainable ultra dense green network (SUDGN) in 5G green communication," *Saudi J. Eng. Technol.*, vol. 5, no. 6, pp. 258–264, 2020.

[7] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2082–2085, Dec. 2015.

[8] L. Shi, L. Zhao, and K. Liang, "Power allocation for wireless powered MIMO transmissions with non-linear RF energy conversion models," *China Commun.*, vol. 14, no. 2, pp. 57–64, Feb. 2017.

[9] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, Sep. 2011.

[10] Z. Ding, R. Schober, and H. V. Poor, "A new QoS-guarantee strategy for NOMA assisted semi-grant-free transmission," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7489–7503, Nov. 2021.

[11] 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, Aug. 2016.

[12] D. Wang et al., "From IoT to 5G I-IoT: The next generation IoT-based intelligent algorithms and 5G technologies," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 114–120, Oct. 2018.

[13] R. C. Daniels, C. M. Caramanis, and R. W. Heath, "Adaptation in convolutionally coded MIMO-OFDM wireless systems through supervised learning and SNR ordering," *IEEE Trans. Veh. Technol.*, vol. 59, no. 1, pp. 114–126, Jan. 2010.

[14] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.
[15] N. C. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.
[16] 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, Apr. 2013.
[17] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.
[18] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. New York, NY, USA, Springer, 2006.
[19] I. El Naqa and M. J. Murphy, *What is Machine Learning?*. Cham, Switzerland: Springer, 2015.
[20] M. Andrychowicz et al., "Hindsight experience replay," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.
[21] R. Jiang, K. Xiong, P. Fan, Y. Zhang, and Z. Zhong, "Optimal design of SWIPT systems with multiple heterogeneous users under non-linear energy harvesting model," *IEEE Access*, vol. 5, pp. 11479–11489, 2017.
[22] E. Boshkovska, R. Morsi, D. W. K. Ng, and R. Schober, "Power allocation and scheduling for SWIPT systems with non-linear energy harvesting model," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–6.
[23] P. Zeng, Q. Wu, and D. Qiao, "Energy minimization for IRS-aided WPCNs with non-linear energy harvesting model," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2592–2596, Nov. 2021.
[24] T.-T. Nguyen, V.-D. Nguyen, J.-H. Lee, and Y.-H. Kim, "Sum rate maximization for multi-user wireless powered IoT network with non-linear energy harvester: Time and power allocation," *IEEE Access*, vol. 7, pp. 149698–149710, 2019.
[25] T. A. Zewde and M. C. Gursoy, "NOMA-based energy-efficient wireless powered communications," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 679–692, Sep. 2018.
[26] L. Li, H. Xu, J. Ma, A. Zhou, and J. Liu, "Joint EH time and transmit power optimization based on DDPG for EH communications," *IEEE Commun. Lett.*, vol. 24, no. 9, pp. 2043–2046, Sep. 2020.
[27] 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, Sep. 2021.
[28] 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 *Proc. IEEE Globecom Workshops*, 2022, pp. 74–79.
[29] 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 Commun. Lett.*, vol. 12, no. 5, pp. 893–897, May 2023.
[30] Z. Wang, L. Ge et al., "Energy efficiency maximization strategy for sink node in SWIPT-enabled sensor-cloud based on optimal stopping rules," *China Commun.*, vol. 18, no. 1, pp. 222–236, Jan. 2021.
[31] S. Boyd et al., *Convex Optimization*. Cambridge, U.K.: Cambridge Univ., 2004.
[32] D. Silver, G. Lever, N. Heess et al., "Deterministic policy gradient algorithms," in *Proc. ICML*, 2014, pp. 387–395.
[33] S. Y. Seidel and T. S. Rappaport, "914 MHz path loss prediction models for indoor wireless communications in multifloored buildings," *IEEE Trans. Antennas Propag.*, vol. 40, no. 2, pp. 207–217, Feb. 1992.
[34] Y. Wang, Y. Wang, F. Zhou, Y. Wu, and H. Zhou, "Resource allocation in wireless powered cognitive radio networks based on a practical non-linear energy harvesting model," *IEEE Access*, vol. 5, pp. 17618–17626, 2017.
[35] Z. Wang, T. Lv, and W. Li, "Energy efficiency maximization in massive MIMO-NOMA networks with non-linear energy harvesting," in *Proc. IEEE WCNC*, 2021, pp. 1–6.

SYED ASAD ULLAH (Student Member, IEEE) received the B.S. degree in electronic engineering from the Baluchistan University of Information Technology Engineering and Management Sciences (BUITEMS), Quetta, and the master's degree in electrical engineering from the National University of Sciences and Technology, Islamabad, Pakistan, where he is currently pursuing the Ph.D. degree. His pioneering research at the Information Processing and Transmission Lab is instrumental in shaping the future of IoT and wireless communication networks while specializing in deep reinforcement learning (DRL)-based optimization and energy-efficient communication protocols for wireless communications. Before embarking on his doctoral journey, he spent five years as a Lecturer with the Electronic Engineering Department, BUITEMS. His academic journey, enriched by extensive teaching experience, innovative research, and active involvement in academic events, highlights his significant impact on the field of wireless communications and his unwavering commitment to academia. His research contributions are substantial, with numerous publications in high-impact IEEE journals and conferences, focusing on DRL for optimizing wireless communication systems. His work spans areas such as non-orthogonal multiple access in IoT networks, vehicular network resource optimization, and energy harvesting technologies. He is also certified as a Microsoft Specialist. His exceptional contributions at BUITEMS earned him the 'Best Faculty Member' Award twice.

AAMIR MAHMOOD (Senior Member, IEEE) received the B.E. degree in electrical engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2002, and the M.Sc. and D.Sc. degrees in communications engineering from the School of Electrical Engineering, Aalto University, Espoo, Finland, in 2008 and 2014, respectively, where he was a Visiting Researcher from 2015 to 2016. He worked as a Research Intern with Nokia Research Center, Helsinki, Finland, in 2014, and a Postdoctoral Researcher with Mid Sweden University, Sundsvall, Sweden, from 2016 to 2018, where he has been an Associate Professor with the Department of Computer and Electrical Engineering, since 2023. His research interests include industrial IoT, 5G-TSN integration, AI/ML for radio network optimization and management, RF interference and coexistence management, network time synchronization, and wireless positioning.

ALI ARSHAD NASIR (Member, IEEE) received the Ph.D. degree in telecommunications engineering from the Australian National University (ANU), Australia, in 2013, where he worked as a Research Fellow from 2012 to 2015. From 2015 to 2016, he was an Assistant Professor with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Pakistan. He joined the Department of Electrical Engineering, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 2016, where he is currently working as an Associate Professor. His research interests are in the area of signal processing in wireless communication systems. He has served as an Editor for IEEE WIRELESS COMMUNICATION LETTERS from 2022 to 2024. He has been an Editor of IEEE COMMUNICATION LETTERS since January 2024.

MIKAEL GIDLUND (Senior Member, IEEE) received the Licentiate of Engineering degree in radio communication systems from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2004, and the Ph.D. degree in electrical engineering from Mid Sweden University, Sundsvall, Sweden, in 2005. From 2008 to 2015, he was a Senior Principal Scientist and Global Research Area Coordinator of Wireless Technologies with ABB Corporate Research, Västeras, Sweden. From 2007 to 2008, he was a Project Manager and a Senior Specialist with Nera Networks AS, Bergen, Norway. From 2006 to 2007, he was a Research Engineer and a Project Manager with Acreo AB, Hudiksvall, Sweden. Since 2015, he has been a Professor of Computer Engineering with Mid Sweden University. He holds more than 20 patents (granted and pending) in the area of wireless communication. His current research interests include wireless communication and networks, wireless sensor networks, access protocols, and security. He is an Associate Editor of the *IEEE Transactions on Industrial Informatics.*

SYED ALI HASSAN (Senior Member, IEEE) received the B.E. degree in electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2004, the first M.S. degree in mathematics from Georgia Tech in 2011, the second M.S. degree in electrical engineering from the University of Stuttgart, Germany, in 2007, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, USA, in 2011. His research interests include signal processing for communications with a focus on cooperative communications for wireless networks, stochastic modeling, estimation and detection theory, and smart grid communications. He is currently working as a Professor with the School of Electrical Engineering and Computer Science, NUST. He was a Visiting Professor with Georgia Tech in 2017. He also held industry positions with Cisco Systems Inc., CA, USA, and the Center for Advanced Research in Engineering, Islamabad.