IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 June 2024, accepted 16 July 2024, date of publication 29 July 2024, date of current version 19 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434619

RESEARCH ARTICLE

FeDRL-D2D: Federated Deep Reinforcement Learning-Empowered Resource Allocation Scheme for Energy Efficiency Maximization in D2D-Assisted 6G Networks

HAFIZ MUHAMMAD FAHAD NOMAN[1], (Graduate Student Member, IEEE), KAHARUDIN DIMYATI[1], (Member, IEEE), KAMARUL ARIFFIN NOORDIN[1], (Senior Member, IEEE), EFFARIZA HANAFI[1], (Senior Member, IEEE), AND ATEF ABDRABOU[2], (Member, IEEE)

[1]Advanced Communication Research and Innovation (ACRI), Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[2]Electrical and Communication Engineering Department, College of Engineering, UAE University, Al Ain, United Arab Emirates

Corresponding authors: Effariza Hanafi (effarizahanafi@um.edu.my) and Kamarul Ariffin Noordin (kamarul@um.edu.my)

This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme under Grant FRGS/1/2020/TK0/UM/02/30.

ABSTRACT Device-to-device (D2D)-assisted 6G networks are expected to support the proliferation of ubiquitous mobile applications by enhancing system capacity and overall energy efficiency towards a connected-sustainable world. However, the stringent quality of service (QoS) requirements for ultra-massive connectivity, limited network resources, and interference management are the significant challenges to deploying multiple device-to-device pairs (DDPs) without disrupting cellular users. Hence, intelligent resource management and power control are indispensable for alleviating interference among DDPs to optimize overall system performance and global energy efficiency. Considering this, we present a Federated DRL-based method for energy-efficient resource management in a D2D-assisted heterogeneous network (HetNet). We formulate a joint optimization problem of power control and channel allocation to maximize the system's energy efficiency under QoS constraints for cellular user equipment (CUEs) and DDPs. The proposed scheme employs federated learning for a decentralized training paradigm to address user privacy, and a double-deep Q-network (DDQN) is used for intelligent resource management. The proposed DDQN method uses two separate Q-networks for action selection and target estimation to rationalize the transmit power and dynamic channel selection in which DDPs as agents could reuse the uplink channels of CUEs. Simulation results depict that the proposed method improves the overall system energy efficiency by 41.52% and achieves a better sum rate of 11.65%, 24.78%, and 47.29% than multi-agent actor-critic (MAAC), distributed deep-deterministic policy gradient (D3PG), and deep Q network (DQN) scheduling, respectively. Moreover, the proposed scheme achieves a 5.88%, 15.79%, and 27.27% reduction in cellular outage probability compared to MAAC, D3PG, and DQN scheduling, respectively, which makes it a robust solution for energy-efficient resource allocation in D2D-assisted 6G networks.

INDEX TERMS 6G, device-to-device communications, double deep Q-network (DDQN), energy efficiency, federated-deep reinforcement learning (F-DRL), resource allocation.

I. INTRODUCTION

The upcoming sixth generation of wireless networks (6G) is envisioned as an indispensable element in various fields of

The associate editor coordinating the review of this manuscript and approving it for publication was Francisco Rafael Marques Lima.

the cyber-physical world. It will address the end-to-end connectivity concerning humans and AI-powered autonomous machines to support ever-present intelligent communication and contribute towards a human-friendly, sustainable, and efficient society [1]. It envisages delivering a truly omnipresent wireless intelligence to unleash various emerging services such as cyber-physical continuum, green communication, zero-touch cognitive networks, extended reality (XR), internet of senses (IoS), holographic projections, etc. [2], [3]. However, ensuring sustainable end-to-end connectivity for massive machine-type communications imposes stringent QoS requirements in 6G wireless network design guidelines that emphasize the significance of energy-efficient resource management to contribute towards self-sustainable networks (SSNs). Furthermore, it requires privacy-aware and intelligent context-aware communication links to accommodate substantial and heterogeneous traffic demands and minimize energy consumption through machine learning-assisted communication protocols. In the sight of green communication, a core economic incentive lies in reducing energy consumption and global carbon emissions [4].

D2D communication paves the way for minimizing the system energy consumption and improving spectrum efficiency by connecting the proximity devices through a direct link without routing the base stations [5], [6]. The dense deployment of multi-tier heterogeneous devices with diverse QoS requirements in 6G wireless networks motivates the implementation of underlay D2D-enabled communication to overcome the challenges of ubiquitous connectivity, energy efficiency, latency minimization, and spectral efficiency. Moreover, the trade-off between spectral and energy efficiency is critical in designing D2D-assisted 6G wireless networks [7]. However, this trade-off can be addressed by considering the energy consumption and minimum spectral efficiency constraints in the energy-efficient resource allocation problem [8]. Additionally, in underlay communication, D2D users share licensed spectrum with cellular users, leveraging spatial diversity to reuse the spectrum, which enhances overall network capacity. Nevertheless, employing multiple DDPs for spectrum reuse without interrupting cellular transmission is challenging because it potentially leads to interference [9]. Therefore, intelligent resource management is essential to mitigate interference among UEs and enhance system performance considering global energy efficiency and network sum rate [10].

Resource management problems in wireless networks are conventionally solved using heuristic or suboptimal strategies. However, the emergence of 6G networks presents joint optimization problems in mixed-integer non-linear programming. It involves joint optimizations of resource and power allocation under computation efficiency constraints. Moreover, global optimization techniques, such as branch-and-bound algorithms, heuristic algorithms, etc., face challenges in solving NP-hard problems due to their exponential complexity and stringent requirements, including robustness, resource efficiency, and reliability. To address this, machine learning (ML) algorithms are considered prospective contenders to bridge the gap between computational complexity and optimal performance [11]. Furthermore, intelligent resource management enables energy-efficient end-to-end (E2E) seamless connectivity in multi-tier heterogeneous 6G networks [12], [13]. Therefore, to jointly optimize network resources in 6G wireless networks, ML-driven intelligent resource management requires a paradigm shift in conventional resource management techniques. In addition, many researchers explored ML-enabled resource management across different network layers in 6G wireless networks [14], [15]. Nevertheless, the optimization challenges associated with mixed integer nonlinear programming and dynamic environment conditions in ultra-dense and heterogeneous networks still need further investigation. It highlights the significance of ML methods for resource allocation problems in D2D communication. Furthermore, selecting an optimal ML technique for joint resource allocation, power control, and computation complexity in dynamic heterogeneous networks is also challenging [16], [17], [18]. Hence, researchers employed model-free reinforcement learning to overcome this challenge [19].

Reinforcement learning (RL) empowers autonomous agents to make sequential decisions by interacting with the environment through trial and error. This iterative process involves optimizing past actions, initiating new ones, and learning from the outcomes dynamically. However, RL suffers from inadequate scalability and high computational complexity to handle large-scale dynamic networks. Deep reinforcement learning (DRL) is used to surmount these constraints by combining the RL strategy with deep neural networks (DNNs). It employs the trained model to calculate the optimal decision while reducing the computational complexity. Hence, DRL is an outstanding tool for addressing NP-hard optimization problems in wireless networks. Its ability to make decisions based on optimal policy to explore feasible solutions with dynamic change renders it a highly adept candidate for addressing resource management problems in 6G wireless networks [20].

Federated learning (FL) enables multiple nodes to contribute towards a single global model training [21]. FL ensures the privacy of local agents through restricted sharing of their data with external entities [22]. Furthermore, federated reinforcement learning allows individual users to explore the environment independently while simultaneously training a global model to leverage the experiences of others. Compared with DRL, F-DRL focuses more on solving collaborative decision-making problems involving multiple agents through distributed machine learning. As a decentralized learning paradigm, it ensures users' data privacy through cooperative learning in which various nodes contribute towards a single global model training. Hence, the devices are trained on local datasets before offloading their models for global aggregation. F-DRL offers a

scalable and adaptive framework to address users' privacy and resource allocation challenges in D2D networks by leveraging local decision-making and collaborative learning among distributed agents. However, implementing F-DRL for joint optimization of resource allocation and power control for energy efficiency maximization in multi-cell D2D heterogeneous networks still needs further investigation. Hence, this work proposes an F-DRL approach to improve energy efficiency for a D2D heterogeneous network. We investigate a decentralized DDQN-based method for joint optimization of power control and channel assignment under QoS constraints for cellular and D2D users.

II. RELATED WORK AND CONTRIBUTION

A. RELATED WORK

Recently, DRL has been used extensively for resource management in D2D-enabled wireless communication. This section presents recent works on DRL-based methods for resource allocation to improve system performance, considering energy efficiency, sum rate, spectrum efficiency, and users' privacy. For example, authors in [23] presented deep deterministic policy gradient (DDPG) to improve the sum rate and fairness in D2D-assisted NOMA networks. In [24], a resource-matching framework for D2D-enabled uplink cellular networks is proposed. The authors used a DDQN for transmit power management while considering the sum rate of D2D users. In [25], an intelligent resource allocation method using DQN is presented. It maximizes the overall throughput in D2D-assisted underlay cellular networks under minimum SINR threshold requirements for CUEs and DDPs. In [26], the authors considered DQN for power allocation to improve the weighted sum rate. In [27], a distributed resource allocation approach using the Stackelberg game is discussed. It guides the learning agents by the Stackelberg Q-value to obtain an optimal policy through the Stackelberg equilibrium. Further, it maximizes the data rate of all CUEs and improves spectrum efficiency by sharing uplink channels with DDPs.

In [28], a priority sampling-enabled dueling-DDQN (PS-D3QN) is presented for co-channel interference management in D2D communications. In this approach, each DDP acts as an agent to share the resources for throughput enhancement. In another work [29], a multi-agent actor-critic (MAAC) approach is proposed for spectrum allocation in a D2D underlay network. In [30], [31], [32], the DDQN is used for resource allocation in D2D communication. The double estimator method in DDQN prevents the overestimation of action values, which results in a fairer value to improve network throughput and spectrum efficiency in these works. In [33], a multi-agent resource allocation problem is investigated using advantage actor-critic (A2C). It ensures the QoS requirements of CUEs and DDPs are satisfied while maximizing the throughput of D2D links. In [34], a framework to improve the throughput in NOMA-enabled D2D networks is investigated. A multi-agent DDPG is proposed for channel assignment to improve the network sum rate considering the SINR constraints of CUEs and DDPs.

Recent studies have employed DRL to maximize energy efficiency in D2D-assisted cellular communications. For instance, in [35], a deep Q-network is presented for system energy efficiency maximization under throughput constraints. It exploits two parallel DQNs for transmit power optimization in a D2D communication environment. In another work [36], the EE optimization for SWIPT-enabled D2D communication is presented by leveraging a multi-agent deep Q-learning model. In [37], subchannel assignment and power splitting are discussed for IoT-assisted energy harvesting (EH) D2D communication. It aims to maximize energy efficiency, subject to minimum data rate constraints. In [38], a multi-agent DQN is presented for system energy efficiency and throughput maximization in UAV-assisted D2D communication. This work adopted a non-linear EH approach to achieve an optimal power-splitting ratio. In [39], the authors investigated subcarrier allocation and power optimization in a D2D underlay network. It exploits DDQN for DDPs transmit power optimization while reducing co-channel interference. However, these works [35], [36], [37], [38], [39] do not consider enhancing energy efficiency under the multi-cell heterogeneous network environment.

Energy efficiency optimization is a critical component of wireless resource management, which can address the trade-off between achievable energy harvesting and network sum rate. For instance, in [40], energy efficiency and fair scheduling optimization are considered. A DRL model is trained to optimize the harvested energy and proportionate fairness among DDPs. In [41], a D2D-assisted cluster association and power optimization for NOMA-based HetNet are discussed. It aims to maximize energy efficiency by exploiting a twin delayed-DDPG (TD3) method. In [42], a power allocation scheme using a decentralized DDPG algorithm is presented. It aims to reduce the energy consumption in a D2D-NOMA-enabled vehicular network. These works [40], [41], [42] provide adept proposals for energy efficiency optimization in D2D heterogeneous networks. However, the proposed models do not consider the co-channel and cross-channel interference, which may affect the overall system performance. In [43], the authors proposed DDPG to maximize the average EE of the D2D links in underlay cellular communication. However, user privacy is not taken into consideration. Furthermore, the issue of non-stationarity can be addressed by considering the essential information exchange among UEs, which is achieved through federated edge learning [44], [45], [46], [47]. Federated learning is cooperative learning, which reduces the communication overhead at the server and preserves users' privacy compared to centralized data aggregation and training [47].

In recent works, F-DRL has been implemented for resource allocation and power optimization in D2D-enabled cellular networks [6], [48], [49]. F-DRL enhances convergence performance by enabling UEs to exchange their experiences. In [48], the authors developed a D2D-aided digital twin architecture for industrial IoT edge networks. It utilizes the Federated-DQN (F-DQN) to maximize the sum throughput

TABLE 1. Comparison of related work with the proposed approach.

Ref.	Network Model			ML type	Approach	Optimization Objective	DCL	CA	EE	User QoS	User Privacy
	D2D	Underlay	HetNet								
[23]	✓	✓		DRL	D3PG	Maximize sum rate		✓		✓	
[24]	✓	✓		DRL	DDQN	Maximize system TP		✓		✓	
[25]	✓	✓		DRL	DQN	Maximize sum rate				✓	
[26]	✓			DRL	DQN	Maximize the DDP sum rate		✓		✓	
[27]	✓	✓		DRL	MADRL	Maximize system TP		✓		✓	
[28]	✓			DRL	PS-D3QN	Maximize DDP TP				✓	
[29]	✓	✓		DRL	MAAC	Maximize sum rate		✓		✓	
[30]	✓	✓		DRL	DDQN	Maximize SE				✓	
[31]	✓	✓		DRL	DDQN	Maximize sum TP				✓	
[32]	✓	✓		DRL	DDQN	Maximize sum TP		✓		✓	
[33]	✓	✓		DRL	MAA2C	Maximize system TP		✓		✓	
[34]	✓	✓		DRL	MADDPG	Maximize sum TP		✓		✓	
[35]	✓	✓		DRL	DQN	Maximize TP and EE			✓	✓	
[36]	✓	✓		DRL	MADQL	Maximize EE			✓	✓	
[37]	✓	✓		DRL	DDQN	Maximize EE		✓	✓		
[38]	✓	✓		DRL	DDQN	Maximize EE and TP			✓	✓	
[39]	✓	✓		DRL	DDQN	Maximize sum TP		✓	✓	✓	
[40]	✓			DRL	DNN	Maximize EE		✓	✓		
[41]	✓		✓	DRL	T-D3PG	Maximize EE			✓		
[42]	✓			DRL	DDPG	Minimize power consumption			✓	✓	
[43]	✓	✓	✓	DRL	DDPG	Maximize EE		✓	✓	✓	
[48]	✓			F-DRL	DQN	Maximize system TP	✓	✓		✓	✓
[49]	✓	✓	✓	F-DRL	DQN	Maximize sum rate	✓	✓		✓	✓
This Work	✓	✓	✓	F-DRL	DDQN	Maximize EE	✓	✓	✓	✓	✓

*CA: Channel Allocation; **DCL:** Decentralized Learning; **EE:** Energy Efficiency; **HetNet:** Heterogeneous Network; **TP:** Throughput*

of DDPs. In [49], F-DQN is presented for network throughput improvement in D2D-enabled HetNet. Nevertheless, the overestimation of action values in DQN and joint optimization of power control and resource allocation to improve energy efficiency in a multi-cell D2D heterogeneous scenario are not considered.

The aforementioned studies consider resource allocation and optimization of energy efficiency in D2D-assisted networks. However, there are still some challenges. Firstly, joint optimization problems of resource allocation, energy, and computation efficiency are NP-hard due to non-convexity. Hence, traditional optimization techniques, such as linear programming, branch-and-bound algorithms, heuristic algorithms, etc., are not scalable to address these problems due to their exponential complexity. On the other hand, ML techniques are a promising tool for addressing the trade-off between computational complexity and optimal performance [11]. However, the requirement of extensive training data in supervised learning (SL) and unsupervised learning (uSL) may limit their applicability to resource management problems in D2D-assisted communication.

In contrast to SL and uSL, DRL eliminates the need for extensive pre-collected training data, making it suitable for dynamic network environments. Despite the advantages of DRL, the existing works [24], [25], [26], [30], [31], [32], [35], [37], [38], [39], [40], [42] exploit centralized DRL schemes, relying on a single global model, which may not adequately capture the inherent local variations and specific conditions experienced by individual users within a heterogeneous network environment. These centralized DRL-based methods have not considered collaborative learning, which could significantly impact system-level performance. For instance, deploying large-scale D2D networks exacerbates centralized DRL approaches' scalability for network resource management, communication overhead, and potentially compromising data privacy [49]. To address these limitations,

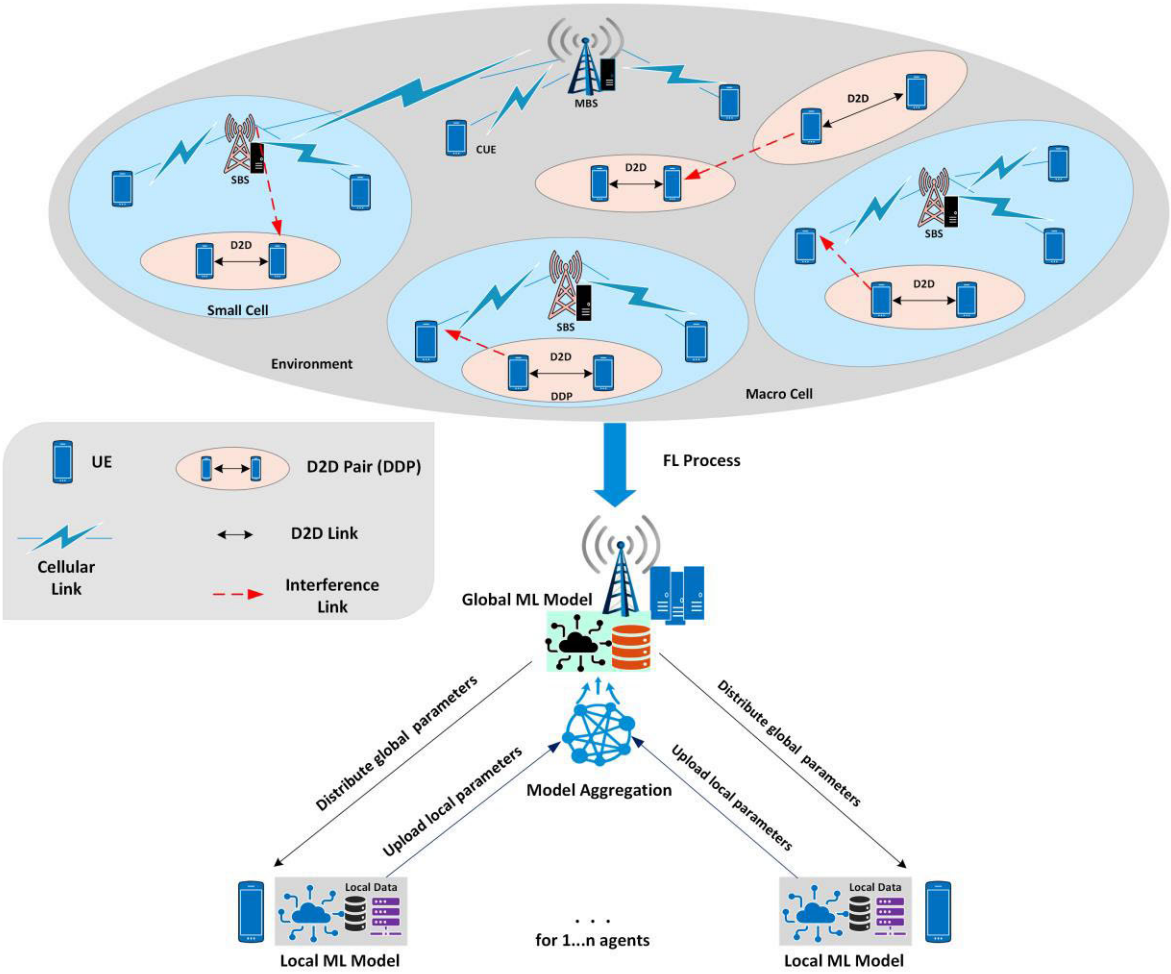

FIGURE 1. Network architecture for F-DRL-empowered D2D-assisted heterogeneous environment.

a decentralized resource allocation framework using F-DRL is a promising solution for a D2D-assisted heterogeneous network.

In F-DRL, UEs explore the environment individually and share their experiences by training a global model. Compared with multi-agent reinforcement learning methods [27], [29], [33], [34], [36], the F-DRL approach allows the UEs for collaborative learning, thus improving the convergence performance [48]. Finally, designing an energy-efficient decentralized framework that jointly optimizes channel allocation and power control to train a global model considering users' privacy is still an open issue. This motivates us to leverage the F-DRL potential for resource allocation and power optimization in a multi-cell D2D-assisted HetNet environment. Table 1 provides a summary of the comparison of the proposed approach with recent works on DRL-enabled resource management in D2D-assisted heterogeneous networks. These related works in Table 1 address the DRL-assisted resource management and user transmit power control to enhance the energy efficiency, spectrum efficiency, and network sum rate. However, the users' privacy-aware energy-efficient resource allocation and power control under stringent QoS requirements in multi-cell underlay D2D heterogeneous networks still need further investigation.

B. CONTRIBUTION AND PAPER ORGANIZATION

To address the challenges mentioned above, we investigate F-DRL for energy efficiency maximization in a device-to-device heterogeneous network. We propose a decentralized DDQN-based method to jointly optimize the power control and channel allocation to ensure users' privacy under QoS constraints. To the best of our knowledge, the proposed method is the first endeavor to exploit the F-DRL as a decentralized learning paradigm for energy efficiency maximization through power control and channel allocation in a multi-cell D2D underlay HetNet environment. This work considers multiple DDPs reusing the same channels with CUEs to maximize energy efficiency under the QoS constraints. We evaluate the uplink transmission of the underlying D2D system because uplink resources are generally

not fully utilized compared to downlink resources in cellular communications. The significant contributions of the paper are given:

- A framework for joint optimization of power control and channel allocation in a D2D-assisted HetNet is proposed. The formulated problem maximizes energy efficiency while considering the users' privacy and QoS requirements for CUEs and DDPs.
- The energy efficiency maximization problem is modeled as reinforcement learning using the Markov decision process (MDP). Specifically, we delineate the state, action, and reward for a D2D-assisted HetNet environment where each UE performs as an agent. It leverages DRL to select the appropriate action to achieve a predefined objective through interaction with the environment.
- A federated deep reinforcement learning-based method is investigated to address users' privacy and joint optimization of transmit power and channel assignment. The proposed method employs FL for decentralized training to ensure user privacy and a double-deep Q-network is used for intelligent resource allocation. This prevents overestimation of the target output by separating the action selection and target output estimation.
- The proposed framework is evaluated in terms of system energy efficiency and network sum rate, considering users' privacy and QoS constraints. The simulation results depict that the proposed F-DDQN framework performs better than the existing DRL schemes, highlighting the proposed solution's scalability and robustness.

The rest of the article is outlined as follows: Section III proposes the system model comprising the network, communication, and energy efficiency model. In Section IV, the EE optimization problem transformed as MDP is discussed. The Federated-DRL-based resource allocation framework is given in Section V. Section VI presents simulation conditions and performance analysis of the proposed approach. Finally, the conclusion is given in Section VII.

III. SYSTEM MODEL

We consider a D2D-enabled heterogeneous network comprising a macro-cell base station (MBS) and small base stations (SBSs) with C cellular user equipment (CUEs) and D D2D pairs (DDPs). Fig. 1 illustrates the network architecture in which each DDP contains a single transmitter and receiver. The sets of CUEs, DDP transmitters, and DDP receivers are given as $M = \{1, 2, \ldots, m\}$, $N = \{1, 2, \ldots, n\}$, $N' = \{1, 2, \ldots, n'\}$ respectively. The list of mathematical notations is given in Table 2.

In the proposed model, UEs can communicate in cellular or D2D mode. In cellular mode, UE establishes communication with the associated base station. Meanwhile, UEs that can set up a D2D link within a specific distance range r are called DDPs. The proposed system model leverages

TABLE 2. List of mathematical notations.

Notations	Description
C	Number of cellular user equipment (CUEs)
D	Number of D2D pairs (DDPs)
$P^c_{mk}(t)$	CUE transmit power
$P^d_n(t)$	Transmit power of DDP transmitter
$\gamma^c_{mbk}(t)$	SINR for CUEs link with base stations
γ_{th}	Minimum SINR value for CUEs and DDPs
$\gamma^d_{nn'}(t)$	SINR for D2D communication
$g^c_{mbk}(t)$	Channel gain for cellular user uplink to the BS
$g^d_{nn'}(t)$	Channel gain for D2D link
$\mu_{mb}(t)$	User association of cellular link to the base station
$\mu_{nn'}(t)$	User association of D2D link
K	Number of channels
σ^c	AWGN power for cellular links
σ^d	AWGN power for DDPs
$Q^\pi(s_t, a_t)$	Q-value of state and action at time t under policy π
ω^m	Main DQN network parameters
ω^r	Target DQN network parameters

FL for the decentralized learning of multiple agents. It distributes the local ML model to shift the computational load to local UEs, reducing the transmission overhead. Furthermore, it enables UEs to act as agents, exploring the environment independently while training a global model through model aggregation to protect users' privacy. After the global model aggregation, the UEs use this trained model for resource allocation.

We propose a multi-cell communication model with uplink transmission for the underlay D2D communication, reusing the same channels with CUE and the base station. The accessible spectrum is partitioned into K channels with bandwidth denoted as β. Each channel can only be assigned to a maximum of one CUE to prevent co-channel interference. Suppose $\phi^m_k(t)$ represents the channel association for the mth CUE on the kth channel, where $\phi^m_k(t) \epsilon \{0, 1\}$. If $\phi^m_k(t) = 1$, it depicts that the kth channel is allotted to the mth CUE. On the other hand, DDPs can transmit using either reuse or cellular mode, leveraging different channels for each transmission mode. Consequently, a DDP autonomously and randomly selects a channel with equal probability from all the accessible channels in the reuse mode. Nevertheless, when each DDP operates in cellular mode, it is restricted to select channels not associated with CUEs. Let $\phi^n_k(t)$ represents the channel assignment indicator for DDPs. If $\phi^n_k(t) = 1$, it shows that the kth channel is allocated to the nth DDP. In this work, the flat-fading channel is used, and the noise power is regarded as a constant at the receiver.

Based on Shannon's theory, the UEs' achievable transmission capacity is affected by the channel capacity and signal-to-interference plus noise ratio (SINR) [50]. Hence, the SINR is influenced by both the signal power and interference. In addition, mutual interference between DDPs and CUEs cannot be avoided due to channel reuse. Consequently,

the proposed model utilizes the maximal SINR. Let $\mu_{mb}(t)$ denote the association relationships of CUEs if the mth user associates with bth BS at time t, $\mu_{mb}(t) = 1$. Similarly, $\mu_{nn'}(t)$ denotes the association relationship between DDPs. When the nth DDP transmitter connects with the n'th DDP receiver to form a DDP link at time t, $\mu_{nn'}(t) = 1$.

With the above communication model, the received SINR at time t from the transmitter of the mth CUE to the bth base station for the kth channel is represented as:

$$\gamma^c_{mbk}(t) = \frac{\mu_{mb}(t) P^c_{mk}(t) G^c_{mbk}(t)}{\sum_{u\in M|\{m\}} \left(\mu_{ub}(t) P^c_{uk}(t) G^c_{ubk}(t)\right) + \sigma^c} \quad (1)$$

where $P^c_{mk}(t)$ and $P^c_{uk}(t)$ denote the mth and uth CUE transmit power for the kth channel, respectively. $G^c_{mbk}(t)$ and $G^c_{ubk}(t)$ depict the channel gain between the mth CUE transmitter and the uth CUE receiver of the bth base station for the kth channel respectively and σ^c represents the noise of the receiver for the cellular link.

Alternatively, the SINR between the n^{th} DDP transmitter to n'th DDP receiver in the assigned DDP channel is expressed as:

$$\gamma^d_{nn'}(t) = \frac{\mu_{nn'}(t) P^d_n(t) G^d_{nn'}(t)}{\sum_{u\in N|\{n\}} \left(\mu_{un'}(t) P^d_u(t) G^d_{un'}(t)\right) + \sigma^d} \quad (2)$$

where $P^d_n(t)$ and $P^d_u(t)$ are the nth DDP transmitter power and the uth DDP transmitter power at time t, respectively. $G^d_{nn'}(t)$ and $G^d_{un'}(t)$ depict the channel gains between the nth DDP transmitter and the uth DDP transmitter to the n'th DDP receiver at time t respectively. σ^d represents the noise of the receiver for the DDP link. The transmission rate of the mth CUE, as determined by Shannon's theorem, can be defined using (1) as follows:

$$R^c_{mbk}(t) = \frac{\beta}{K} \log_2\left(1 + \gamma^c_{mbk}(t)\right) \quad (3)$$

Similarly, the data rate of the D2D link with the nth DDP transmitter and the n'th DDP receiver is defined as:

$$R^d_{nn'}(t) = \beta \log_2\left(1 + \gamma^d_{nn'}(t)\right) \quad (4)$$

Each DDP reuses the aggregate bandwidth assigned to the multiple CUEs; subsequently, the bandwidth for each DDP is given by β. Further, the achievable total sum rate obtained from the sum throughput of both CUE and DDP links can be written as:

$$R_{sum}(t) = \sum_{m\in M}\sum_{b\in B}\sum_{k=1}^{K} R^c_{mbk}(t) + \sum_{n\in N}\sum_{n'\in N'} R^d_{nn'}(t) \quad (5)$$

The energy efficiency of the proposed framework can be modeled as the ratio of the achievable average sum rate and total power consumption, given by [43]:

$$EE(t) = \frac{R_{sum}(t)}{\sum_{m\in M}\sum_{k=1}^{K} P^c_{mk}(t)\mu_{mb}(t) + \sum_{n\in N} P^d_n(t)\mu_{nn'}(t) + P_k} \quad (6)$$

where P_k denotes the circuit power required for baseline operations, i.e., power dissipation by the base station.

Cellular outage probability is a performance indicator that determines the impact of spectrum reuse by D2D communications on cellular users. This indicator measures the likelihood that the cellular users' SINR is less than a threshold value representing the minimum level of service required [33]. The outage probability P_r of the mth cellular user to the bth base station for the kth channel based on (1) can be expressed as:

$$P^{mbk}_r = P_r\left[\gamma^c_{mbk} \le \gamma_{th}\right] \quad (7)$$

where γ_{th} denotes the minimum threshold value of SINR received by CUEs.

IV. PROBLEM FORMULATION

We aim to find an optimal solution for joint optimization of channel allocation and power control to maximize energy efficiency, considering the QoS constraints for CUEs and DDPs. This joint optimization problem is formulated as follows:

$$\mathrm{P}_1: \max_{\gamma^c,\gamma^d,P^c,P^d} EE(t) \quad (8)$$

$$s.t. \quad C_1: \gamma^c_{mbk}(t), \gamma^d_{nn'}(t) > \gamma_{th}, \forall m\in M, \forall n\in N \quad (8a)$$

$$C_2: 0 < P^c_{mk}(t) \le P^c_{max}, \forall m, \forall k \quad (8b)$$

$$C_3: 0 < P^d_n(t) \le P^d_{max}, \forall n \quad (8c)$$

$$C_4: R^c_m(t), R^d_n(t) \ge R_{min}, \forall m, \forall n \quad (8d)$$

$$C_5: \mu_{mb}(t), \mu_{nn'}(t) \in \{0,1\}, \forall m \in M, \forall n\in N \quad (8e)$$

where P^c_{max} and P^d_{max} indicate the maximum transmit power for CUE and DDP transmitters, respectively. Moreover, γ_{th} denotes the minimum SINR requirement, and R_{min} represents the minimum throughput requirement for CUEs and DDPs, respectively. In problem P1, constraint C_1 ensures that the received SINR must be above the minimum received SINR level for both CUEs and DDPs. The constraints C_2 and C_3 guarantee the transmit power of all CUEs and DDPs to be non-negative and must not be greater than the maximum transmit power P^c_{max} and P^d_{max}, respectively. Constraint C_4 restricts the minimum QoS requirements for CUEs and DDPs, i.e., it guarantees that the achievable data rate for the CUEs and DDPs must satisfy the minimum throughput requirements. Constraint C_5 depicts the association between CUEs and base stations or D2D pairs.

We formulate the optimization problem P1 as a dynamic channel assignment and power control problem to maximize energy efficiency subject to QoS constraints for CUEs and DDPs. The objective of problem P1 is to maximize energy efficiency, while the constraints on minimum SINR and minimum data rate requirements ensure that spectral efficiency is maintained. These constraints ensure the efficient use of spectral resources by meeting SINR and throughput requirements.

Moreover, optimal channel allocation and transmission power control mitigate interference, thereby enhancing the overall network capacity.

The optimization problem P1 is characterized as non-convex and belongs to mixed-integer non-linear programming. Hence, using conventional optimization methods to address this NP-hard problem with inequality constraints is challenging. Furthermore, conventional optimization techniques are not suitable to handle dynamic resource constraints or non-linear objective functions, hindering their effectiveness in this energy-efficient resource management for the D2D HetNet context. In contrast, reinforcement learning exploits its dynamic programming ability to learn and adapt to dynamic conditions, potentially leading to better performance than the traditional optimization techniques. Consequently, we formulate problem P1 through the MDP, and DRL is employed to address this problem.

V. FEDERATED-DRL-BASED RESOURCE ALLOCATION

In this work, Federated-DRL is employed for decentralized resource allocation. Specifically, the Federated-Double Deep Q Network (F-DDQN) is used for channel assignment and power allocation. In the subsequent subsections, we briefly discuss an overview of RL, DQN, and DDQN. Further, a detailed description of the proposed framework is provided, including the DDQN algorithm and FL-based model aggregation, to achieve privacy-aware and energy-efficient resource allocation for a D2D-assisted HetNet environment.

A. OVERVIEW OF RL, Q-LEARNING, AND DQN

The reinforcement learning paradigm involves discrete time agent-environment interactions and an MDP is used to model these interactions. The MDP basic components include (S, A, P, π, R), where S depicts the state space and refers to the agent's observations of the environment. It comprises all the possible finite states. A depicts the action space, which includes all the potential decisions. Each decision in the action space is called an action for the agent. P denotes the probability of transition between states, which describes changes in the environment during agent interactions. If an agent observes a state s_t and takes a decision a_t for a time instance t, the transition probability $P_{\mathrm{a}}(s_t, s_{t+1})$ is the likelihood that the state becomes s_{t+1} for the subsequent step. π denotes the agent's decision rule. The likelihood for taking action a_t on the state s_t by an agent is given as $\pi\ (s_t, a_t)$, where $a_t \in A$, $s_t \in S$, and $\sum_{a_t \in A} \pi\ (s_t, a_t) = 1, \forall s_t \in S$. Moreover, R represents rewards depending on the current state and action. When an action is performed, a reward is returned by the environment, which determines the agent's performance in achieving the optimization target. Let the agent perform an action a_t at state s_t; the reward obtained can be expressed as r_{t+1}.

Consider an agent executes an action a_t with the state s_t following a policy π at time t. Then, the environment will determine the agent performance and return a reward r_{t+1} before proceeding to the next state s_{t+1}. Through repeated iterations, the agent receives a sequence of rewards that are utilized to determine the total discounted reward, as given by:

$$R_t = \sum_{i=0}^{T-1} \lambda^i r_{t+i+1} \tag{9}$$

where R_t denotes the cumulative discounted reward, and $\lambda \in [0, 1)$ represents the discount factor to quantify an agent's weight on future rewards compared to the immediate ones. RL aims at finding an optimal policy depicted by π^* for maximizing the cumulative reward based on the objective function [51]. Hence, RL is an adaptable approach to address optimal control problems, making it particularly suitable for complex environments. Moreover, if state transition probabilities in an MDP are known, dynamic programming (DP) can be employed to analyze the MDP [52]. Nevertheless, transition probability acquisition is challenging under dynamic network conditions, such as heterogeneous D2D communication. Consequently, model-free RL techniques are exploited in such scenarios, and Q-learning is widely adopted among these techniques.

The objective function of Q-learning is represented by a Q-function comprising the action-state value function. It signifies the agent's long-term value after executing actions for the current state. The expression for the Q-function with state-action pair (s_t, a_t) under a given policy π can be represented as [28]:

$$Q^{\pi}\ (s_t, a_t) = \mathbb{E}[R_t | s_t = s, a_t = a] \tag{10}$$

where $Q^{\pi}\ (s_t, a_t)$ represents the long-term mathematical expected returns obtained through MDP. Further, using the return function in (9), the Q-function is split into two parts: the first is the immediate reward, and the second part includes the discounted Q-function of the next state. Thus, (10) can be re-written as:

$$Q^{\pi}\ (s_t, a_t) = \mathbb{E}[r_{t+1} + \lambda Q^{\pi}(S_{t+1}, a_{t+1}) | s_t = s, a_t = a] \tag{11}$$

The Q-values are kept in a Q-table in individual state-action pair format, which is used to learn an optimal policy π^* based on the optimal action-value function, denoted by $Q^*\ (s_t, a_t)$. The optimal action-value function is expressed as [30]:

$$Q^*\ (s_t, a_t) = \max_{\pi} Q^{\pi}\ (s_t, a_t) \tag{12}$$

Equation (12) can be further elaborated using the Bellman optimality equation as follows [30] and [35]:

$$Q^*\ (s_t, a_t) = \mathbb{E}\left[r_{t+1} + \lambda \max_{a_{t+1}} Q^*(S_{t+1}, a_{t+1}) | s_t = s, a_t = a \right] \tag{13}$$

Due to the nonlinear optimality, there is no closed-form solution to the Bellman equation in (13). Therefore, in Q-learning, $Q^*\ (s_t, a_t)$ is iteratively updated using the sequence of experience samples through agent-environment interaction. The agent gets experience data in discrete time

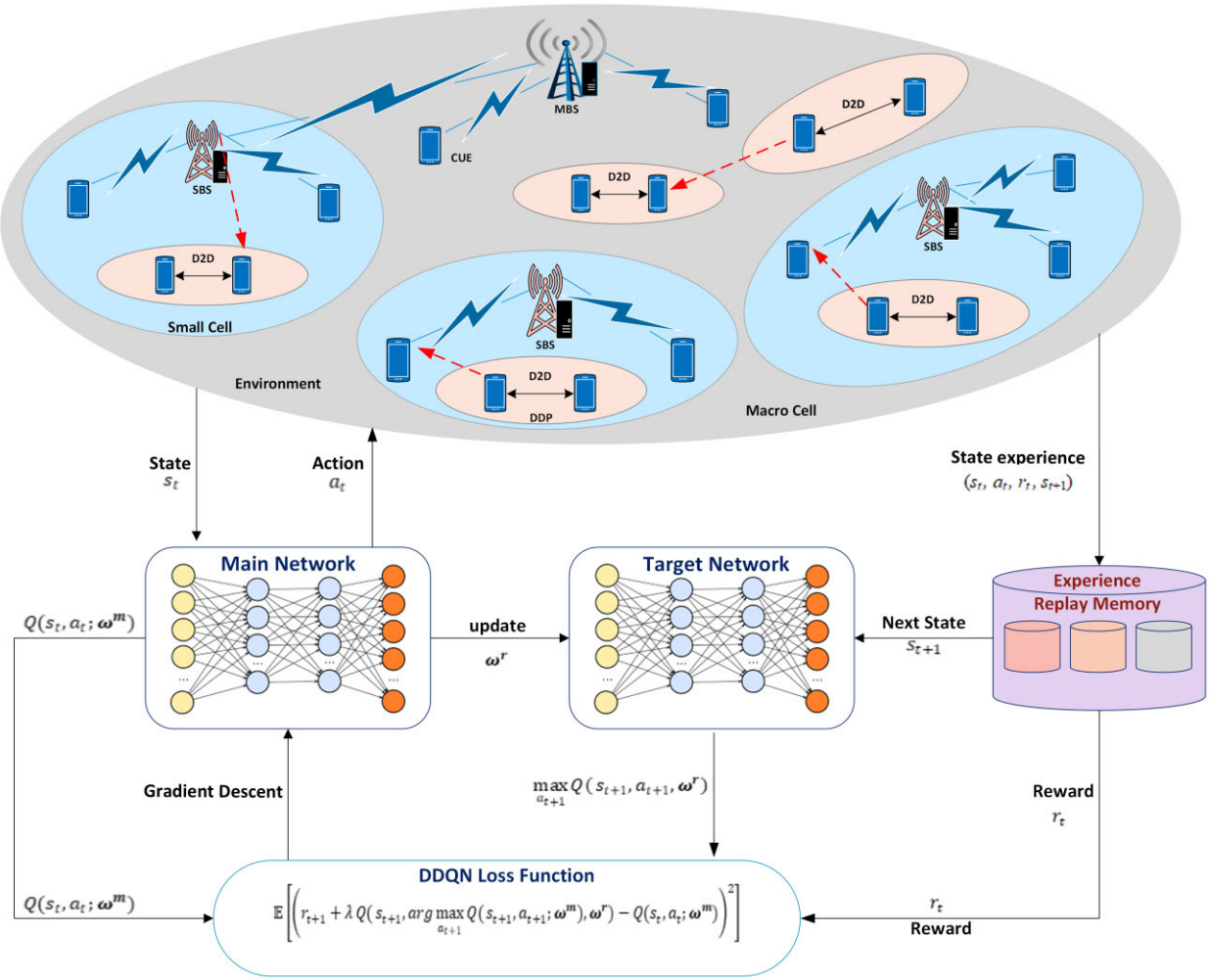

FIGURE 2. The DDQN architecture for resource allocation in D2D HetNet environment.

steps as (s_t, a_t, r_t, s_{t+1}) and updates its learned Q-function recursively as follows [35]:

$$Q(s_t, a_t) \leftarrow Q^*(s_t, a_t) + \alpha \left[r_{t+1} + \lambda \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \tag{14}$$

where $\alpha \in [0, 1)$ represents the learning rate. The agent uses the ξ-greedy algorithm for balancing the exploitation and exploration process. In particular, the agent randomly opts for an action with a probability ξ, or it takes an action to maximize the action value by a probability of $1 - \xi$. This approach has the advantage of directly approximating the optimal action-value function, $Q^*(s_t, a_t)$ through the learned Q-function obtained from (14), regardless of the agents' policy [53].

Q-learning generally shows significant performance with a small state and action space. Nevertheless, with a large state and action space, the size of the Q-value table is increased. It leads to the curse of dimensionality, wherein algorithm convergence becomes more challenging. The DRL approach improves the estimation efficiency of Q-values by using DNNs [36]. The deep Q-network is a DRL technique that uses DNNs rather than Q-tables to determine the optimal policies [54]. The environment state s_t is input to DNN to generate the predicted Q-values in the form of $Q(s_t, a_t; \boldsymbol{\omega})$, $a_t \in A$ where $\boldsymbol{\omega}$ indicates the weights of the DNN. Furthermore, the agents learn the optimal policy for Q-function optimization, which involves the loss function minimization, given as follows [36]:

$$L(\boldsymbol{\omega}) = \mathbb{E}\left[(y_t - Q(s_t, a_t; \boldsymbol{\omega}))^2 \right] \tag{15}$$

whereas y_t denotes the target Q-value represented as:

$$y_t = r_{t+1} + \lambda \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \boldsymbol{\omega}) \tag{16}$$

In the training phase, the DNN parameters are iteratively updated for Q-function optimization. This iterative update process of DNN parameters $\boldsymbol{\omega}$ is expressed as [36]:

$$\boldsymbol{\omega} = \boldsymbol{\omega} + \alpha \mathbb{E}\left[(y_t - Q(s_t, a_t; \boldsymbol{\omega})) \nabla Q(s_t, a_t; \boldsymbol{\omega}) \right] \tag{17}$$

The agent stores experience data in an experience replay pool as a tuple $(s_t,\ a_t,\ r_{t+1},\ s_{t+1})$ rather than employing a single experience data for training at each iteration. Hence, the DNN training involves the selection of a mini-batch of samples randomly from the experience replay memory [32]. The DQN algorithm employs two Q networks: online and target networks. Both online and target networks possess similar architecture but have different weights. The online network's weights are represented as $\boldsymbol{\omega^m}$, while the target network's weights are defined as $\boldsymbol{\omega^r}$. Each T training step replicates online network parameters in the target network. This iterative update process improves convergence. The target Q-value given in (16) is reformulated as follows:

$$y_t = r_{t+1} + \lambda \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \boldsymbol{\omega^r}) \tag{18}$$

Despite its advantages over Q-learning in terms of convergence, DQN still faces some limitations, including overestimation. In DQN, a single *max* mathematical estimator is employed for action selection and evaluation, resulting in a larger Q-value estimation than the actual value. When updating the Q-value function, this positive bias results in an overestimation. Nevertheless, a DDQN algorithm addresses this overestimation challenge [55]. The DDQN algorithm separates the process of selecting an action and evaluating its value by employing two distinct *max* function estimators. The double estimator technique prevents the overestimation of action values, which results in a fairer value. The target Q-value given in (18) can be reformulated for DDQN as follows [32]:

$$y_t = r_{t+1} + \lambda Q\left(s_{t+1}, arg\max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \boldsymbol{\omega^m}); \boldsymbol{\omega^r}\right) \tag{19}$$

From equation (19), it can be seen that the estimation weights $\boldsymbol{\omega^m}$ are still used to define the best action (argmax) in this operation. Accordingly, the greedy strategy value is estimated by the current value given by $\boldsymbol{\omega^m}$ while the second set of weights $\boldsymbol{\omega^r}$ calculates the fairer value of the policy.

In this work, the DDQN algorithm is used to decouple action selection from value estimation. This decoupling reduces the overestimation bias and enhances the learning stability in DDQN. Moreover, employing an experience replay with DDQN improves its learning efficiency [55]. We propose a federated DDQN (F-DDQN) method to jointly optimize energy efficiency and resource allocation while preserving the user's privacy in the D2D HetNet environment. The next section presents the proposed framework.

B. PROPOSED ALGORITHM

This section describes the proposed framework using a Federated-Double Deep Q-Network algorithm. The proposed scheme leverages the F-DDQN framework to ensure users' privacy and prevent overestimating the target output by separating action selection and target output estimation. To implement the F-DDQN approach for maximizing the system energy efficiency formulated as P1, each user equipment acts as an agent to monitor the network state s_t and take appropriate action a_t. Further, a reward is generated for every time slot, and the executed action is evaluated. In the proposed framework, the agent, the network state space, the action space, and the reward function as MDP elements are given as:

- *Agents:* Each UE, i.e., D2D and cellular users, is modeled as an agent.
- *State Space:* The agents examine the state to characterize the environment, which comprises the SINR information at BS and D2D receivers and the QoS satisfaction degree for both CUEs and DDPs. The user's QoS degree of satisfaction is denoted by $\delta_i(t)$, which is defined as:

$$\delta_i(t) = \begin{cases} 1, & R_i(t) \geq R_{min} \\ \dfrac{R_i(t)}{R_{min}}, & R_i(t) < R_{min} \end{cases} \tag{20}$$

  where $R_i(t)$ denotes the data rate of the ith user equipment, and each user equipment is characterized by a minimum rate requirement, indicated by R_{min}. The QoS degree of satisfaction for the ith user, represented as $R_i(t)/R_{min}$, lies within the range [0, 1), where a value of 1.0 indicates the basic level of QoS satisfaction. Formally, the state space S concerning the environmental parameters is expressed as:

$$S(t) = \left\{ \begin{matrix} \gamma_1^c(t), \ldots, \gamma_m^c(t); \gamma_1^d(t), \ldots, \gamma_n^d(t); \\ \delta_1^c(t), \ldots, \delta_m^c(t); \delta_1^d(t), \ldots, \delta_n^d(t) \end{matrix} \right\}; \quad \forall m \in M, \forall n \in N \tag{21}$$

  where $\gamma_m^c(t)$ and $\gamma_n^d(t)$ represent the SINR for the mth cellular user and nth D2D user at time t, respectively. Similarly, δ_m^c and $\delta_n^d(t)$ represent the QoS satisfaction degree for the mth cellular user and nth D2D user at time t, respectively.
- *Action Space:* The agent selects an action a_t considering the current state and decision policy for every iteration. This action is selected from the agent's action space A, which comprises the following permissible actions:

$$A(t) = \left\{ \begin{matrix} \phi_1^c(t), \ldots, \phi_k^c(t); \phi_1^d(t), \ldots, \phi_k^d(t); \\ P_1^c(t), \ldots, P_m^c(t); P_1^d(t), \ldots, P_n^d(t) \end{matrix} \right\}; \quad \forall k \in K, m \in M, \forall n \in N \tag{22}$$

  where $\boldsymbol{\phi}_k^c$ and $\boldsymbol{\phi}_k^d(t)$ represent the kth channel assignment for the cellular user and D2D user at time t, respectively. Similarly, P_m^c and $P_n^d(t)$ represent the transmit power for the mth cellular user and nth D2D user at time t, respectively.
- *Reward Function:* The proposed scheme utilizes the energy efficiency metric defined in (6) as the reward function. This reward function incorporates all constraints associated with power allocation and channel assignment within its formulation. During training, each agent utilizes feedback from the environment to learn the

policies that maximize this reward while satisfying the given constraints. The reward function is expressed as:

$$r(t) = \begin{cases} EE_t(s,a), & \text{if constraints are satisfied} \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where EE_t represents the energy efficiency utility, defined in (6).

The DDQN architecture for resource allocation framework in the D2D-assisted HetNet environment is given in Fig. 2. The proposed DDQN framework comprises two neural networks: a main (online) network, parameterized by ω^{m} for estimating the Q-value defined by $Q(s, a; \omega^{m})$ and a target network with parameters ω^{r} which generates the target Q-value as follows:

$$y_t = r_{t+1} + \lambda Q(s_{t+1}, arg\max_{a} Q(s_{t+1}, a_{t+1}; \boldsymbol{\omega^{m}}), \boldsymbol{\omega^{r}}) \tag{24}$$

The variables s_{t+1} and a_{t+1} represent the state and action of the subsequent step, respectively. During training, the main neural networks' weights are iteratively updated through back-propagation. The process of ω^{m} iteration is expressed as:

$$\boldsymbol{\omega^{m}_{t+1}} = \boldsymbol{\omega^{m}_{t}} + \alpha \mathbb{E}\left[(y_t - Q(s_t, a_t; \boldsymbol{\omega^{m}_{t}}))\nabla Q(s_t, a_t; \boldsymbol{\omega^{m}_{t}})\right] \tag{25}$$

The loss function is defined as:

$$\mathcal{L}(\boldsymbol{\omega^{m}}) = \mathbb{E}\left[\left(y_t - Q(s_t, a_t; \boldsymbol{\omega^{m}})\right)^2\right] \tag{26}$$

In the proposed scheme, each base station is equipped with a local server. Further, an experience replay buffer $\mathfrak{D}$ is used, which stores the state experience $(s^i_t, a^i_t, r^i_t, s^i_{t+1})$ to train the ith agent where $i = 1{:}\ N$. The state space $S = \{\gamma^c_m, \gamma^d_n, \delta^c_m, \delta^d_n\}$ represents the SINR and QoS satisfaction degree status of the CUEs and DDPs in the current environment. The action space $A = \{\phi^c_k, \phi^d_k, P^c_m, P^d_n\}$ comprises the channel assignment and power allocation decision for CUEs and DDPs depending on the environment's current state. Further, a reward r^i_t is obtained after the execution of an action a^i_t in the state s^i_t as per (23). Moreover, the resultant tuple $(s^i_t, a^i_t, r^i_t, s^i_{t+1})$ is then stored within $\mathfrak{D}$ from which the ith agent employs a random selection process to choose mini-batch samples for training the DNN parameters in the subsequent training phase where s^i_{t+1} represents the environment state for the following iteration. The proposed DDQN algorithm employs a rectified linear unit (*ReLU*) as an activation function between layers. The non-linear nature of *ReLU* leads to better exploration of the environment. Furthermore, the ξ-greedy policy is adopted to prevent the local optima while addressing the trade-off of exploration and exploitation. The resource allocation process pseudocode using the DDQN scheme with energy efficiency maximization objective is given in Algorithm 1.

The first step of Algorithm 1 entails initializing the main and target network parameters, with other parameters including the number of maximum episodes, training steps, and replay buffer size. In the next step, the DDQN model receives the system's initial state by observing the network environment. The ith agent observes a state s^i_t at a time t and

Algorithm 1 Resource Allocation Algorithm for D2D-HetNet using DDQN

Initialize:
Initialize the main DQN parameters $\boldsymbol{\omega^{m}}$
Initialize the target DQN parameters $\boldsymbol{\omega^{r}}$
Initialize training episodes, training steps, and replay buffer size as $E_{\max}$, $T_{\max}$, and N_c, respectively.
while training episode $= 1{:}\ E_{\max}$ **do**
  Update the network environment parameters $\{s^i_1\}\ i^N_1$
  for $t = 1 : T_{\max}$ **do**
    for $i = 1 : N$ **do**
      Observe the state s^i_t
      Select an action a^i_t using ξ-greedy policy
    end for
    Obtain immediate reward $r^i_t\ i = 1{:}\ N$ and the subsequent state s^i_{t+1} observation
    Store state experience $\{s^i_t, a^i_t, r^i_t, s^i_{t+1}\}$ into experience replay memory
    Randomly sample some mini-batches from the experience replay buffer
    for each sample, $e = \{s^i_t, a^i_t, r^i_t, s^i_{t+1}\}$ **do**
      Calculate the target Q-value using (24) in target network
      Update $\boldsymbol{\omega^{m}}$ using the semi-gradient of Q-learning using (25)
    end for
    Train $\boldsymbol{\omega^{m}}$ to minimize the loss function using (26)
    Update $\boldsymbol{\omega^{r}}$ using soft updates
  end for
end while

Algorithm 2 FL-Enabled Distributed Training

Initialize the maximum FL iterations to K.
for each model aggregation round, $k = 1{:}\ K$ **do**
  Initialize the global model parameters $\boldsymbol{\omega^{g}}$
  Distribute initial parameters to all agents
  Each agent performs Algorithm 1 using DDQN
  Each agent uploads local model parameters for weighted aggregation
  Global model parameters update through aggregation server using (27)
  Distribute updated parameters to all agents through the aggregation server
end for

executes an action a_t^i, resulting in a change in environment to obtain a reward r_t^i by (23), then the subsequent state s_{t+1}^i is observed by the ith agent. Moreover, the agent executes an action using the ξ-greedy policy. It selects a random action for exploration by a probability ξ, or the action with the largest Q-value is opted by a probability 1-ξ, estimated by the main network. Next, the target network generates the target Q-value using (24), i.e., the target network values the selected action.

Furthermore, the obtained state experience $\{s_t^i, a_t^i, r_t^i, s_{t+1}^i\}$ is copied within the experience replay buffer. After obtaining the immediate reward, the agent observes the next stage. For each time till E_{max} training episodes, some randomly selected mini-batch samples are extracted from the replay memory for the subsequent training. A semi-gradient descent method is adopted for updating the main network parameters $\boldsymbol{\omega^m}$ by (25). Further, the square error function given in (26) reduces the loss between the main and the target network.

The primary network parameters are updated in each iteration, whereas the target network parameters are updated iteratively using soft updates, i.e., by setting $\boldsymbol{\omega^r} = \boldsymbol{\omega^m}$. Furthermore, during this training phase, the UEs do not update the global model online because it uses the data from the previous environment. Consequently, this training procedure is termed offline DRL training. Further, the central server sends the resource allocation decision to the UE. Then, the UE can perform the power allocation and channel assignment according to the corresponding transmission mode.

After local model training on UEs, algorithm 2 executes FL-enabled global model aggregation in the subsequent phase. We use federated learning to average local models and update the global model. The aggregation server is located at MBS and acts as a centralized server. It iteratively updates the global model parameters $\boldsymbol{\omega^g}$ using local model parameters obtained from UEs. This model aggregation process can be defined as [48]:

$$\boldsymbol{\omega^g}(t) = \frac{\sum_{i=1}^{K} {}_i\boldsymbol{\omega}\eta_i^{\boldsymbol{m}}(t)}{\sum_{i=1}^{K} \eta_i} \tag{27}$$

where η denotes the size of the training batch of each agent, $\boldsymbol{\omega^m}$ and $\boldsymbol{\omega^g}$ represent the neural network weights of the local and global model, respectively. As the neural network weights are directly related to the experience and memory of each UE, this aggregation process allows each UE to share experience and memory with other UEs. Moreover, following the completion of the global model aggregation, the UEs train the local model by downloading the updated global model parameters from the server. These steps are iterated till convergence is achieved. Further, on completion of the training, UEs use the trained model to allocate resources for corresponding communication modes based on the network states. This process is given in Algorithm 2.

C. COMPUTATIONAL COMPLEXITY ANALYSIS

The computational complexity for DRL-based methods comprises the model training phase and the model inference phase [23]. In algorithm 1, the training complexity refers to the DDQN training phase, while the inference complexity arises when the trained DDQN is used to make decisions directly at runtime. After completing model training, inferences for optimal network configuration can be made within milliseconds [43]. The time computational complexity of the proposed DDQN-based scheme in algorithm 1 can be directly calculated by the neural network structure [39]. The proposed DDQN has three fully connected layers: the input, hidden, and output layers. Let n_i be the number of neurons in the input layer, and n_h and n_o represent the number of neurons in the hidden layer and output layer, respectively. The computational complexity of each layer is determined through matrix operation and activation function calculation [36]. The time complexity for computing the input layer to the hidden layer can be expressed as $O(n_i n_h)$, and the amount of calculation from the hidden layer to the output layer can be described as $O(n_h n_o)$. Therefore, the time complexity for the forward propagation is calculated as $O(n_h(n_i + n_o))$. Moreover, the time complexity of the backward pass (back-propagation and weight update) is similar to the forward propagation [28]. Since each iteration involves a forward and a backward pass and let the proposed algorithm converges after E episodes with I iterations per episode. Hence, the overall time complexity of algorithm 1 can be represented as $O(EIn_h(n_i + n_o))$. Similarly, based on the neural network weights and biases, the total space complexity can be described as $O(n_h(n_i + n_o))$.

In the proposed approach, each agent trains a local model, and these local models are then aggregated into a global model using federated learning. Since all agents are computed in parallel, the time computational complexity of a single agent is the same for all agents [49]. In algorithm 2, the model aggregation step comprises averaging the parameters (sum of weights and biases) of all agents. Hence, the time complexity for model aggregation for K agents is expressed as $O(Kn_h(n_i + n_o))$. Furthermore, the space complexity for storing the global model is equivalent to storing the parameters of a single model, denoted as $O(n_h(n_i + n_o))$.

VI. PERFORMANCE EVALUATION

In this section, we discuss simulation results to evaluate the performance of the proposed method. It comprises simulation parameters, a performance comparison of the proposed scheme with existing DRL approaches, and system performance analysis for different performance indicators, including network sum rate, energy efficiency, QoS satisfaction degree, and cellular outage probability.

TABLE 3. List of simulation parameters.

Parameter	Value	Parameter	Value
No. of Cellular UEs	10	Discount factor	0.99
No. of D2D UEs	5,10,15,20,25	Learning rate	0.01
Radius of MBS	500 m	Batch size	64
Radius of SBS	250 m	Experience buffer size	1000
Uplink Bandwidth	5 MHz	Gradient decay	0.9
Max Transmit power D2D pair	20 dBm	Initial epsilon	1
Max. transmit power of CUE	25 dBm	Min. epsilon	0.01
min. SINR CUEs and DDPs	3 dBm	Epoch's no. of steps	20
Path loss of MBS	128.1 + 37.6 log (*d*)	Squared gradient decay	0.999
Path loss of SBS	140.7 + 36.7 log (*d*)	L2 regularization	0.0001
Path loss of D2D links	148 + 40 log (*d*)	Epsilon decay	0.005
Shadow fading	8 dB	Sample Time	1
Noise Power	-174 dBm/Hz	Optimizer	Adam
Min. Data rate (QoS) CUEs and DDPs	1 Mbps	Activation function	ReLu
DDP link max. distance	20 m	Target smooth factor	0.001

A. SIMULATION SETUP

The proposed network model comprises one MBS, three SBSs, ten CUEs, and twenty-five D2D pairs. The MBS radius is 500m, and the SBS radius is given as 250m. UEs in the simulation are randomly located, and they can establish cellular links to communicate with MBS and SBS through cellular mode or communicate with other UEs through D2D mode. The maximum transmission distance between the DDP transmitter and receiver is 20m. The value of path loss is specified as 128.1+ 37.6 log10 d for the MBS link,140.7+36.7 log10 d for the SBS link, and 148+40 log10 d for the D2D link (where d in km) [4]. The CUE maximum transmit power is given as 20 dBm; in the case of DDPs, it is specified as 25 dBm. The shadow fading is given by an 8 dB lognormal distribution [12], [13]. The value of thermal noise power is specified as −174 dBm. The minimum SINR requirement for each UE is 3 dBm.

Furthermore, a DDQN is constructed with three fully connected layers: the input, hidden, and output layers. Each layer has 150, 250, and 200 neurons, respectively. The activation function is used as *ReLU*, and *Adam* is utilized as an optimizer in the proposed architecture. The proposed mechanism is evaluated using MATLAB 2023a, installed in a PC with Intel Core™ i5-11400F CPU, 16 GB random access memory, and a GPU specified as NVIDIA GeForce RTX 3060. The simulation time of the proposed approach is 0.021265 s with 6.12510^{10} number of operations corresponding to the given time complexity analysis. The discount factor value is 0.99, while the experience buffer size is 1000. The epoch's no. of steps is 20, and the learning rate is 0.01. The initial value of the exploration ξ–greedy policy algorithm is 1, which gradually drops to 0.005 with a decay rate of 0.9. Table 3 summarizes the parameters used in the simulations.

B. SIMULATION RESULTS AND DISCUSSION

In this subsection, we discuss the performance analysis of the proposed F-DDQN scheme and compare it with other existing DRL methods, including MAAC [29], Distributed DDPG (D3PG) [42], and DQN scheduling [35] in terms of network sum rate, energy efficiency, QoS satisfaction degree, and outage probability of cellular users. The MAAC-based approach [29] considers a distributed multi-agent framework using an actor-critic algorithm. It utilizes the global historical states, actions, policies, and user cooperation to improve the system sum rate. In the D3PG-based approach [42], a power allocation scheme using a decentralized DDPG algorithm is considered. The DDPG technique integrates the deterministic policy gradient with the actor-critic method. By employing decentralized learning, it observes the environment to determine the actions. Thus, an optimal policy is obtained to reduce the overall power consumption. In DQN scheduling [35], two parallel DQNs are proposed for transmit power optimization with dynamic rewards. It considers system energy efficiency maximization under throughput constraints and QoS requirements for CUEs and DDPs.

In the first step, we evaluate the network performance by analyzing the system sum rate of the proposed method and comparing it with other DRL approaches based on a different number of users. Fig. 3 shows the network sum rate for various numbers of DDPs. The figure illustrates that the network sum rate increases as the DDPs increase. This is because more DDPs would be able to explore the channel conditions to reuse the available channels and consequently achieve better performance. However, when the number of DDPs reaches 20, there is no significant increase in the network sum rate. This occurs because the co-channel interference significantly increases by increasing the number of DDPs. Furthermore, the proposed approach gives 15.83%, 27.24%,

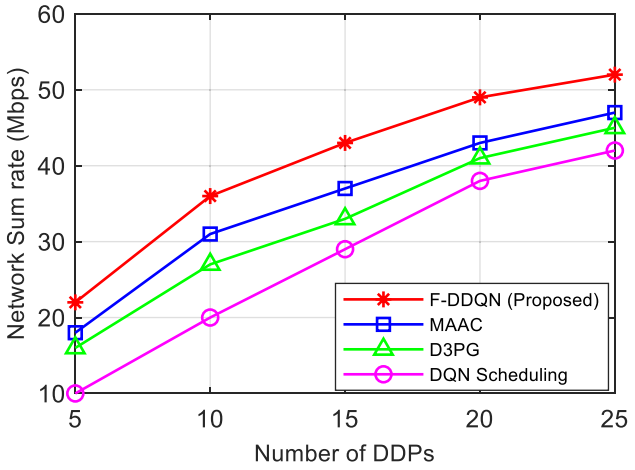

FIGURE 3. Sum rate achieved with different number of DDPs.

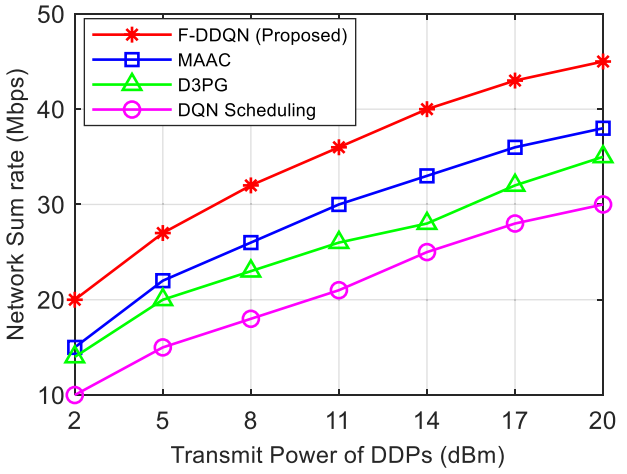

FIGURE 5. Sum rate versus different transmit power of DDPs.

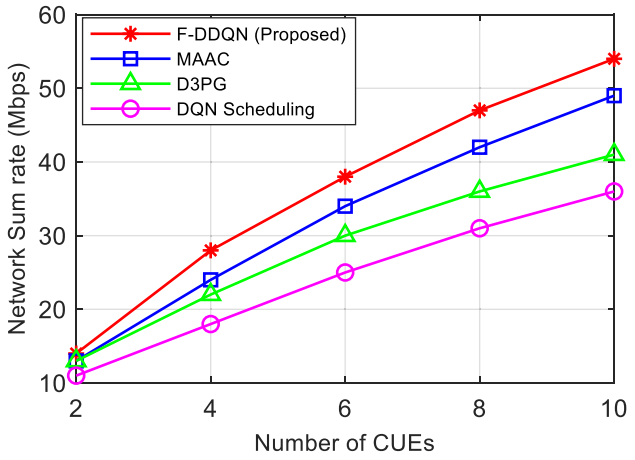

FIGURE 4. Sum rate achieved with different number of CUEs.

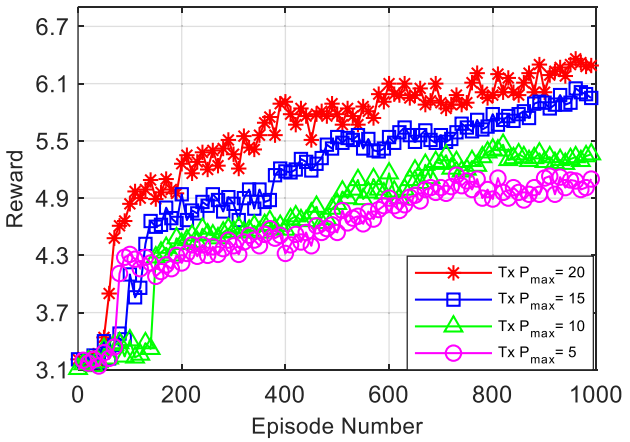

FIGURE 6. Convergence of proposed scheme with different transmit power of DDPs.

and 60.21% higher average sum rates than the MAAC, D3PG, and DQN scheduling, respectively, due to the involvement of a double Q-learning technique, which exploits two separate Q-networks for action selection and target estimation. The proposed approach rationalizes the dynamic channel selection in which DDPs as agents could reuse the uplink channels of CUEs to mitigate the interference between UEs.

Fig. 4 shows the system sum rate according to different numbers of CUEs. It indicates that the network sum rate increases by increasing the number of CUEs. This is because DDPs can reuse channels with higher channel gains to satisfy their QoS requirements with less mutual interference by increasing CUEs in a cell. Moreover, the prioritized experience replay memory allows each UE to learn its samples faster. As shown in Fig. 4, the proposed scheme significantly performs better than the existing methods with respect to network sum rate and achieves 11.65%, 24.78%, and 47.29% higher average sum rate than the MAAC, D3PG, and DQN scheduling, respectively. Fig. 5 presents the network sum rate under the distinct transmit power of DDPs. Fig. 5 shows that the network sum rate gradually increases by increasing the transmit power of DDPs. Moreover, the proposed algorithm enhances performance efficiency, achieving a significant improvement in the sum rate over the existing DRL methods.

Fig. 6 depicts the impact of different transmit power of DDPs on the convergence of energy efficiency versus training episodes. Results reveal that average rewards converge to a higher value as transmit power increases. This demonstrates that increased transmit power leads to improved system energy efficiency. Furthermore, Fig. 6 shows that at the beginning of the learning phase, the trend for average energy efficiency is approximately the same under specific values of maximum transmit power. In this case, the agents learn from their actions and experiences (through trial and error) in response to feedback signals from the interactive environment. Consequently, the initial episodes have less knowledge of the environment. Nevertheless, an adequate number of episodes allows the agents to acquire sufficient information about the network model to reach convergence toward the optimal value of the average energy efficiency.

In Fig. 7, the proposed system energy efficiency is compared with the existing methods under the different transmit

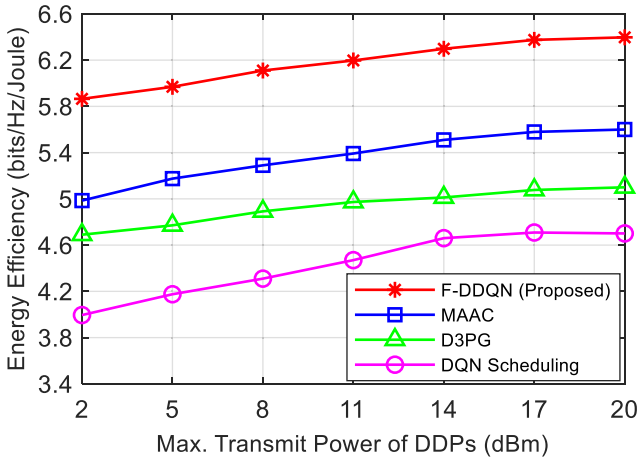

FIGURE 7. Energy efficiency versus maximum transmit power of DDPs.

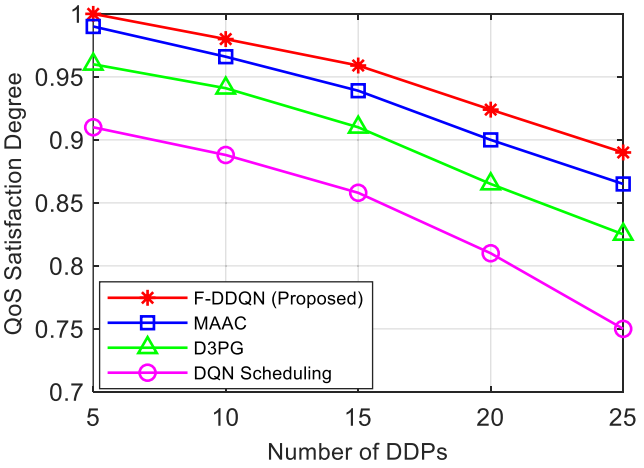

FIGURE 9. QoS satisfaction degree versus number of DDPs.

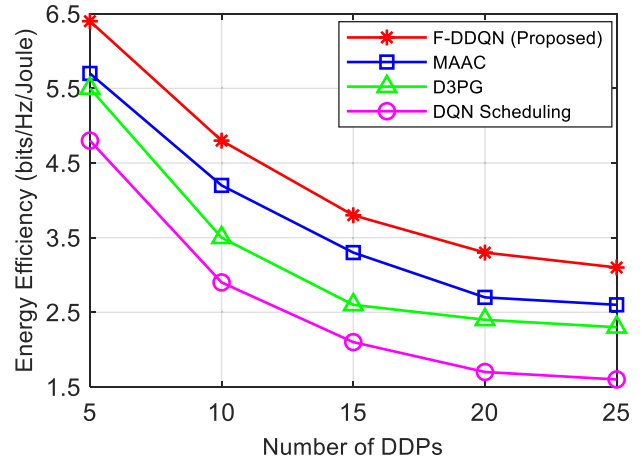

FIGURE 8. Energy efficiency versus number of DDPs.

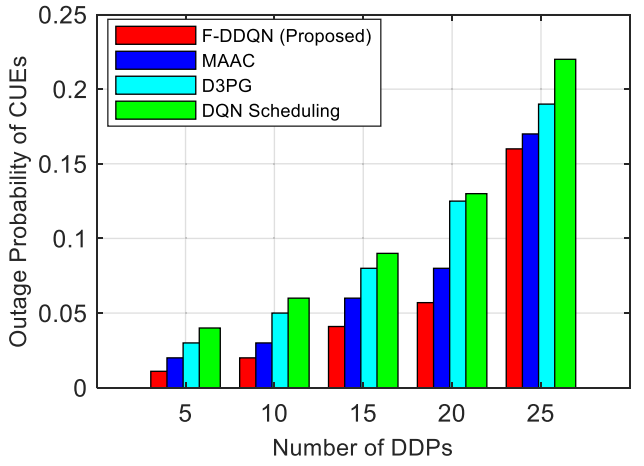

FIGURE 10. CUEs outage probability comparison of proposed method with different DRL schemes.

power of DDPs. Results depict that the energy efficiency increases with higher maximum transmit power. This occurs because higher transmit power increases the signal strength between DDPs, allowing them to achieve higher SINR and improve energy efficiency. Nevertheless, the energy efficiency converges once the maximum transmit power exceeds 20 dBm. Hence, the further increase in the transmit power (exceeding 20 dBm) causes more interference for other DDPs or CUEs, leading to a convergence or deterioration in energy efficiency. Furthermore, the results indicate that the proposed method optimizes the transmit power of DDPs to reduce the co-channel interference better than MAAC and D3PG methods because the transmit power is quantized at fixed levels in the action space. Moreover, the proposed method decouples the action selection from value estimation using two separate *max* function estimators. This decoupling mitigates the overestimation of action values, which results in a fairer value and enhances the learning stability of the proposed scheme compared with the DQN method.

Fig. 8 presents the impact of the number of DDPs on the system's energy efficiency. It can be seen from Fig. 8 that adding more DDPs leads to a decrease in energy efficiency. This happens because the increase in the number of DDPs leads to a corresponding increase in co-channel interference, resulting in a degradation of the system sum rate. Hence, the energy efficiency of a system is significantly impacted by the number of users. Due to this, when a network has a massive number of users, network operators need to configure more resources to enhance energy efficiency. Moreover, the simulation results indicate that, compared with MAAC, D3PG, and DQN scheduling, the average energy efficiency of the sys- tem is improved by 16.63%, 34.39%, and 73.53%, respectively.

Fig. 9 presents the variation in QoS satisfaction degree with respect to the number of DDPs. The results indicate that QoS is well-preserved with smaller DDP deployments. However, adding more DDPs to the environment reduces the QoS satisfaction degree. This is due to the mutual interference between the DDPs, which becomes more prevalent as the number of DDPs rises. Thus, the network can only handle a small number of DDPs due to its limited radio resources. Nevertheless, the results indicate that when the number of DDPs reaches 25, the proposed method still outperforms

the MAAC, D3PG, and DQN scheduling by 2.89%, 7.88%, and 18.67%, respectively. This implies that the proposed approach maintains a good QoS satisfaction level, indicating efficient use of spectral resources.

Fig. 10 demonstrates the comparison of the cellular outage probability of the proposed F-DDQN method with existing schemes for different numbers of DDPs. Cellular outage probability determines the communication quality of CUEs by calculating the impact of spectrum reuse by DDPs. It measures the likelihood that the received SINR by CUEs is less than the minimum SINR required for the threshold level of communication quality. Fig. 10 shows that the proposed F-DDQN scheme outperforms the existing DRL schemes with respect to cellular outage probability and provides a better communication quality to cellular users, i.e., it minimizes disruptions to cellular users by avoiding excessive interference in the presence of DDPs, which leads to efficient spectrum sharing between cellular and D2D users. Furthermore, when the number of DDPs reaches 25, the proposed scheme achieves a 5.88% decrease in outage probability as compared to MAAC, and it also shows a 15.79% and 27.27 % reduction in outage probability compared to D3PG and DQN scheduling, respectively, which highlights the scalability and robustness of the proposed solution. This scalability and robustness could be advantageous in scenarios where the system needs to handle a massive number of DDPs.

VII. CONCLUSION

This paper presents a privacy-aware and energy-efficient resource allocation scheme for D2D-assisted 6G networks. We investigate the joint optimization of channel allocation and power control to maximize the system's energy efficiency subject to QoS requirements for CUEs and DDPs. The proposed framework employs federated learning-based decentralized training with global model aggregation to address the users' privacy. Further, a double-deep Q-network is used for intelligent resource allocation. The proposed approach prevents the overestimation bias of the target value through the decoupling of action selection and target estimation, which enhances the learning stability of agents. Simulation results show that the average energy efficiency of the system is improved by 16.63%, 34.39%, and 73.53% compared to MAAC, D3PG, and DQN scheduling, respectively. Moreover, the proposed method achieves 11.65%, 24.78%, and 47.29% higher sum rates than MAAC, D3PG, and DQN scheduling, respectively. In conclusion, the proposed F-DDQN method maintains QoS satisfaction requirements for cellular and D2D users and shows superior performance compared with existing DRL schemes, which makes it a robust and scalable solution for energy-efficient resource allocation in D2D-assisted communication towards self-sustainable networks. The proposed framework could be extended to multi-objective learning for spectrum-energy-efficient resource allocation in future studies. However, when a massive number of agents are deployed, the complexity of both training and aggregation can be significant which could be an interesting investigation to achieve optimal spectrum-energy efficiency.

REFERENCES

[1] N. A. Alhaj, M. F. Jamlos, S. A. Manap, S. Abdelsalam, A. A. Bakhit, R. Mamat, M. A. Jamlos, M. S. M. Gismalla, and M. Hamdan, "Integration of hybrid networks, AI, ultra massive-MIMO, THz frequency, and FBMC modulation toward 6G requirements: A review," *IEEE Access*, vol. 12, pp. 483–513, 2024, doi: 10.1109/ACCESS.2023.3345453.

[2] V. K. Quy, D. C. Nguyen, D. Van Anh, and N. M. Quy, "Federated learning for green and sustainable 6G IIoT applications," *Internet Things*, vol. 25, Apr. 2024, Art. no. 101061, doi: 10.1016/j.iot.2024.101061.

[3] M. Xu, H. Du, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, A. Jamalipour, D. I. Kim, X. Shen, V. C. M. Leung, and H. V. Poor, "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1127–1170, 2nd Quart., 2024, doi: 10.1109/comst.2024.3353265.

[4] H. Li, X. Li, M. Zhang, and B. Ulziinyam, "System-wide energy efficient computation offloading in vehicular edge computing with speed adjustment," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 2, pp. 701–715, Jun. 2024, doi: 10.1109/TGCN.2023.3349273.

[5] K. Z. Shen, D. K. C. So, J. Tang, and Z. Ding, "Power allocation for NOMA with cache-aided D2D communication," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 529–542, Jan. 2024, doi: 10.1109/TWC.2023.3279266.

[6] T. Chen, X. Zhang, M. You, G. Zheng, and S. Lambotharan, "Federated learning enabled link scheduling in D2D wireless networks," *IEEE Wireless Commun. Lett.*, vol. 13, no. 1, pp. 89–92, Jan. 2024, doi: 10.1109/LWC.2023.3321500.

[7] Z. Zhou, M. Dong, K. Ota, J. Wu, and T. Sato, "Energy efficiency and spectral efficiency tradeoff in device-to-device (D2D) communications," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 485–488, Oct. 2014, doi: 10.1109/LWC.2014.2337295.

[8] A. Bhardwaj and S. Agnihotri, "Energy- and spectral-efficiency tradeoff for D2D-multicasts in underlay cellular networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 546–549, Aug. 2018, doi: 10.1109/LWC.2018.2794353.

[9] C. Zhang, C. Wu, M. Lin, Y. Lin, and W. Liu, "Proximal policy optimization for efficient D2D-assisted computation offloading and resource allocation in multi-access edge computing," *Future Internet*, vol. 16, no. 1, p. 19, Jan. 2024, doi: 10.3390/fi16010019.

[10] H. Yin, Y. Lyu, L. Xu, and T. A. Gulliver, "Intelligent optimization algorithm for green IoV networks based on SSA," *IEEE Trans. Veh. Technol.*, vol. 1, no. 1, pp. 1–10, Nov. 2024, doi: 10.1109/TVT.2023.3347744.

[11] X. Li, J. Zhang, and C. Pan, "Federated deep reinforcement learning for energy-efficient edge computing offloading and resource allocation in industrial Internet," *Appl. Sci.*, vol. 13, no. 11, p. 6708, May 2023, doi: 10.3390/app13116708.

[12] Z. Ji, Z. Qin, and X. Tao, "Meta federated reinforcement learning for distributed resource allocation," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7865–7876, Jul. 2024, doi: 10.1109/twc.2023.3345363.

[13] M. Banafaa, I. Shayea, J. Din, M. Hadri Azmi, A. Alashbi, Y. Ibrahim Daradkeh, and A. Alhammadi, "6G mobile communication technology: Requirements, targets, applications, challenges, advantages, and opportunities," *Alexandria Eng. J.*, vol. 64, pp. 245–274, Feb. 2023, doi: 10.1016/j.aej.2022.08.017.

[14] Z. Li, C. Xu, Z. Zhang, and R. Wu, "Deep reinforcement learning based trajectory design and resource allocation for task-aware multi-UAV enabled MEC networks," *Comput. Commun.*, vol. 213, pp. 88–98, Jan. 2024, doi: 10.1016/j.comcom.2023.11.006.

[15] Y. Dai, J. Zhao, J. Zhang, Y. Zhang, and T. Jiang, "Federated deep reinforcement learning for task offloading in digital twin edge networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 3, pp. 2849–2863, May 2024, doi: 10.1109/TNSE.2024.3350710.

[16] H. M. F. Noman, E. Hanafi, K. A. Noordin, K. Dimyati, M. N. Hindia, A. Abdrabou, and F. Qamar, "Machine learning empowered emerging wireless networks in 6G: Recent advancements, challenges and future trends," *IEEE Access*, vol. 11, pp. 83017–83051, 2023, doi: 10.1109/ACCESS.2023.3302250.

[17] Y. Song, Y. Xiao, Y. Chen, G. Li, and J. Liu, "Deep reinforcement learning enabled energy-efficient resource allocation in energy harvesting aided V2X communication," presented at the *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2022.

[18] P.-G. Ye, Y.-G. Wang, and W. Tang, "S-MFRL: Spiking mean field reinforcement learning for dynamic resource allocation of D2D networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 1032–1047, Jan. 2023, doi: 10.1109/TVT.2022.3203050.

[19] A. Yarali, "Artificial intelligence and machine learning in the era of 5G and 6G technology," in *From 5G to 6G: Technologies, Architecture, AI, and Security*. IEEE, 2023, pp. 65–72, doi: 10.1002/9781119883111.ch4. [Online]. Available: https://ieeexplore.ieee.org/document/10174776

[20] W. Jiang, D. Feng, Y. Sun, G. Feng, Z. Wang, and X.-G. Xia, "Joint computation offloading and resource allocation for D2D-assisted mobile edge computing," *IEEE Trans. Services Comput.*, vol. 1, no. 2, pp. 1–14, Oct. 2022, doi: 10.1109/TSC.2022.3190276.

[21] S. Zafar, S. Jangsher, and A. Zafar, "Federated learning for resource allocation in vehicular edge computing-enabled moving small cell networks," *Veh. Commun.*, vol. 45, Feb. 2024, Art. no. 100695, doi: 10.1016/j.vehcom.2023.100695.

[22] S. Liu, P. Guan, J. Yu, and A. Taherkordi, "FedSSC: Joint client selection and resource management for communication-efficient federated vehicular networks," *Comput. Netw.*, vol. 237, Dec. 2023, Art. no. 110100, doi: 10.1016/j.comnet.2023.110100.

[23] V. Vishnoi, I. Budhiraja, S. Gupta, and N. Kumar, "A deep reinforcement learning scheme for sum rate and fairness maximization among D2D pairs underlaying cellular network with NOMA," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 1–17, Mar. 2023, doi: 10.1109/TVT.2023.3276647.

[24] Y. Yuan, Z. Li, Z. Liu, Y. Yang, and X. Guan, "Double deep Q-network based distributed resource matching algorithm for D2D communication," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 984–993, Jan. 2022, doi: 10.1109/TVT.2021.3130159.

[25] D. Ron and J.-R. Lee, "DRL-based sum-rate maximization in D2D communication underlaid uplink cellular networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11121–11126, Oct. 2021, doi: 10.1109/TVT.2021.3106398.

[26] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021, doi: 10.1109/TWC.2020.3032991.

[27] D. Shi, L. Li, T. Ohtsuki, M. Pan, Z. Han, and H. V. Poor, "Make smart decisions faster: Deciding D2D resource allocation via Stackelberg game guided multi-agent deep reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 21, no. 12, pp. 4426–4438, Dec. 2022, doi: 10.1109/TMC.2021.3085206.

[28] H. Xiang, Y. Yang, G. He, J. Huang, and D. He, "Multi-agent deep reinforcement learning-based power control and resource allocation for D2D communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1659–1663, Aug. 2022, doi: 10.1109/LWC.2022.3170998.

[29] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1828–1840, Feb. 2020, doi: 10.1109/TVT.2019.2961405.

[30] B. Gu, X. Zhang, Z. Lin, and M. Alazab, "Deep multiagent reinforcement-learning-based resource allocation for Internet of Controllable things," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3066–3074, Mar. 2021, doi: 10.1109/JIOT.2020.3023111.

[31] J. Huang, Y. Yang, G. He, Y. Xiao, and J. Liu, "Deep reinforcement learning-based dynamic spectrum access for D2D communication underlay cellular networks," *IEEE Commun. Lett.*, vol. 25, no. 8, pp. 2614–2618, Aug. 2021, doi: 10.1109/LCOMM.2021.3079920.

[32] J. Huang, Y. Yang, Z. Gao, D. He, and D. W. K. Ng, "Dynamic spectrum access for D2D-enabled Internet of Things: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17793–17807, Sep. 2022, doi: 10.1109/JIOT.2022.3160197.

[33] X. Li, G. Chen, G. Wu, Z. Sun, and G. Chen, "Research on multi-agent D2D communication resource allocation algorithm based on A2C," *Electronics*, vol. 12, no. 2, p. 360, Jan. 2023, doi: 10.3390/electronics12020360.

[34] M. A. A. Khan, H. M. Kaidi, N. Ahmad, and M. U. Rehman, "Sum throughput maximization scheme for NOMA-enabled D2D groups using deep reinforcement learning in 5G and beyond networks," *IEEE Sensors J.*, vol. 23, no. 13, pp. 15046–15057, Jul. 2023, doi: 10.1109/JSEN.2023.3276799.

[35] Z. Ji, A. K. Kiani, Z. Qin, and R. Ahmad, "Power optimization in device-to-device communications: A deep reinforcement learning approach with dynamic reward," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 508–511, Mar. 2021, doi: 10.1109/LWC.2020.3035898.

[36] S. Muy, D. Ron, and J.-R. Lee, "Energy efficiency optimization for SWIPT-based D2D-underlaid cellular networks using multiagent deep reinforcement learning," *IEEE Syst. J.*, vol. 16, no. 2, pp. 3130–3138, Jun. 2022, doi: 10.1109/JSYST.2021.3098860.

[37] A. Omidkar, A. Khalili, H. H. Nguyen, and H. Shafiei, "Reinforcement-learning-based resource allocation for energy-harvesting-aided D2D communications in IoT networks," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16521–16531, Sep. 2022, doi: 10.1109/JIOT.2022.3151001.

[38] M. A. Ouamri, G. Barb, D. Singh, A. B. M. Adam, M. S. A. Muthanna, and X. Li, "Nonlinear energy-harvesting for D2D networks underlaying UAV with SWIPT using MADQN," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1804–1808, Jul. 2023, doi: 10.1109/LCOMM.2023.3275989.

[39] C. Kai, X. Meng, L. Mei, and W. Huang, "Multi-agent reinforcement learning based joint uplink–downlink subcarrier assignment and power allocation for D2D underlay networks," *Wireless Netw.*, vol. 29, no. 2, pp. 891–907, Feb. 2023, doi: 10.1007/s11276-022-03176-6.

[40] E.-J. Han, M. Sengly, and J.-R. Lee, "Balancing fairness and energy efficiency in SWIPT-based D2D networks: Deep reinforcement learning based approach," *IEEE Access*, vol. 10, pp. 64495–64503, 2022, doi: 10.1109/ACCESS.2022.3182686.

[41] X. Wang, H. Shi, Y. Li, Z. Qian, and Z. Han, "Energy efficiency resource management for D2D-NOMA enabled network: A dinkelbach combined twin delayed deterministic policy gradient approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 1–16, Apr. 2023, doi: 10.1109/TVT.2023.3267452.

[42] D. Long, Q. Wu, Q. Fan, P. Fan, Z. Li, and J. Fan, "A power allocation scheme for MIMO-NOMA and D2D vehicular edge computing based on decentralized DRL," *Sensors*, vol. 23, no. 7, p. 3449, Mar. 2023, doi: 10.3390/s23073449.

[43] T. Zhang, K. Zhu, and J. Wang, "Energy-efficient mode selection and resource allocation for D2D-enabled heterogeneous networks: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1175–1187, Feb. 2021, doi: 10.1109/TWC.2020.3031436.

[44] J. Lee, F. Solat, T. Y. Kim, and H. V. Poor, "Federated learning-empowered mobile network management for 5G and beyond networks: From access to core," *IEEE Commun. Surveys Tuts.*, vol. 1, no. 1, pp. 1–16, 2nd Quart., 2024, doi: 10.1109/comst.2024.3352910.

[45] J. Taghia, F. Moradi, H. Larsson, X. Lan, A. Orucu, M. Ebrahimi, and A. Johnsson, "Congruent learning for self-regulated federated learning in 6G," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, no. 2, pp. 129–149, Aug. 2024, doi: 10.1109/TMLCN.2023.3347680.

[46] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Data-importance aware user scheduling for communication-efficient edge machine learning," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 1, pp. 265–278, Mar. 2021, doi: 10.1109/TCCN.2020.2999606.

[47] J. Du, B. Jiang, C. Jiang, Y. Shi, and Z. Han, "Gradient and channel aware dynamic scheduling for over-the-air computation in federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1035–1050, Apr. 2023, doi: 10.1109/JSAC.2023.3242727.

[48] Q. Guo, F. Tang, and N. Kato, "Federated reinforcement learning-based resource allocation for D2D-aided digital twin edge networks in 6G industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 19, no. 5, pp. 7228–7236, May 2023, doi: 10.1109/TII.2022.3227655.

[49] Q. Guo, F. Tang, and N. Kato, "Federated reinforcement learning-based resource allocation in D2D-enabled 6G," *IEEE Netw.*, vol. 2, no. 1, pp. 1–7, Jul. 2022, doi: 10.1109/MNET.122.2200102.

[50] L. Guo, J. Jia, J. Chen, A. Du, and X. Wang, "Deep reinforcement learning empowered joint mode selection and resource allocation for RIS-aided D2D communications," *Neural Comput. Appl.*, vol. 35, no. 25, pp. 18231–18249, Sep. 2023, doi: 10.1007/s00521-023-08745-0.

[51] İ. Yazici, I. Shayea, and J. Din, "A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems," *Eng. Sci. Technol., Int. J.*, vol. 44, Aug. 2023, Art. no. 101455, doi: 10.1016/j.jestch.2023.101455.

[52] Y. Xiao, Y. Song, and J. Liu, "Multi-agent deep reinforcement learning based resource allocation for ultra-reliable low-latency Internet of Controllable Things," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5414–5430, May 2023, doi: 10.1109/twc.2022.3233853.

[53] P. Kulkarni, "Introduction to reinforcement and systemic machine learning," in *Reinforcement and Systemic Machine Learning for Decision Making*. Hoboken, NJ, USA: Wiley, 2012, pp. 1–21.

[54] B. Banerjee, R. C. Elliott, W. A. Krzymień, and M. Medra, "Access point clustering in cell-free massive MIMO using conventional and federated multi-agent reinforcement learning," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, no. 1, pp. 107–123, Nov. 2023, doi: 10.1109/TMLCN.2023.3283228.
[55] J. Miao, X. Chai, X. Song, and T. Song, "A DDQN-based energy-efficient resource allocation scheme for low-latency V2V communication," in *Proc. IEEE 5th Int. Electr. Energy Conf. (CIEEC)*, May 2022, pp. 53–58, doi: 10.1109/CIEEC54735.2022.9846189.

HAFIZ MUHAMMAD FAHAD NOMAN (Graduate Student Member, IEEE) received the B.E. degree in electronic engineering from COMSATS University, Islamabad, Pakistan, in 2012, and the M.E. degree in computer engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, in 2016. He is currently pursuing the Ph.D. degree in electrical engineering (major in artificial intelligence and wireless communication) with the Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia. He has more than ten years of research and development experience in academia and industry. His research interests include machine learning, wireless networks, D2D communication, and energy-efficient 6G networks. He has been awarded the HEC Pakistan Overseas Scholarship for the Ph.D. degree, in 2022.

KAHARUDIN DIMYATI (Member, IEEE) received the B.E. degree from the University of Malaya, Malaysia, in 1992, and the Ph.D. degree from the University of Wales Swansea, U.K., in 1996. He is currently a Professor with the Department of Electrical Engineering, Faculty of Engineering, University of Malaya. Since joining the university, he has been actively involved in teaching, postgraduate supervision, research, and administration. To date, he has supervised 15 Ph.D. students and 32 master's students by research students. He has published over 100 journal articles. He is also a Professional Engineer and a Chartered Engineer. He is a member of IET and IEICE.

KAMARUL ARIFFIN NOORDIN (Senior Member, IEEE) received the B.Eng. (Hons.) and M.Eng. degrees from the University of Malaya, Kuala Lumpur, Malaysia, in 1998 and 2001, respectively, and the Ph.D. degree in communication systems from Lancaster University, U.K., in 2009. He is currently an Associate Professor with the Department of Electrical Engineering, University of Malaya. His research interests include resource allocation in wireless networks, cognitive radio networks, device-to-device communications, network modeling, and performance analysis.

EFFARIZA HANAFI (Senior Member, IEEE) received the B.Eng. degree (Hons.) in telecommunications from The University of Adelaide, Adelaide, Australia, in 2010, and the Ph.D. degree in electrical and electronic engineering from the University of Canterbury, Christchurch, New Zealand, in 2014. She joined the University of Malaya, Kuala Lumpur, Malaysia, where she is currently a Senior Lecturer of electrical engineering. She has authored or co-authored journal articles in ISI-indexed publications and refereed conference papers. Her research interests include multiple antennas systems, RFID, cognitive radio networks, cooperative communications in wireless communications, the Internet of Things, and 5G networks and beyond.

ATEF ABDRABOU (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada.

In 2010, he joined the Electrical Engineering Department, UAE University, Al-Ain, UAE, where he is currently a Professor. His research interests include smart grid communication, vehicular communication, self-organizing wireless networks, radio resource allocation, the Internet of Things/machine-to-machine communication, and machine learning applications in wireless communication. He was a technical program committee member and the session chair of many IEEE international conferences. He is an Associate Editor of IEEE TRANSACTIONS ON MOBILE COMPUTING.