Received 1 May 2024, accepted 15 June 2024, date of publication 24 June 2024, date of current version 9 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3417935

RESEARCH ARTICLE

Joint Optimization of Packet Scheduling and Energy Harvesting for Energy Conservation in D2D Networks: A Decentralized DRL Approach

SENGLY MUY[1], **EUN-JEONG HAN**[1], **AND JUNG-RYUN LEE**[1,2], **(Senior Member, IEEE)**
[1]School of Intelligent Energy and Industry, Chung-Ang University, Seoul 06974, South Korea
[2]School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Jung-Ryun Lee (jrlee@cau.ac.kr)

This work was supported in part by Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE), Republic of Korea, under Grant 20214000000280; in part by the National Research Foundation of Korea (NRF) Grant funded by Korean Government through MSIT under Grant NRF-RS-2023-00251105; and in part by the Chung-Ang University Research Scholarship Grants, in 2021.

ABSTRACT This study investigates the optimization of proportional fair (PF) and energy efficiency in simultaneous wireless information and power transfer (SWIPT)-based device-to-device (D2D) networks considering the residual battery levels of D2D users to increase the network lifetime. We establish an optimization model that determines the subchannel allocation and transmission power levels for D2D users, to maximize an objective function that combines user fairness and energy efficiency. To tackle this problem in a distributed manner, we propose a multi-agent deep reinforcement learning (DRL) model. Given that fairness considerations necessitate information about other agents, we employ the long short-term memory (LSTM) algorithm to estimate the parameters of other D2D pairs within the state space of the multi-agent DRL model. Through simulations, we compare the performance of our proposed algorithm with that of existing iterative algorithms, namely, exhaustive search (ES) and gradient search (GS). The results demonstrate that the proposed multi-agent DRL approach achieves a solution that is nearly globally optimal, while maintaining a lower computational complexity due to the parallel computing of multi-agent DRL. Furthermore, the proposed algorithm reduces the standard deviation of residual battery levels among D2D pairs and contributes to an increased network lifetime.

INDEX TERMS Distributed D2D, proportional fair, energy efficiency, joint optimization, multi-agent DRL.

I. INTRODUCTION

Increasing demands of next generation networks for higher transmission rates, energy efficiency, cellular coverage, and spectral efficiency have triggered the emergence of a device-to-device (D2D) networks, which enables more energy-efficient communication through direct communication among mobile nodes, [1], [2], [3]. Recently, researchers have been investigating D2D networks with SWIPT functionality, as these can improve the spectral efficiency and throughput by letting D2D devices communicate with each other while simultaneously harvesting energy from other nodes. Wireless networks feature a well-known trade-off between the total throughput (i.e., energy efficiency) and user fairness; it has been controlled using various scheduling schemes such as the proportional fair (PF) scheduling algorithm. The PF scheduling algorithm manages this trade-off in a way to maximize the sum of the logarithmic average received data rates of users, [4]. For improving the D2D network performance, it is critical to not only improve the energy efficiency of specific D2D nodes but also maximize network lifetime, because the latter ensures efficient use of

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.

limited wireless resources, maintains the quality of service (QoS) of D2D users, and provides network scalability by supporting more D2D users over time.

In recent years, there has been a lot of research into improving the performance of SWIPT-based D2D networks. In [5], the joint optimization of the transmit power and power-splitting coefficients was proposed for SWIPT-based multi-user in distributed networks. In [6], the research focused on enhancing D2D communication through the integration of full-duplex (FD) relaying, simultaneous wireless information and power transfer (SWIPT), and non-orthogonal multiple access (NOMA), demonstrating superior efficiency and connectivity compared to classical SWIPT-FD-orthogonal multiple access (SWIPT-FD-OMA) methods. In [7], a game-theoretic model for D2D power allocation with SWIPT is introduced, offering mechanisms and strategies that boost energy efficiency and user mobility responsiveness. In [8], the study explored SWIPT-enhanced mode selection for D2D communications through stochastic geometry, presenting an energy-efficient methodology that surpasses traditional mode selection strategies, particularly in ultra-dense cellular environments. However, these advancements leverage novel strategies to effectively address the challenges inherent in SWIPT-based D2D networks.

Moving forward, machine learning has gained popularity among researchers as an intelligent solution to wireless network challenges. In [9], the authors highlight the crucial role of machine learning in advancing next-generation wireless networks, achieving high data rates, and supporting new applications through intelligent adaptive learning across various 5G technologies. The authors of [10] proposed a deep reinforcement learning (DRL)-based resource allocation strategy for wireless sensor networks (WSNs) aimed at maximizing throughput through optimized power and time management, which outperformed conventional policies. In [11], the authors presented a DRL-based optimization for wireless powered IIoE networks, focusing on age-energy efficiency by optimizing device scheduling and power, which significantly outperformed simulation benchmarks. The authors of [12] proposed the multi-agent DRL model to optimize the dynamic radio resource allocation in a distributed system. In [13], the multi-agent DRL framework was developed to optimize the load-aware distributed resource allocation.

Therefore, we develop a PF scheduling problem for a SWIPT-based D2D network. Our problem differs from the conventional PF scheduling problems in that it considers not only the trade-off between user throughput and fairness but also the effect of the energy harvesting functionality of D2D users on scheduling by considering the residual battery lives of D2D devices for prolonging the network lifetime. It is noted that PF scheduling operate in a centralized way where each user provides the central coordinator with information necessary for scheduling, and the central coordinator collects this information and controls the amount of user data transmitted at each time. However, using a centralized algorithm is not suitable in the D2D networks owing to the increased computational load imposed on each D2D device having low computing capacity, consequent energy consumption for D2D users with limited battery life, and increased signaling overhead in the network. Furthermore, previous PF scheduling problems only required information about the channel quality of each user and the amount of data transmitted during a given past interval. By contrast, the our PF scheduling problem needs additional information, such as the energy harvesting ratio and residual energy level of each user, which inevitably increases algorithm complexity. To solve this problem, we propose a decentralized machine learning algorithm considering its low computation complexity.

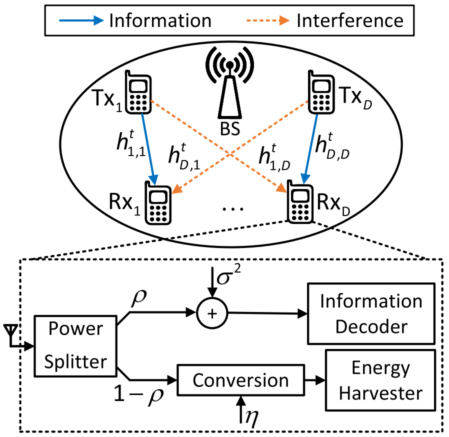


FIGURE 1. System model.

The contributions of this study can be summarized as follows. First, we develop the optimization of packet scheduling, energy efficiency, and network life-time maximization in distributed SWIPT-based D2D networks. After that, we focus on applying DRL to optimize energy efficiency and PF in D2D communication networks. Considering the impracticality of centralized DRL for requiring a single agent to collect entire information in D2D communication networks, we propose integrating long short-term memory (LSTM) networks into the multi-agent DRL design. Here, LSTM is employed for each agent to estimate important information about the other agent, such as channel gain, transmit power, and subchannel allocation. Despite the limitations in information availability, this predictive capability provided each agent with critical knowledge for decision-making. Finally, we simulate our proposed decentralized DRL framework with LSTM by comparing its performance with the gradient search (GS) and exhaustive search (ES). Results indicate that our proposed method significantly improved the performance in terms of optimality (energy efficiency and PF) and real time implementation (low computation complexity) for SWIPT-based D2D networks.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

In this study, we consider a D2D communication network consisting of D transmitter (Tx) and receiver (Rx) pairs with index $i \in \mathcal{D} = \{1, 2, \ldots, D\}$, where all D2D pairs are deployed randomly within the coverage area of a base station (BS), as shown in Fig. 1. The effective channel between Tx i and Rx j at time slot t is denoted as $\left|h_{i,j}^{t}\right|^{2}$, and the channel gain $h_{i,j}^{t}$ is assumed to be an independently and identically distributed (i.i.d.) Rician random variable with mean $\mu_{i,j}$. The communication channel is divided into L equally distributed subchannels with index $l \in \mathcal{L} = \{1, 2, \ldots, L\}$, and $q_{i,l}^{t}$ is the channel allocation indicator where $q_{i,l}^{t} = 1$ if the subchannel l is allocated to the Tx i at time slot t, and $q_{i,l}^{t} = 0$ otherwise.

Let B and p_{i}^{t} denote the maximum battery capacity, and the transmitted power of Tx i at time slot t, respectively. Then, the received signal to noise ratio (SINR) for D2D pair i is given as

$$\Gamma_{i}^{t} = \frac{\rho p_{i}^{t}\left|h_{i,i}^{t}\right|^{2}}{\sigma^{2} + \rho\left(\sigma_{A}^{2} + I_{i',i}^{t}\right)}, \tag{1}$$

where σ^2 and σ_A^2 represent the antenna noise and base-band noise power at the Rx, respectively. And, ρ is the energy conversion ratio. Here, $I_{i',i}^{t}$ represents the interference from other Tx i' during time slot t, and it is expressed as

$$I_{i',i}^{t} = \sum_{l\in\mathcal{L}} \sum_{i'\in\mathcal{D}\setminus\{i\}} q_{i',l}^{t} p_{i'}^{t} \left|h_{i',i}^{t}\right|^{2}. \tag{2}$$

By using the Shannon capacity, the data rate of D2D pair i during time slot t can be written as

$$\mathrm{DR}_{i}^{t} = \sum_{l\in\mathcal{L}} q_{i,l}^{t} \log_2\left(1 + \Gamma_{i}^{t}\right). \tag{3}$$

Furthermore, the average data rate of D2D pair i during time window T is expressed as

$$\overline{\mathrm{DR}}_{i}^{t} = \begin{cases} \frac{1}{t-1}\sum_{\tau=1}^{t-1} \mathrm{DR}_{i}^{\tau}, & t < T, \\ \frac{1}{T}\sum_{\tau=t-T}^{t-1} \mathrm{DR}_{i}^{\tau}, & t \geq T. \end{cases} \tag{4}$$

To ensure network fairness, we formulate the PF scheduling function so that user fairness is built into the objective function using a sum-of-logarithmic [4]. The logarithmic function prioritizes equitable resource distribution by adjusting user throughput or data rates, balancing total network throughput with user fairness, and aligning with the principle of proportional fairness in network resource allocation algorithms. The PF scheduling function can be formulated as follows:

$$\mathrm{PF}^{t} = \sum_{i\in\mathcal{D}} \log_2 \overline{\mathrm{DR}}_{i}^{t}. \tag{5}$$

The energy consumption of D2D Tx i during time window T is given by

$$\mathrm{EC}_{i}^{t} = \sum_{\tau=1}^{t} P_C + p_{i}^{\tau}, \tag{6}$$

where P_C is the power consumption in the circuit. The total energy harvested at Rx i from all Txs during time window T is given by

$$\mathrm{EH}_{i}^{t} = \sum_{\tau=1}^{t}\sum_{j\in\mathcal{D}} (1-\rho)\,\eta p_{j}^{\tau}\left|h_{j,i}^{\tau}\right|^{2}. \tag{7}$$

Then, the residual energy of D2D pair i at the time slot t can be calculated as

$$\mathrm{ER}_{i}^{t} = \min\left(\max\left(\mathrm{ER}_{i}^{t-1} - \mathrm{EC}_{i}^{t-1}, 0\right) + \mathrm{EH}_{i}^{t-1}, B\right). \tag{8}$$

Finally, we can calculate the total residual energy of the system at time slot t as following

$$\mathrm{ERT}^{t} = \sum_{i\in\mathcal{D}} \mathrm{ER}_{i}^{t}. \tag{9}$$

B. PROBLEM FORMULATION

From the definitions of PF^t and ERT^t, we define the objective function so that it considers both proportional fairness and energy efficiency with the residual battery, which is given by

$$f\left(\boldsymbol{p}^{t}, \boldsymbol{q}^{t}\right) = \frac{\mathrm{PF}^{t}\left(\boldsymbol{p}^{t}, \boldsymbol{q}^{t}\right)}{B - \mathrm{ERT}^{t}\left(\boldsymbol{p}^{t}, \boldsymbol{q}^{t}\right)}, \tag{10}$$

where the transmitted power vector is $\boldsymbol{p}^{t} = \left\{p_1^t, p_2^t, \ldots, p_D^t\right\}$. The subchannel allocation indicator $\boldsymbol{q}^t$ is defined by the following matrix:

$$\boldsymbol{q}^{t} = \begin{bmatrix} q_{1,1}^{t} & \cdots & q_{1,L}^{t} \\ \vdots & \ddots & \vdots \\ q_{D,1}^{t} & \cdots & q_{D,L}^{t} \end{bmatrix}. \tag{11}$$

Then, our target becomes to find $\boldsymbol{p}^t$ and $\boldsymbol{q}^t$ that maximize the objective function. Therefore, the optimization problem can be formulated as

$$\begin{aligned} \max_{\boldsymbol{p}^{t},\boldsymbol{q}^{t}} \quad & f\left(\boldsymbol{p}^{t}, \boldsymbol{q}^{t}\right) \\ \text{s.t.} \quad & C1: 0 \leq p_{i}^{t} \leq P_{\max}, \\ & C2: q_{i,l}^{t} \in \{0, 1\}, \\ & C3: \overline{\mathrm{DR}}_{i}^{t} \geq \mathrm{DR}_{\min}, \\ & C4: \mathrm{ER}_{i}^{t} \leq B, \\ & \text{for } i \in \mathcal{D}, \text{ and } l \in \mathcal{L}, \end{aligned} \tag{12}$$

where $P_{\max}$ is the maximum transmission power and $\mathrm{DR}_{\min}$ is the minimum data rate for guaranteeing the QoS.

III. PROPOSED MULTI-AGENT DRL WITH LSTM

In real-world implementation, centralized control would be impractical due to environmental scale or privacy concerns. Therefore, we propose a multi-agent DRL model, as illustrated in Fig. 2, to determine the optimal variables $\boldsymbol{p}$ and $\boldsymbol{q}$. The proposed multi-agent DRL model's decentralized design makes it more scalable, adaptable to network changes, and practical for real-world use. It is noticed that all agents in our proposed multi-agent DRL are processing in parallel. In the context of multi-agent DRL, the optimization problem can be formulated as a Markov decision process (MDP). As defined in [14], the MDP is represented by the tuple

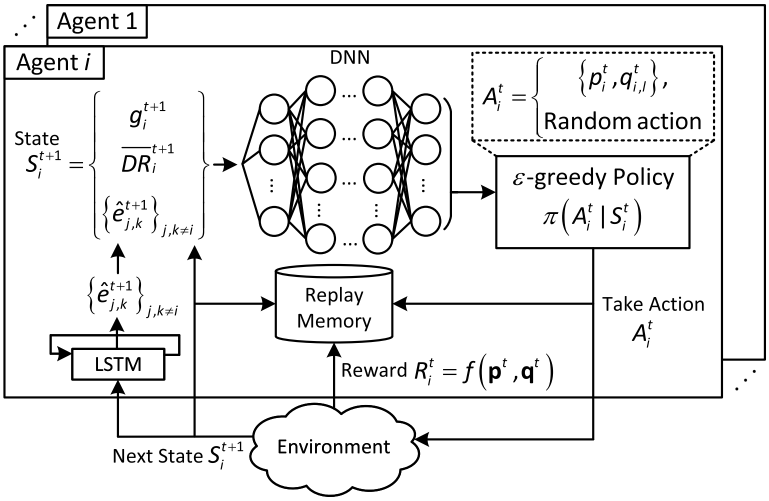

Agent 1
Agent i
DNN
State $S_i^{t+1} = \left\{ g_i^{t+1}, \overline{DR}_i^{t+1}, \left\{ \hat{e}_{j,k}^{t+1} \right\}_{j,k\neq i} \right\}$
$A_i^t = \left\{ \{p_i^t, q_{i,l}^t\}, \text{Random action} \right.$
ε-greedy Policy
$\pi\left(A_i^t \mid S_i^t\right)$
$\left\{ \hat{e}_{j,k}^{t+1} \right\}_{j,k\neq i}$
Replay Memory
LSTM
Take Action A_i^t
Reward $R_i^t = f\left(\mathbf{p}^t, \mathbf{q}^t\right)$
Next State S_i^{t+1}
Environment

FIGURE 2. The proposed Multi-Agent DRL model.

$\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P} \rangle$, where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R}$ respectively denote the finite sets of states, actions, and reward functions for each agent, and $\mathcal{P}$ represents the transition probability from the current state $S_i^t \in \mathcal{S}$ to the next state $S_i^t \in \mathcal{S}$ when following the policy $P^{\pi}\left(S_i^{t+1} \mid S_i^t\right)$. Specifically, $P\left(S_i^{t+1} \mid S_i^t, A_i^t\right)$ signifies the transition probability from the current state S_i^t to the subsequent state S_i^{t+1} given action A_i^t, and $\pi\left(A_i^t \mid S_i^t\right)$ represents a mapping (or policy) from the current state S_i^t to the action A_i^t. Therefore, $P^{\pi}\left(S_i^{t+1} \mid S_i^t\right)$ is essentially the transition probability $P\left(S_i^{t+1} \mid S_i^t, A_i^t\right)$ weighted by the policy $\pi\left(A_i^t \mid S_i^t\right)$. The primary goal of the MDP is to identify the optimal policy π^* that maximizes the reward function $\mathcal{R}$.

A. MULTI-AGENT DRL DESIGN

In the proposed algorithm, each D2D pair is assumed to be adapted with an agent; therefore, the system consists of D agents, and each agent is required to observe the information required to take an action by controlling the transmit power and subchannel allocation. Subsequently, the agent is rewarded from the objective function obtained by the channel conditions, interference levels, and QoS specifications of the D2D pairs. The actions of an agent may influence the decisions of other agents, resulting in the coordinated and efficient operation of the overall system in real time. The state, action, and reward of the proposed multi-agent DRL model are designed as following.

- **State**: The state space of agent i at time slot t is defined as

$$S_i^t = \left\{ g_i^t, \hat{\boldsymbol{e}}_i^t, \overline{\mathrm{DR}}_i^t \right\}, \tag{13}$$

where $g_i^t = \left| h_{i,i}^t \right|^2$ is the channel gain of D2D pair i, $\overline{\mathrm{DR}}_i^t$ is the average data rate of D2D pair i at time slot t during time window T, and $\hat{\boldsymbol{e}}_i^t = \left\{ \hat{e}_{j,k}^t \right\}_{j,k\neq i}$ is the estimated channel gain information of the other D2D pairs.

- **Action**: To optimize the objective function, each agent needs to adjust its transmission power and the subchannel allocation. The action space of agent i at time slot t is defined as the set of the transmit power and subchannel allocation indicator given by

$$A_i^t = \left\{ p_i^t, q_{i,l}^t \right\}. \tag{14}$$

It is noted that $p_i^t \in \left\{ 0, \frac{p_{\max}}{M-1}, \frac{2p_{\max}}{M-1}, \ldots, p_{\max} \right\}$, where M is the number of discrete levels of the maximum transmission power, and $q_{i,l}^t \in \{0, 1\}, \forall l \in \{1, 2, \ldots, L\}$ in the subchannel allocation indicator.

- **Reward**: In DRL, the reward function connects an agent to its environment by evaluating its actions at each step, helping it associate positive or negative outcomes, and enabling better decision-making to achieve the maximum accumulated reward. In this study, the reward function is defined as the objective function while satisfying all constraints. In this case, the reward function of agent i at time slot t is given by

$$R_i^t = f\left(\boldsymbol{p}^t, \boldsymbol{q}^t\right) = \frac{\mathrm{PF}^t\left(\boldsymbol{p}^t, \boldsymbol{q}^t\right)}{B - \mathrm{ERT}^t\left(\boldsymbol{p}^t, \boldsymbol{q}^t\right)}. \tag{15}$$

When one or more constraints are not satisfied, the reward given is negative.

- **Deep Q-Network (DQN)**: A common strategy in DRL is to employ a Q-value function, that estimates the expected cumulative reward for taking a specific action in a given state. And, the updated Q-value function can be expressed as, [15],

$$Q^{\mathrm{new}}\left(S_i^t, A_i^t\right) = (1-\alpha)\, Q^{\mathrm{old}}\left(S_i^t, A_i^t\right) + \alpha \left[R_i^t + \gamma \max_{A_i \in \mathcal{A}} Q^{\mathrm{old}}\left(S_i^{t+1}, A_i\right) \right], \tag{16}$$

where $\alpha > 0$ is the learning rate, and $0 \leq \gamma \leq 1$ is the discount factor. Owing to the large size of the state space, DQN are used to estimate the Q-value function. Furthermore, when the replay memory is full, the data will be split into multiple minibatch samples to train a DQN so as to ensure diverse, representative, and efficiently processed training data.

- Policy: The ϵ-greedy policy is a strategy for selecting actions based on Q-values to improve the balance between exploration and exploitation during the decision-making process for each agent. The action selection mechanism, which is based on a DQN with weights θ to approximate the Q-value function, can be defined as follows:

$$A_i^{t+1} = \begin{cases} \underset{a}{\operatorname{argmax}}\, Q\left(S_i^t, a, \theta\right), & \text{with Prob. } 1-\epsilon, \\ \text{Random action}, & \text{with Prob. } \epsilon. \end{cases} \tag{17}$$

This approach ensures that while the algorithm primarily exploits known information about the environment to make decisions that maximize immediate rewards, it also explores alternative actions with a probability (Prob.) of ϵ, thereby avoiding convergence to sub-optimal solutions.

B. THE PROPOSED MULTI-AGENT DRL WITH LSTM

In the proposed multi-agent DRL, a D2D pair cannot easily obtain channel information of other D2D pairs in real time, making multi-agent DRL decision-making challenging. To address this issue, we apply LSTM algorithm, a type of recurrent neural network (RNN) architecture, to estimate long-term dependencies in the sequence data of channel gains in the other D2D pairs. Here, RNN is particularly useful in situations where past information needs to be remembered and taken into account when making predictions or decisions. The architecture of the proposed multi-agent DRL algorithm with LSTM is illustrated in Fig. 2. The number of inputs in the sequence layer is equal to the number of input data, and the size of the fully connected layer is equal to the number of responses. Moreover, the number of outputs in the output layer is equal to the total number of channel gains from other D2D pairs, $\hat{\mathbf{e}}_i^t = \left\{\hat{e}_{j,k}^t\right\}_{j,k\neq i}$.

Results in Figs. 3 and 4 validate the LSTM's effectiveness in accurately predicting transmit power and subchannel allocation. These graphs demonstrate the LSTM's ability to estimate the transmit power and subchannel allocation for an agent based on the observed data from a previous time slot related to other agents. This similarity in pattern changes between the LSTM's predictions and the actual data indicates the model's accuracy in estimating transmit power and subchannel allocation. This indicates that the LSTM model is effective in capturing the underlying patterns and dynamics of the system, making it a reliable tool for optimizing resource allocation in wireless communication networks.

The operational procedure of the proposed multi-agent DRL algorithm with LSTM is as follows. Each D2D pair (i.e., an agent) controls the current transmission power $\vec{p}^{(t_C)}$, and current subchannel allocation $s^{(t_C)}$ of the D2D transmitter. Each agent is assumed to send information necessary for distributed computing, including the transmit power, subchannel allocation, data rate, and channel gain between the D2D transmitter and D2D receiver with the BS during time window T. Then, the BS broadcasts this information to all D2D agents. Because an agent cannot obtain the necessary information for every time slot t but just one time slot during time window T, the channel gains of other D2D pairs are estimated in each D2D agent by using LSTM. Then, each agent calculates the reward and determines its own transmission power $\vec{p}^{(t_C+1)}$ and subchannel allocation $s^{(t_C+1)}$ in every time t by using the estimated state information of other D2D pairs obtained from LSTM. At the next time slot $t+1$, each agent observes the new state and reward from the environment and then takes an action (control transmission power and subchannel allocation) by using the ϵ-greedy policy.

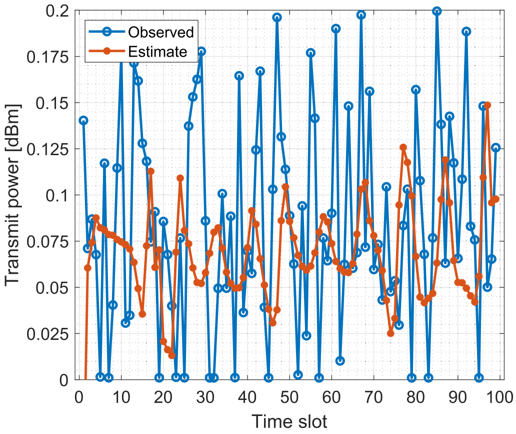


FIGURE 3. Transmit power prediction.

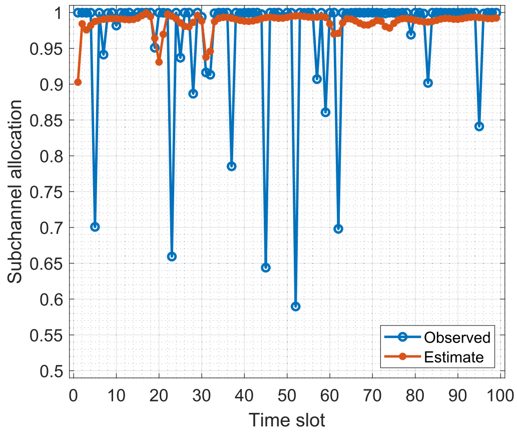


FIGURE 4. Subchannel allocation prediction.

IV. TIME COMPLEXITY ANALYSIS

In this study, we employ the following two existing iterative-based optimization algorithms, such as ES and GS, in order to compare the performance of the proposed multi-agent DRL algorithm. The following represents the time complexity analysis for ES, GS, and the proposed multi-agent DRL.

- **Exhaustive search (ES)** is a type of global optimization algorithm that identifies the global solution by examining every possible case. In employing ES algorithm to tackle a given problem, quantizing the control variables is essential. This process enables a thorough exploration of all potential variable permutations. In our study, the transmit power $\boldsymbol{p}$, is quantized into M equal levels, and the subchannel allocation indicator $\boldsymbol{q}$, into L equal levels, with both $\boldsymbol{p}$ and $\boldsymbol{q}$ operating within a D-dimensional space. Consequently, the total number of viable combinations for evaluation is $O\left((M \times L)^D\right)$, which represents the computational complexity of the ES algorithm.

- **Gradient search (GS)** is a technique for discovering locally optimal solutions. In GS algorithm, solutions are iteratively approached by advancing in increments defined by the learning rate, which are in turn directed by the gradient of the objective function. This process continues until the magnitude of the error is reduced to less than a specified error tolerance, denoted as ϵ. According to the studying in [16], the computational complexity of the GS algorithm can be analytically expressed as $O\left(\epsilon^{-2}\right)$. This denotes that the complexity increases inversely with the square of the error tolerance, highlighting a fundamental trade-off between computational demand and the precision of the solution obtained.
- **Proposed multi-agent DRL algorithm**: In the proposed multi-agent DRL algorithm, each agent is equipped with a DQN designed to approximate the Q-value function. The architecture of DQN consists of an input layer, H hidden layers forming a fully connected network, and an output layer, employing the ReLU activation function. The number of neurons in the input layer corresponds to the dimensionality of the state space, while the number of neurons in each hidden layer is equal to the number of neurons in the output layer. The dimensionality of this output layer is determined by the number of quantized actions $(M \times L)$ available to each agent. Given that the neuron count in the output layer exceeds that of the input layer, the time complexity of the proposed multi-agent DRL algorithm can be articulated as $O\left(H \times M \times L\right)$, [17]. It is noted that this time complexity is calculated for the deployment phase of the proposed multi-agent DRL (not the training phase), where the processing of computing the LSTM is constant.

In general, the ES algorithm can achieve the global optimal result because it exhaustively searches all possible solutions; however, the complexity rises exponentially as level D is quantized. Furthermore, the GS algorithm complexity is small due to the square of the tolerable ϵ; however, it does not guarantee the global optimal. On the other hand, our proposed multi-agent DRL only requires deep neural network computation, which is very time-consuming and suitable for real-time implementation. As shown in Table 1, the time complexity analysis of our proposed multi-agent DRL is more efficient and scalable compared to ES and GS. This makes it a promising approach for real-time applications where computational resources are limited. Additionally, the multi-agent aspect of the DRL can lead to more robust and efficient solutions compared to single-agent algorithms.

TABLE 1. Computational complexity comparison.

Algorithms	Computation Complexity
ES	$O\left((M \times L)^D\right)$
GS	$O\left(\epsilon^{-2}\right)$
Proposed multi-agent DRL	$O\left(H \times M \times L\right)$

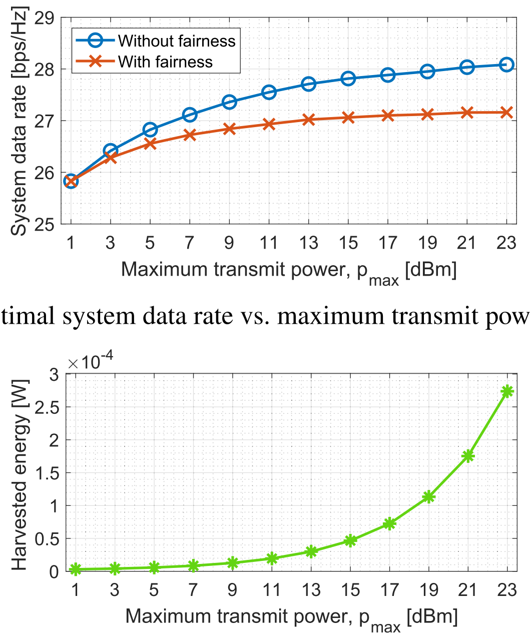
TABLE 2. Simulation parameters.

Parameters	Values
Number of subchannels, L	3
Number of D2D pairs, D	6
Energy conversion efficiency [18], η	0.5
Distance between D2D pairs	10m-20m
Energy consumption of circuit, P_C	20dBm
Rician factor	5dB
Base-band noise power spectrum, σ^2	-70dBm
AWGN power spectrum, σ_A^2	-100dBm
Maximum transmission power, $P_{\max}$	1dBm-23dBm
Window size, T	5
Path-loss exponent	3.6
Power splitting ratio, ρ	0.5
Battery size, B	10dB

V. PERFORMANCE EVALUATION

For performance evaluation, we consider a scenario with three subchannels and six D2D pairs. The energy conversion efficiency η is set to 0.5, [18]. D2D pairs are spatially distributed according to a normal distribution, with an average direct link distance of 10m and an interference link distance of 20m. The constants for the energy consumption of the circuit, baseband noise power spectrum, and additive white Gaussian noise power spectrum are set to 20dBm, -70dBm, and -100dBm, respectively. The path loss and Rician small-scale fading gain are 3.6 dB and 5dB, respectively. The power-splitting ratio and battery size are held constant at 0.5 dB and 10 dB, respectively. The simulation parameters are summarized in Table 2.

Fig. 5 illustrates the optimized system data rate and the energy harvesting in terms of increasing the maximum transmit power. Here, Fig. 5a shows the optimal system data rates with and without considering the PF scheduling function as a function of the maximum transmit power. When the PF scheduling function is not considered, the numerator of the objective function becomes the sum of the received data rate, and the system data rate increases as the maximum transmit power increases. Moreover, we can also verify the trade-off relation between fairness and data rate; specifically, if a communication system prioritizes fairness, the overall system data rate can be lower, and vice versa. Additionally, Fig. 5b shows the total harvested energy under the proposed algorithm as the transmission power $P_{\max}$ increases. This result shows that the harvested energy increases exponentially as $P_{\max}$ increases. Figs. 5a and 5b show that as maximum transmit power increases, system data rate increases logarithmically, and energy harvesting increases exponentially. This result indicates a strategy to limit the maximum transmission power of a D2D transmitter



a. Optimal system data rate vs. maximum transmit power.

b. Energy harvested vs. maximum transmit power.

FIGURE 5. **Performance results for various maximum transmit power levels.**

with a high residual battery to be determined only by its data rate requirements.

Fig. 6 shows the objective function obtained by ES, the proposed multi-agent DRL with and without LSTM, and GS. The graph shows that our proposed multi-agent DRL with LSTM achieves up to 98% of the global optimum that is obtained by ES. It is noted that the overall performance of the proposed algorithm is evaluated from both the performance of ES as an upper bound and the performance of GS serving as a benchmark for suboptimal performance. It also demonstrates that the result obtained by the proposed multi-agent DRL is very close to that obtained by ES, indicating that the former achieves a near-global-optimal solution and outperforms the optimization based iteration method obtained by GS. Moreover, we observe that the proposed multi-agent DRL with LSTM obtains a higher performance than that of the proposed multi-agent DRL without LSTM, thus verifying the effectiveness of the LSTM network for estimating other D2D pairs.

Through the simulation, we also evaluated the energy conservation for our objective function with and without considering the residual battery of devices in the D2D network, as shown in Fig. 7. It is notice that the average residual battery without considering the residual battery is obtained using the object function, which is given as follows:

$$f\left(\boldsymbol{p}^t,\boldsymbol{q}^t\right)=\frac{\mathrm{PF}^t\left(\boldsymbol{p}^t,\boldsymbol{q}^t\right)}{\sum_{i\in\mathcal{D}}\mathrm{EC}_i^t\left(\boldsymbol{p}^t,\boldsymbol{q}^t\right)-\sum_{i\in\mathcal{D}}\mathrm{EH}_i^t\left(\boldsymbol{p}^t,\boldsymbol{q}^t\right)}. \quad (18)$$

The graph in Fig. 7 shows that the average residual battery of D2D pairs is enhanced if our proposed objective function considering residual battery is used.

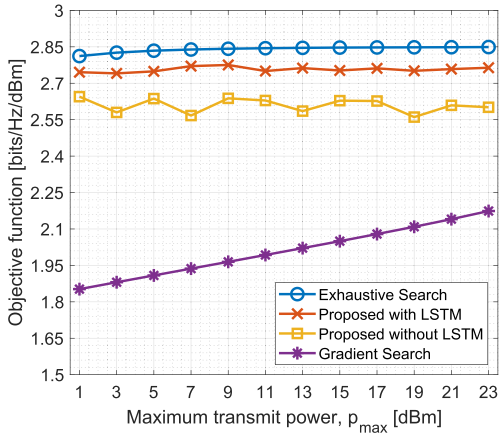


FIGURE 6. **Optimal objective function vs. maximum transmit power.**

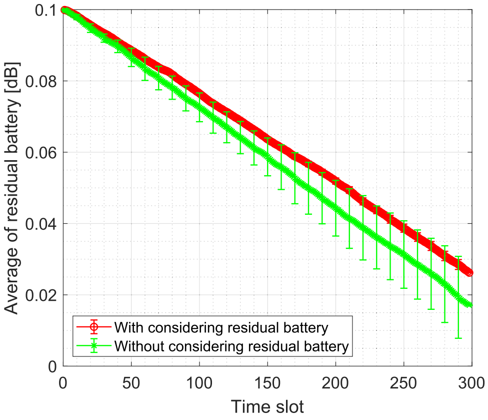


FIGURE 7. **Average residual battery of D2D pairs with and without considering residual battery in target function.**

TABLE 3. **Comparison for network lifetime with/without considering residual battery.**

Considerations	Network lifetime (Number of time slot)
With considering residual battery	607
Without considering residual battery	411

TABLE 4. **Standard deviation of residual battery of D2D pairs.**

Considerations	Average standard deviation
With considering residual battery	9.36×10^{-5}
Without considering residual battery	0.78×10^{-3}

The result in Table 3 indicates that the average residual battery of D2D pairs is prolonged when the proposed algorithm is applied, which shows the network lifetime improvement. And Table 4 shows that the average standard deviation of the data rate. It is seen that the average standard deviation of data rate when considering the residual battery is much smaller than that when not considering the residual battery, thus indicating the fairness of the residual battery among D2D pairs.

VI. CONCLUSION

In this study, we investigated an optimization problem by considering PF scheduling and energy efficiency considering residual batteries in SWIPT-based D2D networks. To solve this problem in a distributed manner, a multi-agent DRL model that can determine the best transmission power and subchannel allocation indicator in a way to maximize the reward function is proposed. To enhance the performance of the proposed algorithm, an LSTM network that estimates the states of other agents is applied to the proposed multi-agent DRL model. The use of LSTM was found to enhance the performance of the proposed multi-agent DRL. Simulation results showed that the proposed algorithm outperformed GS and achieved a near-global optimal solution with lower time complexity. In addition, the average residual battery of D2D pairs and network lifetime increased, and the fairness of the residual battery among D2D pairs was enhanced by considering the residual battery in the optimization model. For our future work, we plan to investigate our proposed algorithm for the large scale of the network environment.

REFERENCES

[1] M. S. M. Gismalla, A. I. Azmi, M. R. B. Salim, M. F. L. Abdullah, F. Iqbal, W. A. Mabrouk, M. B. Othman, A. Y. I. Ashyap, and A. S. M. Supa'at, "Survey on device to device (D2D) communication for 5GB/6G networks: Concept, applications, challenges, and future directions," *IEEE Access*, vol. 10, pp. 30792–30821, 2022.

[2] Z. Su, W. Feng, J. Tang, Z. Chen, Y. Fu, N. Zhao, and K.-K. Wong, "Energy-efficiency optimization for D2D communications underlaying UAV-assisted industrial IoT networks with SWIPT," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 1990–2002, Feb. 2023.

[3] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2133–2168, 3rd Quart., 2018.

[4] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett.*, vol. 9, no. 3, pp. 210–212, Mar. 2005.

[5] Y. Xu, H. Sun, and Y. Ye, "Distributed resource allocation for SWIPT-based cognitive ad-hoc networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 4, pp. 1320–1332, Dec. 2021.

[6] I. Budhiraja, N. Kumar, S. Tyagi, S. Tanwar, and M. Guizani, "SWIPT-enabled D2D communication underlaying NOMA-based cellular networks in imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 692–699, Jan. 2021.

[7] J. Huang, C.-C. Xing, and M. Guizani, "Power allocation for D2D communications with SWIPT," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2308–2320, Apr. 2020.

[8] J. Huang, J. Cui, C.-C. Xing, and H. Gharavi, "Energy-efficient SWIPT-empowered D2D mode selection," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3903–3915, Apr. 2020.

[9] 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.

[10] B. Zhao and X. Zhao, "Deep reinforcement learning resource allocation in wireless sensor networks with energy harvesting and relay," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2330–2345, Feb. 2022.

[11] H. Zheng, K. Xiong, M. Sun, H. Wu, Z. Zhong, and X. Shen, "Maximizing age-energy efficiency in wireless powered industrial IoE networks: A dual-layer DQN-based approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 2, pp. 1276–1292, Feb. 2024.

[12] X. Liao, X. Hu, Z. Liu, S. Ma, L. Xu, X. Li, W. Wang, and F. M. Ghannouchi, "Distributed intelligence: A verification for multi-agent DRL-based multibeam satellite resource allocation," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2785–2789, Dec. 2020.

[13] S. Zhang, Z. Ni, L. Kuang, C. Jiang, and X. Zhao, "Load-aware distributed resource allocation for MF-TDMA ad hoc networks: A multi-agent DRL approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 6, pp. 4426–4443, Nov. 2022.

[14] P. Sunehag, R. Evans, G. Dulac-Arnold, Y. Zwols, D. Visentin, and B. Coppin, "Deep reinforcement learning with attention for slate Markov decision processes with high-dimensional states and actions," 2015, *arXiv:1512.01124*.

[15] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao, "Deep reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5064–5078, Apr. 2024.

[16] C. Cartis, N. I. M. Gould, and P. L. Toint, "On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems," *SIAM J. Optim.*, vol. 20, no. 6, pp. 2833–2852, Jan. 2010.

[17] K. Lee, J.-R. Lee, and H.-H. Choi, "Learning-based joint optimization of transmit power and harvesting time in wireless-powered networks with co-channel interference," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3500–3504, Mar. 2020.

[18] V.-D. Nguyen, T. Q. Duong, H. D. Tuan, O.-S. Shin, and H. V. Poor, "Spectral and energy efficiencies in full-duplex wireless information and power transfer," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2220–2233, May 2017.

SENGLY MUY received the B.S. degree from the Institute of Technology of Cambodia (ITC), Phnom Penh, Cambodia, in 2018. He is currently pursuing the integrated M.S. and Ph.D. degree with the School of Intelligent Energy and Industry, Chung-Ang University, Republic of Korea. His current research interests include performance optimization in wireless networks, artificial intelligence, and machine learning.

EUN-JEONG HAN received the B.S. degree from the School of Electrical and Electronics Engineering, College of ICT Engineering, Chung-Ang University, Seoul, South Korea, in 2021, and the M.S. degree from the School of Intelligent Energy and Industry, Chung-Ang University, in 2023. Her research interests include optimization problem using machine learning in wireless networks, federated learning, and artificial intelligence.

JUNG-RYUN LEE (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from Seoul National University, in 1995 and 1997, respectively, and the Ph.D. degree in electrical and electronics engineering from Korea Advanced Institute of Science and Technology (KAIST), in 2006. From 1997 to 2005, he was a Chief Research Engineer with LG Electronics, South Korea. From 2006 to 2007, he was a full-time Lecturer of electronic engineering with the University of Incheon. Since 2008, he has been a Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, South Korea. His research interests include energy-efficient networks and algorithms, bioinspired autonomous networks, and artificial intelligence-based networking. He is a Regular Member of IEICE, KIISE, and KICS. He received the Excellent Paper Award at ICUFN 2012, the Best Paper Award at ICN 2014, the Best Paper Award at QSHINE 2016, and the Excellent Paper Award at ICTC 2018.