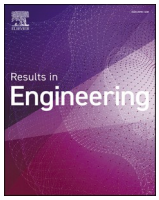
Results in Engineering 23 (2024) 102462

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Distributed resource optimisation using the Q-learning algorithm, in device-to-device communication: A reinforcement learning paradigm

Steffi Jayakumar , S. Nandakumar *

School of Electronics Engineering, Vellore Institute of Technology, Vellore, India

ARTICLE INFO

Keywords:
B5G
D2D communication
Resource allocation
Interference management
Machine learning-RL
Q-learning

ABSTRACT

In the context of wireless systems going forward, particularly in the Beyond 5G (B5G) era, where high data rates and low latency are critical, D2D communication is a pivotal technology that has many advantages. In D2D communication, the allocation of resources plays a critical role in achieving higher throughput while ensuring interference management, improved spectrum and energy efficiency, and system fairness. Conventional resource allocation methodologies encounter challenges in dynamically changing and diverse communication environments. To deal with the dynamic and unpredictable character of channel characteristics, we utilize a distributed iterative resource allocation technique based on the reinforcement learning (RL) approach that empowers the system to learn and adapt to the wireless environment autonomously. In this paper, we formulate a distributed Q learning-based RL method as its real-time learning capabilities, reduced communication overhead, and exploration efficiency make it well-suited for adapting to dynamic D2D environments compared to other RL methods. By applying the Q-learning method, the D2D devices act as learning agents striving to maximize the cumulative rewards. Through interactions with the environment and continuous learning from feedback, these agents adapt to real-time resource allocation decisions over time. Comparing our proposed Q-learning method to the state-of-the-art RL techniques, simulation results show improvements in energy and spectrum efficiency, latency, an increase in Jain's fairness index, and improvements in overall system throughput of about 6 %–8 %. The scalability is found to be 1.69 interpreting that Q-learning exhibits a good scalability as the throughput does not abruptly decrease for an increasing number of devices.

1. Introduction

In the current world, cellular and mobile technology have grown rapidly over the past 20 years, demonstrating their indispensable role in our daily lives [1]. The mobile traffic in the wireless cellular network has surged due to the growing popularity of wireless devices and ubiquitous wireless connections [2]. This accelerating growth in mobile traffic is expected to rise even more in the upcoming next-generation wireless networks where the world is expected to deploy smarter networks thereby accumulating many autonomous devices [3]. In this scenario, direct communication among the mobile users can offload the load and the traffic from the central entity of the cellular network architecture [4]. D2D communication is an emerging technology in which devices that are closer to each other geographically send data while partially or fully interacting with the base station. The D2D communication facilitates direct data transmission between the devices, eliminating the need for the Base station to relay data [5].

D2D communication is energy-efficient as it requires less time and power to transfer data due to the shorter distances between the communicating devices resulting in proximity gain. D2D devices ensure spectrum efficiency by utilizing the limited licensed spectrum resources allotted to cellular users [6]. D2D communication also assures system fairness, increased network capacity as well as enhanced network performance and improved throughput. D2D finds its applications in local data services like multimedia content transmissions, IoT-based services like Vehicle-to-Vehicle (V2V) and Machine-to-Machine (M2M) communication and acts as an ad-hoc network in disaster-hit areas [7]. However, the performance advantages of D2D communication are best realized in underlay mode where the spectrum is shared among the users as the networks are experiencing spectrum constraints because of the manifold increase in user demands and application services. Nonetheless, underlay D2D communication causes interferences to the cellular networks if the radio resources are not distributed effectively as shown in Fig. 1. Also, given that the future networks demand a greater number

* Corresponding author.
E-mail addresses: julianasteffi@gmail.com (S. Jayakumar), snandakumar@vit.ac.in (S. Nandakumar).

https://doi.org/10.1016/j.rineng.2024.102462
Received 17 April 2024; Received in revised form 13 June 2024; Accepted 23 June 2024
Available online 28 June 2024

of resources and data rates due to the employment of applications like multimedia and online gaming, video streaming and conferencing, vehicle-to-vehicle communication, machine-to-machine communication, personalized TVs, smart devices, Internet of Things (IoT), self-driven automatic cars, the limited resources should be managed effectively [8]. In today's smart world, autonomous vehicles in a smart city need to communicate with each other for collision avoidance, traffic management and cooperative driving to reduce latency and to improve reliability. Numerous devices such as traffic lights, autonomous vehicles and IoT sensors communicate with each other to enhance urban living. Effective resource allocation in such a dense network is crucial for minimizing interference and maximizing the efficiency of communication. IoT devices in a smart home or industrial IoT setup use RL to optimize communication schedules and routes. In 5G networks, user devices in a crowded area learn to allocate resources dynamically, selecting optimal channels and power settings to maximize throughput and minimize interference, leading to better user experience and network performance. Therefore, proper and efficient allocation of the limited radio resources is important to ensure reliable communication that increases the capacity and efficiency of the cellular network consequently reducing the interferences [9]. RL methods are practical in real-time wireless communication systems because they allow devices to make intelligent decisions in a dynamic and uncertain environment. Through interactions with the environment and continuous learning from feedback, the communication devices known as agents adapt to decisions related to real-time resource allocation over time.

In this paper, we propose a Q-learning method employing RL principles in allocating the optimal resources due to its faster exploration of resource allocation strategies, potentially resulting in higher throughput as it considers a wider range of actions and their consequences, which can be crucial for adapting to dynamic and changing wireless environments. Q-learning can handle many users of today's wireless network without performance degradation which makes it suitable for dense urban environments and large-scale deployments. The dynamic adaptation of Q-learning to real-time conditions helps in reducing interferences, leading to a more stable and more efficient network. Q-learning offers good scalability, improved latency, robustness, security, and mobility. Our research focuses on increasing throughput while establishing constraints to reduce interferences brought on by spectrum sharing. Q-learning is a method that uses a trial-and-error technique to get to the best optimal solution without requiring any prior knowledge of the environment. Our work in this paper extends our research work by tackling the research allocation problem by constraint optimization method. We in this paper, have compared our proposed Q-learning method with our previous work, which is one of the state-of-the-art techniques like State-Action-Reward-State-Action (SARSA) [10] and other conventional resource allocation methods.

1.1. Contributions

This paper focuses on the user-centric learning method of resource allocation. The significant contributions of this paper to the field of D2D communication are listed below.

- For the joint resource optimization of spectrum and power in a real-time, dynamic D2D-enabled wireless cellular communication system, a decentralized, distributed, and learning framework-based Q-learning method is proposed. The constraint that the interferences resulting from frequency reuse stay within limits is taken into consideration in our study.
- We assess the efficiency of the proposed Q-learning technique, considering the method's convergence, spectrum and power

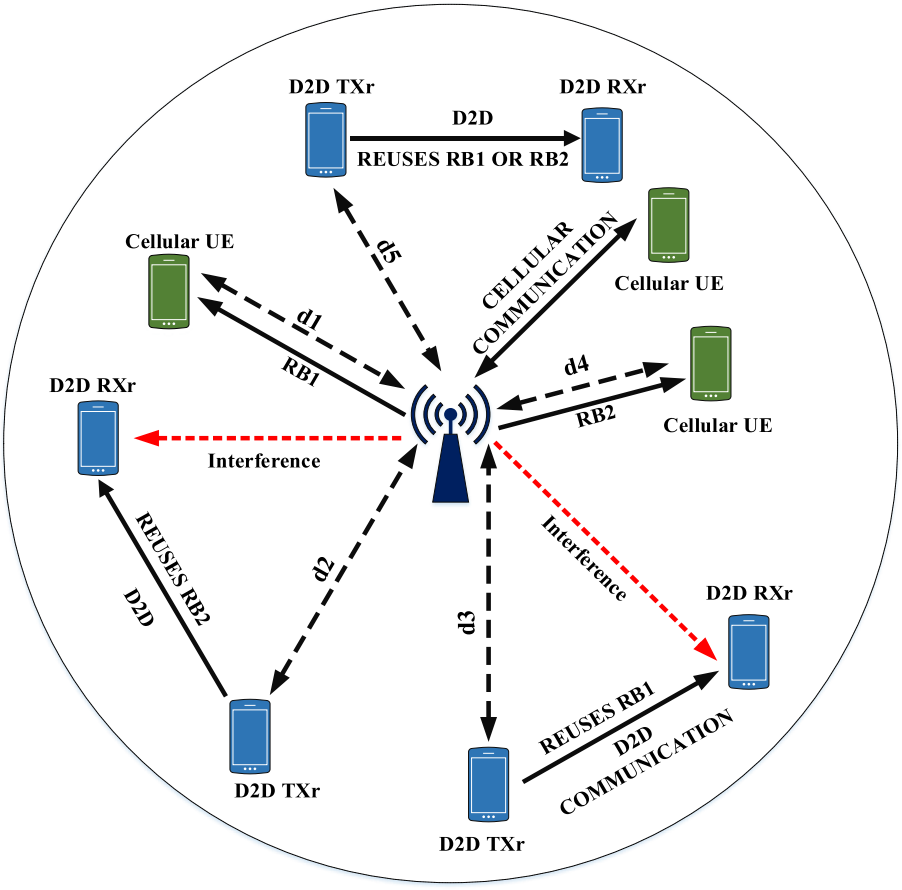


Fig. 1. System model for D2D-enabled wireless cellular communication that depicts the image of an underlay resource allocation scenario.

efficiency, fairness, and system throughput. The proposed approach is assessed against various resource allocation strategies including the state of art approaches like SARSA.

- It has been demonstrated that the efficiency of the proposed reinforcement learning-based Q-learning technique, which is applied in the joint assignment of spectrum and power resources, increases the throughput of D2D users as well as the overall throughput of the wireless system. The simulation results indicate that the suggested strategy achieves a good level of fairness in addition to better spectrum and power efficiencies. The communication system's fairness is determined using Jain's fairness index.
- Our proposed Q-learning method is designed to operate in dynamic environments, where channel conditions and user demands change over time. Our learning-based approach allows the system to maintain low latency even under varying network conditions.
- We address the scalability of our proposed method thus demonstrating its efficient implementation in a large-scale D2D network.
- By showcasing the efficacy of machine learning-based resource allocation, the future insight into the potential application of AI-based strategies for tackling D2D challenges in the communication environment is strongly emphasized.

The rest of this paper proceeds as follows: The recent literature on resource allocation techniques in D2D communication is reviewed in Section 2. In Section 3, the system model is illustrated and the optimization problem is formulated. The proposed spectrum and power allocation method is detailed in Section 4, along with an explanation of its background. The effectiveness of the proposed method is illustrated in Section 5. In Section 6, the results of the work are summarized and further potential possibilities for research are listed.

2. Related work

The rapid proliferation of wireless devices and the escalating demand for high-speed data transmissions have placed immense pressure on traditional cellular networks, recently. These greater requirements lead to higher bandwidth requirements in 5G and B5G because of which the researchers are driven to focus more on optimizing the scarce or limited resources. Effective research allocation lies at the heart of optimizing D2D communication systems, ensuring that spectrum resources are allocated judiciously to maximize performance. Table 1 presents an analysis and tabulation of the most recent resource allocation strategies and Table 2 gives the notations and symbols used in this paper.

To optimize the use of scarce resources and enhance energy and spectrum utilization, D2D communication is combined with uplink Non-Orthogonal Multiple Access (NOMA) as in Ref. [11], hence offering multiple access to D2D transmitters. It has been suggested to increase energy efficiency and decrease computations based on the energy of the UE by using a two-stage game approach and energy-aware screening techniques. In the first step of the method, energy harvesting is optimized, and in the second stage, resource allocation, network efficiency, and performance are evaluated while taking interference limits into account. By the application of NOMA as in Ref. [12], resource allocation is performed by user clustering, power control and mode selection. A swarm intelligence-based whale optimization algorithm is applied to find the optimal solution for the assignment power allocation in the NOMA (JRPAN) algorithm [13]. In addition to improving the data rate and fairness of the system compared to the Orthogonal Frequency division multiplexing (OFDM) approach, D2D user groups are created for resource reuse based on distance limitations.

By allowing systems to learn optimal strategies [10], demonstrates the potential of reinforcement learning in enhancing resource utilization and network performance. The SARSA method is used to optimize resource allocation by letting the agents—in this case, the D2D transmitters—learn how to assign the best resources through interactions with the environment. Energy efficient channel allocation for the D2D devices engaged in underlay multicast communication scenarios is discussed in Ref. [14] where a two-stage semi-distributed approach has been applied that gives better performance when compared to the pure optimal centralized method. For the multicast device-to-device (D2MD) communications group, a coalitional game model is used in the first stage to allocate channels followed by the fractional programming method that is used in the energy optimization that is handled by the central entity. To maximize network performance while maintaining Quality of Service (QoS), both the uplink and downlink resources are cooperatively controlled and distributed to the D2D users as in Ref. [15] that discusses the joint Uplink Downlink Resource allocation technique (JUDRA).

The sequential geometric programming methodology is applied to assign subcarriers followed by the optimal power allocation procedure. Resource allocation based on the Channel State Information (CSI) is said to be more effective and easier to manage. For networks with an uncertain CSI, a robust strategy for allocating resources is needed. In Ref. [16], a method based on Support Vector Clustering (SVC) is applied to address the problem of uncertain CSI in a D2D underlaying cellular network by designing the uncertain CSI as a complex uncertainty set. Uncertain CSI problem encountered during the assignment of both channel and power allocation is addressed in Ref. [17]. Both single and multiple antenna configurations are used in a centralized and distributed manner to perform joint power and channel assignments.

By focusing on these strategies, D2D communication can be significantly optimized, leading to improved network performance, user experience, and resource utilization. Throughput and latency in the context of network performance and resource allocation, is an outcome of an optimization problem. Optimization involves finding the best solution from a set of possible solutions. There are a variety of real-world situations where optimization is useful to accomplish tasks in the best possible way. Manufacturing, production, engineering, transportation, communication, maintenance, and scheduling are a few real-world instances of the optimal approach [20]. Engineering optimization problems are suitable for large dimensional situations with imprecise data [21]. Optimization tasks in various industries also aim at overcoming the challenges like complex, convergence speed and proneness to local optima [22,23]. Likewise, reinforcement learning methods that are a solution to optimization problems are applied to explore more complex and dynamic environments. The reinforcement learning-based method finds a broad range of applications in communication, electricals and electronics [24], aircraft designing, shipping, and engineering, etc [25]. Several computational and feedback mechanisms play an important role in diverse scientific and engineering domains to attain optimization [26].

The studies surveyed collectively underscore the capacity of distributed resource allocation algorithms to dynamically optimize resource allocation decisions. These algorithms consider factors such as throughput, spectrum efficiency, interference management, fairness, and user satisfaction in a dynamic environment. Though RL has been implemented in a remarkable number of works, to the best of the author's knowledge, very few literature sources have integrated power and spectrum allocation for the enhancement of throughput and fairness while maintaining good scalability and remarkable latency in a D2D-enabled system integrated with constrained learning techniques. Thus, our primary objective is to present a distributed reinforcement learning (RL) based Q-learning method that chooses the best resources to satisfy the interference constraints and outperform in terms of performance metrics.

3. System model and problem formulation

Our work focuses on assigning resources for a D2D-enabled wireless cellular communication system. We have considered a single-cell wireless network. The D2D and cellular user devices share the same limited channel resources. The system consists of one central base station to

Table 1
Comparative analysis of recent literature on Resource optimization in D2D communication.

Papers	Resource Block	Power	Throughput	Uplink	Downlink	Energy Efficiency	Spectrum Efficiency	Centralized RA	Distributed RA	QoS/ QoE	Fairness	Remarks/Discussion
[10]	✓	✓	✓			✓	✓	×	✓	✓	✓	The SARSA resource allocation method, which is based on distributed reinforcement learning, is put forth. The D2D transmitters perform better when choosing the ideal RB-power level by applying trial and error approach
[11]	✓	✓		✓		✓	✓	×	✓	✓		In a D2D communication system integrated with Uplink NOMA, the joint distributed resource allocation problem is investigated by a two-stage game strategy. To minimize the computational complexities and signalling overheads, an approximation strategy is developed by creating a noncooperative game between the D2D-cellular user groups. The computation complexities are further reduced by applying an energy-aware screening technique.
[12]	✓	✓	✓		✓		✓			✓		The joint challenges of D2D resource allocation for power control, user clustering, and mode selection are analysed. To solve the optimization problem with effective solutions, the whale optimization algorithm, a form of swarm intelligence, is used.
[13]	✓	✓	✓		✓					✓	✓	A joint allocation of resources and power in the NOMA-enabled D2D system is proposed. The cellular users are first assigned resources based on the user's applications to ensure QoS. The performance evaluation is carried out based on the throughput, fairness, QoS and QoE.
[14]	✓	✓		✓		✓			✓	✓	✓	A two-stage semi-distributed approach is proposed. In the first stage, channel allocation is performed for the multicast D2D group using a cooperative

(*continued on next page*)

Table 1 (*continued*)

Papers	Resource Block	Power	Throughput	Uplink	Downlink	Energy Efficiency	Spectrum Efficiency	Centralized RA	Distributed RA	QoS/ QoE	Fairness	Remarks/Discussion
												coalitional game framework that allows co-channel transmission over a set of shared RBs. In the second stage, the central entity uses fractional programming to determine the optimal transmission power for each user in the system. The results are proven to be more in line with the optimal centralized method.
[15]	✓	✓	✓	✓	✓		✓			✓		The Joint uplink-downlink resource allocation (JUDRA) strategy aims to ensure that the near-optimal solution in polynomial time is obtained by assigning the subchannels and power by a sequential geometric programming method.
[16]	✓	✓	✓	✓		✓	✓			✓		The SVC-based approach is used to model the channel state information of the system that aids the robust resource allocation framework that aids in maximizing the cellular throughput.
[17]	✓	✓	✓	✓		✓	✓				✓	The co-channel interference and energy efficiency problem is addressed by the proposed fuzzy clustering and game theory methods. The D2D devices are grouped by fuzzy clustering such that the system throughput is improved while the interference is reduced. The energy efficiency is improved as the user transmission power is optimized by game theory.
[18]	✓	✓	✓	✓				✓	✓	✓	✓	Both centralized and distributed resource allocation techniques have been applied for performing joint channel and power allocation under imperfect CSI conditions. The main objective is to improve the throughput and fairness of the system while the QoS of the system is maintained.
[19]	✓	✓	✓	✓			✓		✓	✓		Stackelberg game theory is applied for interference coordination in a joint

(*continued on next page*)

Table 1 (*continued*)

Papers	Resource Block	Power	Throughput	Uplink	Downlink	Energy Efficiency	Spectrum Efficiency	Centralized RA	Distributed RA	QoS/ QoE	Fairness	Remarks/Discussion
												channel and power allocation of a D2D-enabled cellular communication system. The channel allocation is performed by the Base station while the interference management is performed in a distributed manner.

RA- QoS- Quality of Service QoE- Quality of Experience.

Table 2
Summary of notations.

Notation	Description
f_c	Carrier frequency
P_D	Transmission power
$P^r_{d_u}$	The transmission power of the uth D2D user employing the rth RB
$P^r_{CU_c}$	The transmission power of the cth cellular user employing the rth RB
$P^r_{d_j}$	The transmission power of the jth D2D user employing the rth RB
$PL_{x,y}$	Pathloss between the transmitting device, x and the receiving device, y
$PL_{d_u d_v}$	Pathloss between the D2D user devices u and v
$PL_{eNB,j}$	Pathloss between the eNB and the user device, j in the network
$g_{x,y}$	Channel gain between the communicating devices, x and y
$g_{eNB,j}$	Channel gain between the base station, eNB and the user device, j in the communication system
$g_{d_u d_v}$	Channel gain between the communicating D2D user devices, d_u and d_v
$G(d_{T,}d_R)$	Channel gain between the communicating devices in D2D mode
$G^r_{d_{uT}d_{uR}}$	Channel gain between the D2D transmitter d_{uT} and receiver d_{uR} in rth RB
$G^r_{CU_c d_{uR}}$	Channel gain between the cth cellular UE and the D2D user receiver, d_{uR} using rth RB
$G^r_{d_{jT}d_{uR}}$	Channel gain between the D2D communicating devices, d_{jT} and d_{uR} operating in rth RB
G^r_{eNB,CU_c}	Channel gain between the central entity, *eNB* and the cellular user, CU_c
$G^r_{d_{uT}eNB}$	Channel gain between D2D user d_{uT} and *eNB* operating in rth RB
$G^r_{d_{jT}eNB}$	Channel gain between the transmitter in D2D mode, d_{jT} and the eNB in rth RB
γ_{Drx}	SINR of *Drx*, D2D receiver
$\gamma^r_{D_u}$	SINR of the D2D user, D_u over the RB, r.
$\gamma^r_{d_{uT}d_{uR}}$	SINR between the transmitting and receiving devices, d_{uT} and d_{uR} over the resource block, r
$\gamma^r_{CU_C}$	SINR of the cellular user, CU_C
γ_c, γ_d	SINR of the communication system in cellular and D2D modes
$R^r_{d_{uT}}$	The reward of the of the agent, D2D transmitter, d^r_{uT}
τ_0, τ_1, τ_2	Minimum threshold values of the channel gain and SINR
σ^2	Noise variance
$Z^{(r,p)}_{d_N}$	Binary variable for decision
I_{D_R}	Interference raised by D2D users because of resource sharing
I_{d_N}	Interference raised by D2D transmitter because of resource sharing
I_{TH}	Threshold Interference value
α	Learning rate
$\gamma, \gamma_1, \gamma_2$	Discount factors

which each of the M cellular user devices and N D2D user devices are connected. The set of cellular users and the D2D user devices are denoted as $C = \{C_1, C_2, \ldots, C_M\}$ and $D = \{D_1, D_2, \ldots, D_N\}$ wherein, C_M and D_N represent the highest number of cellular and D2D users in the system. The user devices in the communication system model are distributed randomly in the cell as shown in Fig. 2. A D2D transmitter, D_T and a D2D receiver, D_R comprise a D2D communication pair. Every D2D transmitter and receiver pair is equipped with a single antenna. As the cellular and D2D devices share the spectrum resources, the number of resource blocks is assumed to be equal to the number of cellular devices in the cell. The set of resource blocks available for assignment is given as $B = \{b_1, b_2, \ldots, b_M\}$. Each resource block is assigned to one

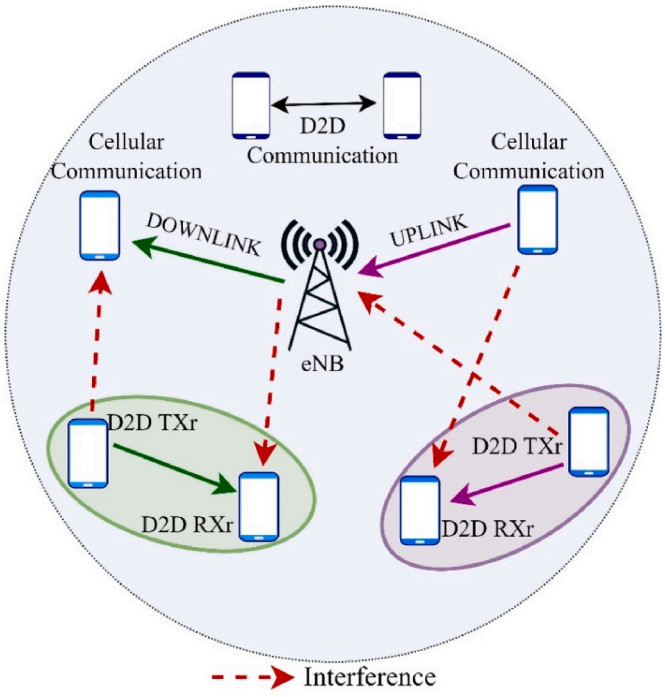


Fig. 2. Cross-tier Interferences in a D2D-enabled cellular wireless communication system due to spectrum reuse.

cellular user in the cell and to one or many D2D pairs in the cell. A D2D pair can be assigned with one resource block, which is the smallest unit of spectrum resource. The power needed to transmit data is chosen from a limited range of power values, $P = \{P_1, P_2, \ldots, P_M\}$. Each D2D transmitter, d_T is allocated with the power and spectrum combinations from the power and resource block sets. The available spectrum levels are of the scarce quantity, therefore reuse of spectrum resources is unavoidable. The spectrum reuse strategy also leads to interference [27]. The interference threshold value, I_{TH} is set such that its value is very low so that it is negligible and the QoS of the system is maintained as well. Interferences can be cross-tier and co-tier. When the uplink resource is being shared, by the D2D transmitter, d_{uT} causes Cross-tier interference at the base station. Co-tier interference is present at the receiver side of the D2D device, d_{uR} by other D2D device transmitters, d_{jT} from d_{jT} $\forall$j, n = 1,2, …, N and $j \neq n$. d_{uR} and d_{jT} cause co-tier interference when they are sharing the same resource of the cellular devices operating at the resource block, r. When a downlink resource-sharing scenario is applied, the signal from the base station causes interference to the D2D receiver as in Fig. 2.

In D2D communication, Channel State Information (CSI) plays a

crucial role in optimizing network performance and resource allocation. In a dynamically changing environment, when CSI is unknown, it presents significant challenges in optimizing wireless communication performance. In a reinforcement learning scenario like our proposed Q-learning method, even with unknown CSI, the algorithm can still learn the optimal resource allocation strategies. The agent interacts with the environment directly, taking actions (allocating resources like channel and power) and receiving rewards like latency and throughput. Over time, the agent can learn which actions yield the best long-term rewards, implicitly adapting to the channel conditions without explicit CSI.

Let us assume that the *i*th D2D pair is denoted as D2D *(i)*, where, $i \varepsilon \{1,2, \ldots, k\}$. The location of one D2D pair is given as (x^i_{TX}, y^i_{TX}) and (x^i_{RX}, y^i_{RX}). The Euclidean formula as given in Eqn (1), is used to determine the distance between the *i*th D2D pair.

$$dist = \sqrt{(x^i_{TX}, - x^i_{RX})^2 + (y^i_{TX} - y^i_{RX})^2} \tag{1}$$

We assume that the communication environment is urban and apply the Rayleigh fading path-loss model expressed in dB which is given as

$$PL\ in\ dB = 36.7\ log_{10}(dist) + 22.7 + 26\ log_{10}(f_c) \tag{2}$$

where, *dist* is the distance in meters between the D2D users and f_c is the carrier frequency in GHz.

The channel gain between the communicating devices in the system is calculated using the formula

$$g_{x,y} = 10^{-PL_{x,y}/10} \tag{3}$$

Let d_u and d_v be the transmitting uth and vth D2D user devices involved in data transmission. The channel condition is calculated between the central entity and the user device and between the D2D devices.

$$g_{eNB,j} = 10^{-PL_{eNB,j}/10} \tag{4}$$

$$g_{d_u d_v} = 10^{-PL_{d_u d_v}/10} \tag{5}$$

Let us consider that the devices are communicating in the rth RB. Due to this spectrum sharing, interference arises. The Signal to interference plus noise ratio (SINR) values are used to determine the quality of the communication link.

According to the below equation, the SINR of the D2D receiver, γ_{Drx} is given as

$$\gamma_{Drx} = \frac{Transmit\ power, P_D * channel\ gain, G(d_T, d_R)}{\sigma^2 + I_{D_R}} \tag{6}$$

The SINR between the transmitting and the receiving devices, d_{uT} and d_{uR} over the resource block *r* is given as

$$\gamma^r_{d_{uT}d_{uR}} = \frac{P^r_{d_u} G^r_{d_{uT}d_{uR}}}{\sigma^2 + P^r_{CU_c} G^r_{CU_c d_{uR}} + \sum_{j=1, j\neq d}^{N} P^r_{d_j} G^r_{d_{jT}d_{uR}}} \tag{7}$$

Additionally, the SINR of the cellular user involved in conventional communication is determined and is given by

$$\gamma^r_{CU_C} = \frac{P^r_{CU_c}\ G^r_{eNB,CU_c}}{\sigma^2 + \sum_{d\in D} P^r_{d_u} G^r_{d_{uT}eNB} + \sum_{j=1, j\neq d}^{N} P^r_{d_j} G^r_{d_{jT}eNB}} \tag{8}$$

The noise variance σ^2 in the SINR equations is given as

$$\sigma^2 = NoBW_{RB} \tag{9}$$

The interference value should be maintained below a threshold value I_{TH} to maintain the performance of the system and for efficient transmission of data. The interference that is created by the D2D transmitter is given as

$$I_{d_N} = \sum_{D=1}^{N} Z_{d_N}\ P^r_{d_u} G^r_{d_{uT}eNB} \leq I_{TH} \tag{10}$$

To analyse the performance of the system, throughput values are important. The higher the throughput value, the greater the system performance. The SINR values are used in the calculation of the system throughput [28]. The throughput for a specific user operating over the resource block *r*, can be mathematically computed by applying Shannon's formula as follows

$$T_{cellular} = BW_{RB} \sum_{c=1}^{C} log_2(1+\gamma_c) \tag{11}$$

$$T_{D2D} = BW_{RB} \sum_{d=1}^{D} log_2(1+\gamma_d) \tag{12}$$

$$T_{total} = BW_{RB} \left[\sum_{c=1}^{C} log_2(1+\gamma_c) + \sum_{d=1}^{D} log_2(1+\gamma_d) \right] \tag{13}$$

The binary form of the RB selection by the D2D user can be expressed as

$$Z^{(r,p)}_{d_N} = \begin{cases} 1, if\ the\ transmitting\ device\ is\ transmitting\ over\ the\ \mathrm{r^{th}}\ RB\ with\ power\ level, p \\ 0, otherwise \end{cases} \tag{14}$$

The achievable throughput is

$$T_{D2D} = BW_{RB} \sum_{d=1}^{D} Z_{D_N}\ log_2(1+\gamma_d) \tag{15}$$

The main objective is to improve the system performance and throughput. Given below is the optimization problem of the system

$$\max_{r\epsilon R} \sum_{r=1}^{R} Z_{D_N}\ BW\ log_2(1+\gamma_d) \tag{16}$$

To improve the performance of the system, interference constraint is considered. The interferences caused by the D2D transmitters are maintained less than or equal to a threshold value while keeping the SINR values above a minimal SINR threshold value.

The optimal value is obtained when the following constraints are met. The first two constraints create a limitation that only one RB can be used by a D2D user at a particular instant of time.

$$\text{C.1 : Interference } I_{d_N} \leq I_{TH}\ \forall \mathrm{r} \in \mathrm{R}, d_u \in D \tag{17a}$$

$$\text{C.2 : Binary decision variable } Z^{(r,p)}_{d_N} \varepsilon \{0,1\} \forall \mathrm{r} \in \mathrm{R}, d_u \in D \tag{17b}$$

$$\text{C.3 : SINR value } \gamma^r_{CU_C} \geq \gamma_{min}\ \forall \mathrm{r} \in \mathrm{R}, CU_C \in C \tag{17c}$$

The resource allocation problem can be expressed as

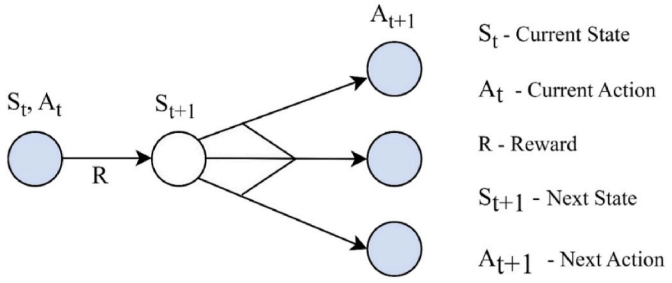


Fig. 3. Q learning.

$$max \sum_{r=1}^{R} BW * \left\{ \sum_{c=1}^{C} log_2(1+\gamma_c) + \sum_{d=1}^{D} log_2(1+\gamma_d) \right\} \text{ for } I^r < I_{th}^r \text{ for all the values of } r \tag{18}$$

4. Reinforcement learning strategy for allocation of resources

To achieve the maximum throughput value in a dynamic environment, fixed resource allocation strategies are challenging. In a real-time scenario, distributed and dynamic resource allocation methods are the most significant and efficient methods [29]. One such autonomous method that learns from past trials and conditions, and adapts to varying system conditions, is the machine learning-based resource allocation Q-learning method as shown in Fig. 3. Machine learning is a recent and trending method that appears to have a promising future in wireless communication systems [30]. This section provides a brief definition and application of reinforcement learning, which is one of the subsets of machine learning.

Integrating network cellular communication with machine learning D2D communication enhances the efficiency and effectiveness of D2D operations. Cellular networks provide a structured framework for resource allocation, interference management, and QoS assurance, ensuring D2D communication can seamlessly integrate with conventional cellular traffic. Machine learning further augments this integration by enabling an intelligent autonomous decision-making process that predicts network and user conditions, optimizes resource allocations dynamically, and minimizes interference. Hybridization of D2D and cellular communication using machine learning involves integrating direct communication between devices with traditional cellular networks to enhance overall network efficiency. Devices collect real-time data about network conditions, such as signal strength, path loss, channel gain, interference levels, and device locations. Channel quality and device proximity are calculated. Our proposed Q-learning method is applied to dynamically allocate resources between D2D and cellular communication. These models train the devices to decide the best resource allocation strategy to optimize communication efficiency. By intelligently managing resources, the system maximizes throughput, reduces latency, and improves fairness, leading to robust and efficient connectivity in complex and dynamic network environments.

As mentioned earlier, Reinforcement learning offers the advantages of adaptability to dynamic environments, real-time optimization, and the ability to handle complex decision-making processes, making it a powerful robust approach for optimizing resource allocation in Device-to-Device communication networks [31]. The traditional methods rely on fixed, stable, and well-defined environments while the proposed Q-learning-based RA learns, and updates continuously based on real-time interactions making it suitable for the real-time, more complex, and ever-changing nature of modern wireless communication networks. This allows it to handle varying interference levels, user mobility and fluctuating traffic demands more efficiently. The proposed learning algorithm determines the optimal policy by which the decisions are taken without any prior knowledge of the environment. This makes the method to be a trial-and-error one.

For a wireless communication system which keeps on changing over time is practically impossible to predict and model. For such a dynamic system, our proposed Q-learning is a significant technique for decision-making. The agents in this case are the D2D users who learn about the environment and take actions [32]. The actions are a set of channel and power values that must be assigned to the users. The Q-learning analyses the quality of the action from its feedback. The feedbacks are in the form of rewards which is the throughput in our case. The system gets a positive reward for an optimal assignment of the resources and a negative or no reward at all otherwise. Consequently, the user learns from the feedback about the environment and the optimal policy that gives a higher reward. The scenario that we have considered is a multi-agent system which consists of multiple cellular and D2D devices that are distributed randomly in the cell. In the realm of D2D communication, Q-learning serves as a powerful framework for dynamic and intelligent decision making whose framework is demonstrated in Fig. 4.

Initially, a Q-table is established encompassing state-action pairs representing the complex D2D communication environment. States include information on D2D device locations, signal quality and other factors while the actions could involve choices like transmission power levels and channel selections. The iterative process begins by observing the current state and then selecting an action based on its observation. Once the action is taken, it is executed within the communication environment. The system then measures the effectiveness of this action through rewards, which consider factors like signal quality, interference with other devices energy consumption and data transmission quality.

These rewards provide essential feedback on how well the selected action performs in the current state. Subsequently, using the Q-learning algorithm, the Q-value for the current state-action pair is updated which incorporates the obtained reward and the expected future rewards. Through this iterative process, D2D devices learn and adapt their strategies to make more informed decisions over time, ultimately improving communication efficiency, reliability, and overall network performance. This cycle repeats until the algorithm converges allowing the D2D communication system to operate optimally in each environmental context. The algorithm maintains a Q-table that is built and iteratively updated with Q-values representing the expected future rewards for each state-action pair. Once the Q-values have converged, the agents use this table to make decisions, always selecting the actions that have the highest Q-value for the current state. The flowchart that outlines the experimental procedure of the resource allocation of D2D communication by the application of Q-learning is provided in Fig. 5.

The elements of Q-learning include Agents, State (S), Environment, Action (A), Reward (R), state-action pair (S, A), Q-Table (Q (S, A)), Learning rate, Discount factor and policy. The Q-learning elements work together in the Q-learning framework to enable the agent to learn sequential decisions in an uncertain and dynamic communication environment to maximize the rewards as shown in Fig. 6.

AGENT: The main responsibility of the agent in the proposed work is to find the optimal resources for operation. The agent makes interactions with the operating environment, learns, retunes, and makes decisions. The multiple agents in the system have the goal of selecting the optimal policy simultaneously. Here, all the D2D transmitters are agents that are involved in the learning and decision-making procedures. Therefore, the agents are the ones that are responsible for the learning-based allocation process.

STATE: The state of the agent is its operating environmental condition. At a specific time, t the channel and the power level of D2D communication describe the state. The state space of our proposed method is three-dimensional which includes the number of D2D pairs, spectrum, and power level of the D2D transmitter and receiver pairs. Therefore, the state is defined as

$$S_i^t = \gamma_{D_u}^r \cup G_{D_{uk}} \cup G_{eNB,j} \tag{19}$$

The parameters that are used in determining the QoS of the system

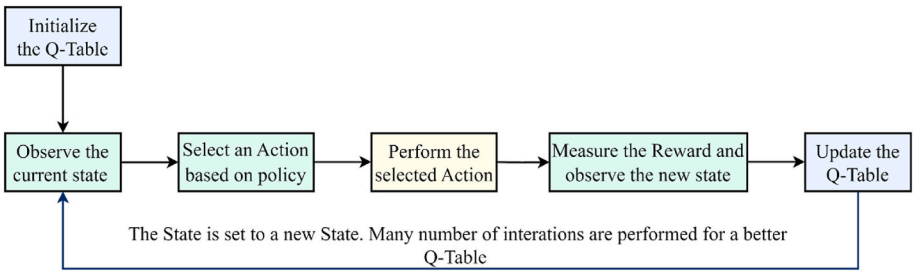


Fig. 4. Proposed Q-learning based resource allocation framework.

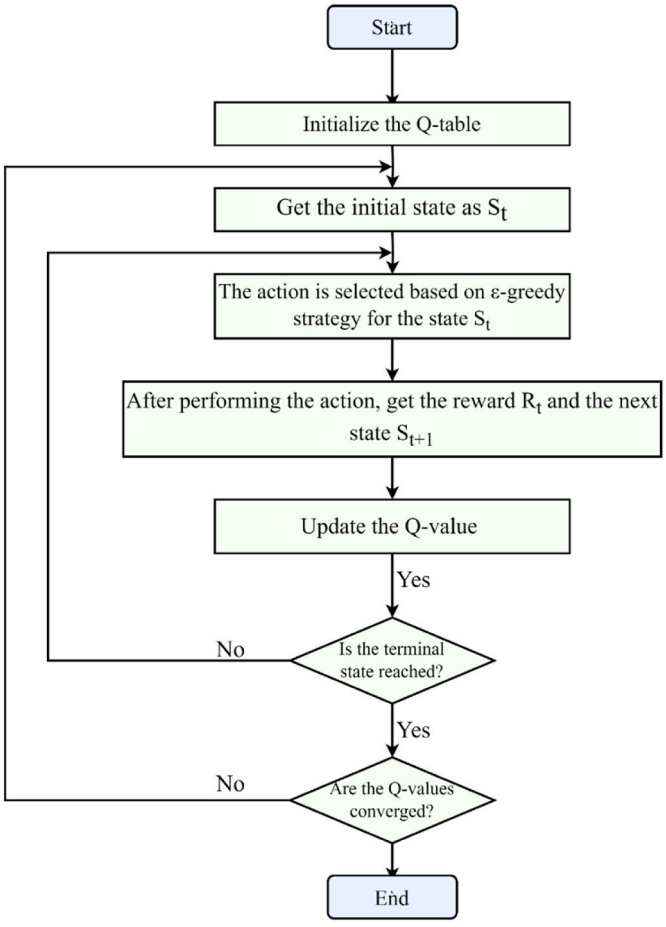


Fig. 5. Proposed Q-learning based resource allocation framework.

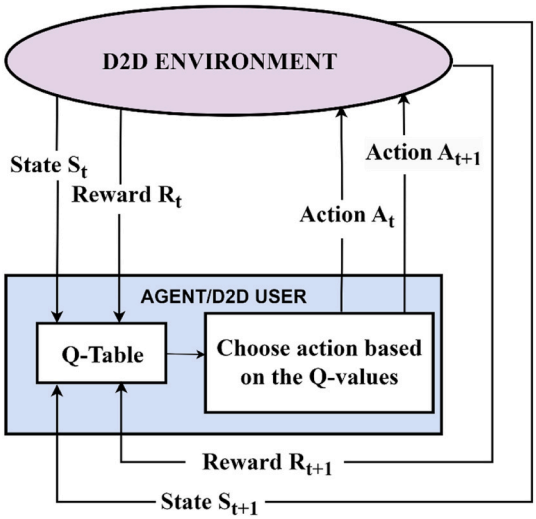


Fig. 6. Agent-environment interactions in Q-learning.

Table 3
Classification of state values.

State value = 0	State value = 1
$\gamma^r_{D_u} < \tau_0$	$\gamma^r_{D_u} \geq \tau_0$
$G_{D_{uk}} < \tau_1$	$G_{D_{uk}} \geq \tau_1$
$G_{eNB,j} < \tau_2$	$G_{eNB,j} \geq \tau_2$

are considered in states. The variables that are used in state representation are $\gamma^r_{D_u}$, $G_{D_{uk}}$ and $G_{eNB,j}$. Here, $\gamma^r_{D_u}$ is the SINR of the agent which in our work is the transmitting device in D2D mode employing the rth RB, $G_{D_{uk}}$ is the channel gain over the link that is engaged in D2D data transmission between the devices j and k and $G_{eNB,j}$ is the channel gain over the link that is engaged in communication between the base station and the device j and I_{d_N} implies the interference level.

The state-defining variables are assigned a threshold value that is set as τ_0, τ_1 and τ_2 , where, τ_0 is the minimum value of SINR while τ_2 and τ_3 are the minimum threshold values of the channel gain. These threshold values are fixed in such a way that the system performs well. The state set values take the value 1 when the three variables are greater than the threshold value and 0 when it is lesser as presented in Table 3.

ACTION: The task that is being carried out by the agent. The action here is the assignment of the resources.

$$A(t) = \{A_1(t), A_2(t)\} \tag{20a}$$

$$A = a^r_p = a^r_1, a^r_2, a^r_3, \ldots, a^r_{pl} \tag{20b}$$

$$A_1(t) = \mathrm{r(t)} \tag{20c}$$

where, at time, t, the action A_t is taken and $r = \{1, 2, \ldots r_{max}\}$.

The actions can be chosen in different ways that can be random or according to some policy or strategy like the ε-greedy strategy.

$$\pi^s_a = arg\ max_{a \in A} Q_i(s_i, a_i) \tag{21}$$

REWARD: The reinforcement learning-based Q-learning algorithm is proposed to improve the throughput of the devices in the system individually such that the overall system throughput is also increased. The reward that the agent receives depends on the actions that the agent performs and can be both positive and negative. Therefore, the better the action that is taken, the greater the reward.

The D2D transmitter d^r_{uT} which here is the agent receives the reward of optimal throughput value as

$$R^r_{d_{uT}} = BW_{RB} \sum_{d=1}^{D} log_2\left(1 + \gamma^r_{d_{uT} d_{uR}}\right) \tag{22a}$$

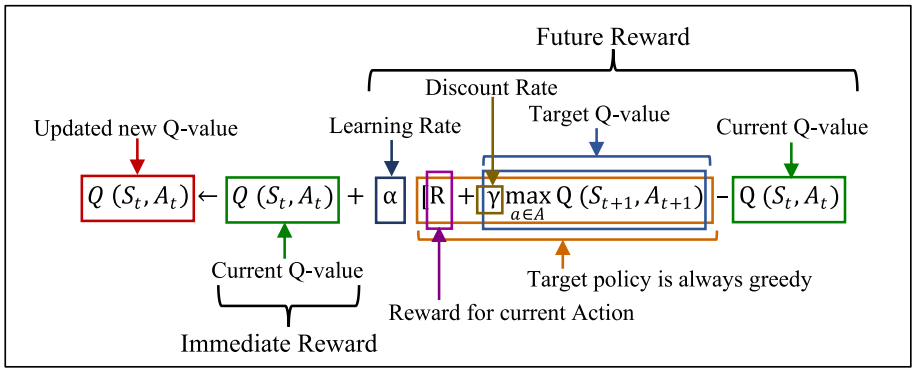


Fig. 7. Q-Learning: updated Q-value and Rewards relation.

$$R^{r}_{d_{uT}} = \begin{cases} log_2\left(1+\gamma^{r}_{d_{uT}d_{uR}}\right) & if\ S_t \epsilon\ S, A_t \epsilon\ A \\ -1 & otherwise \end{cases} \tag{22b}$$

$$Q\ (S_t, A_t) \leftarrow\ Q\ (S_t, A_t) + \alpha\left[\mathrm{R} + \gamma \underset{a\in A}{\max} \mathrm{Q}\ (S_{t+1}, A_{t+1}) - \mathrm{Q}\ (S_t, A_t)\right] \tag{23}$$

The Q-learning equation breaks down as demonstrated in Fig. 7.

Algorithm 1. explains how the Q-values are updated iteratively based on the observed rewards and transitions in the D2D environment. The Q-learning algorithm aims to find the optimal policy by learning the Q-values that maximize the expected cumulative rewards over time. The agent uses these Q-values to make decisions by selecting the action with the highest Q-value for a given state. The input parameters are taken in such a way that the system achieves good performance. The maximum allowable distance that we have considered for D2D communication is 100 m.

System fairness is the fair distribution of resources among the cellular and D2D user devices. System fairness becomes challenging when the number of D2D pairs in the network increases with only available limited resources in the cell. Our proposed Q-learning scheme achieves a good level of fairness among multiple D2D users to achieve the same average throughput.

The fairness index of the proposed Q-learning method is measured using Jain's Fairness index. It is given as

$$f(D_1, D_2, \ldots D_N) = \frac{\left(\sum_{i=1}^{N} D_i\right)^2}{N\left(\sum_{i=1}^{N} {D_i}^2\right)} \tag{24}$$

where, D is the obtained throughput by each D2D device and N is the number of devices.

Latency is a critical performance metric in D2D, particularly for applications in real-time data exchange. Delay-sensitive applications are impacted by latency which is a combination of propagation delay (D_p), transmission delay (D_t), processing delay (D_p) and queuing delay (D_q).

$$Latency, L = \frac{dist}{s} + \frac{l}{R} + \mathrm{D_{p+}}\ \mathrm{D_q} \tag{25}$$

where, s is the speed of transmission, l is the length of the packet in bits and R is the transmission rate of the link. Here, the propagation delay remains constant as the distance between the devices is approximately 100 m, the transmission delay depends on the bandwidth and data rate, which are influenced by the number of devices sharing the resource blocks, processing delay is a fixed value set for the hardware involved in communication and the queuing delay increases linearly with the number of devices.

In larger communication networks with a higher number of D2D devices, scalability becomes a critical concern. Scalability in the context of D2D networks refers to the ability of a resource allocation algorithm to maintain performance levels even as the number of user devices (both cellular and D2D) increases. In other words, it is the ability to handle an increasing network size efficiently while maintaining or improving performance metrics such as throughput, latency, and interference management. Scalability can be measured in terms of the scalability index and is given as

$$\text{Scalability index}, S = \frac{Th_{max}}{Th_{min}} \tag{26}$$

where, Th_{max} is the throughput at the maximum number of devices and Th_{min} is the throughput at the minimum number of devices. Q-learning exhibits strong scalability in D2D-enabled cellular networks and maintains their performance as the number of D2D devices increases in the network due to its fast convergence, high adaptability, and efficient learning making it well-suited for dense environments. Random allocation, while showing some scalability, is not as effective as the learning-based algorithms. Q-learning in this paper exhibits a scalability index of approximately 1.65, indicating that it performs well in maintaining throughput even as the number of D2D devices increases which can in other words be interpreted that Q-learning exhibits a good scalability as the throughput does not abruptly decrease for increasing number of devices.

Table 4
Simulation parameters.

Parameters	Values
Bandwidth, BW	30 MHz
Cell radius	1000 m
D2D radius	100 m
Number of Cellular users	30–100
Number of D2D users	10–120
Number of resource Blocks	35
Pmax	23 dBm
Pathloss parameter	3.5
Єmax	0.3
Єmin	0.1
τ_0	0.004
τ_1	0.2512
τ_2	0.2512
$I^{(r)}_{th}$	0.001
W_{RB}	180 kHz
Initial Q (S,A)	0
Initial e(S,A)	0
ρ	1
k	0.25
γ	0.9
γ_1	0.5
λ	0.5
Learning rate.	0.2

5. Simulation parameter and result analysis

The proposed resource allocation strategy is simulated using MatLab and is numerically analysed in this section. The simulation system parameters are summarized in Table 4.

We consider a single-cell environment of about 1000 m radius, where the central entity, eNB is positioned at the cell centre and the devices are scattered in the cell randomly. The Cellular users, CUs and the D2D users, DUs are assumed to be within the coverage of the central entity. We have considered 10 to 130 number of D2D devices and 35 RBs for allocation.

In this section, the effective performance of the proposed Q-learning technique is evaluated with the efficacy of state-of-the-art methods like SARSA, distributed RA, location-based RA, and random RA approaches to analyse the performance evaluation of the proposed RL-based D2D communication. SARSA and Q-learning are both similar in operation except that they differ in the way they update their action-value functions. Considering the actions being performed in the next state, SARSA modifies its Q-values based on the current and the next state. In contrast, Q-learning updates its Q-values by considering the best optimal action in the next state. Distributed RA is a device-centric method where the devices directly communicate with each other to assign the available resources like channel and power. Location-based D2D RA involves the assignment of resources based on the physical position of the D2D user devices. Random resource allocation in D2D communication is a simple-to-implement scheme that does not have a specific plan or logic. We compare our proposed Q-learning with the above-listed methods to analyse its performance. Implementing Q-learning-based D2D resource allocation for the given system mode is practical and efficient in terms of performance metrics and system parameters.

5.1. Throughput analysis

Figs. 8–10 show the variation in throughput for the D2D users, cellular users, and the overall system, respectively for the varying number of devices. We can infer from the figures that the overall system throughput increases with the increasing number of devices. This increase is higher in our proposed Q-learning method compared to the considered existing state-of-the-art techniques. From Fig. 8, it is evident that the throughput is maximum for a greater number of devices. For 130 devices, 170 Mbps of throughput is achieved for the Q-L method while, 160Mbps, 140Mbps, 135Mbps and 110Mbps are achieved for the other methods as mentioned earlier. Furthermore, when system throughput is considered as shown in Fig. 10, for 130 devices, 200Mbps

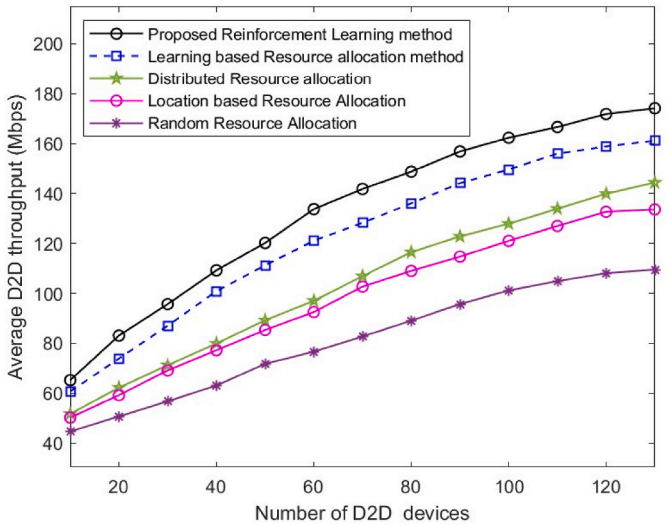


Fig. 8. Number of D2D devices vs the average D2D throughput.

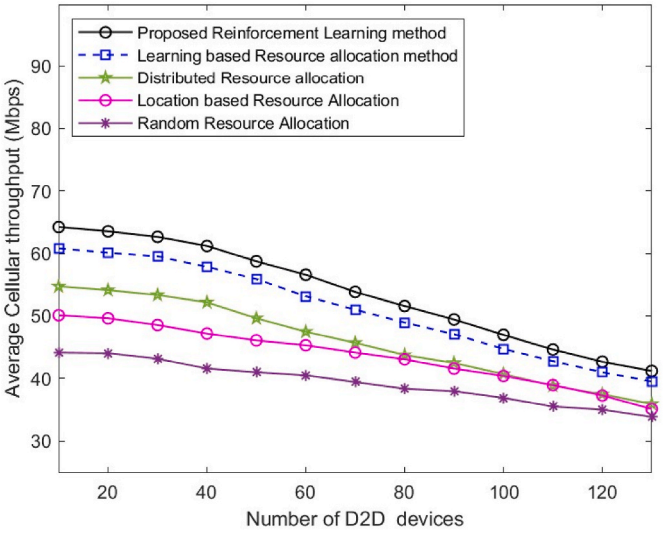


Fig. 9. Number of D2D devices vs the average Cellular throughput.

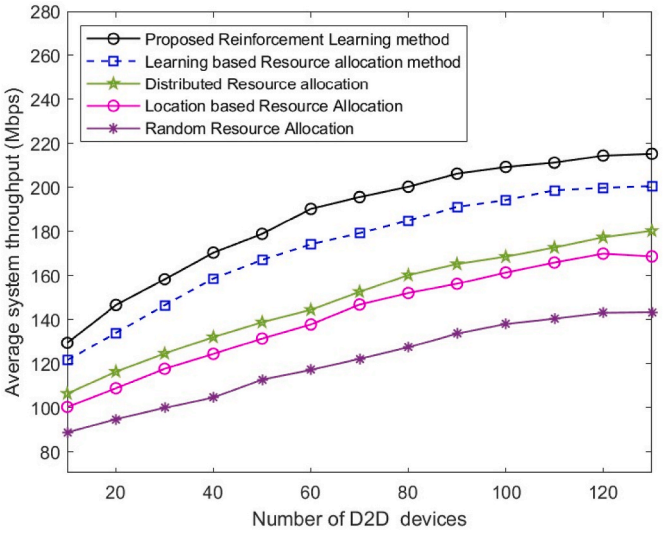


Fig. 10. Number of D2D devices vs the average system throughput.

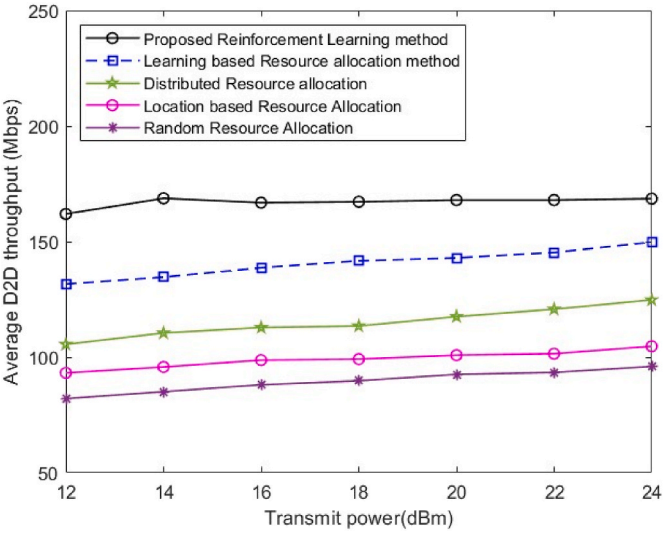


Fig. 11. Transmit power vs the average D2D throughput.

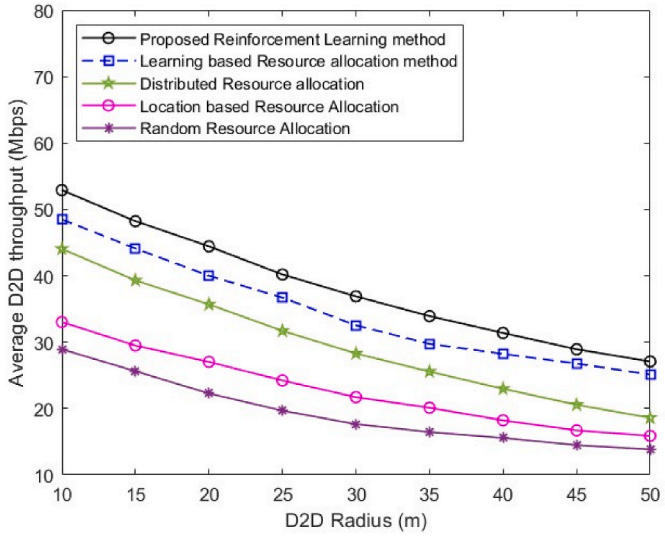


Fig. 12. D2D radius vs Average D2D throughput.

of throughput is exhibited while the other RA methods exhibit 190Mbps, 170Mbps, 160Mbps and 140Mbps respectively.

Fig. 11 presents the transmission power of the devices vs the average D2D throughput. 160Mbps of throughput is displayed by the Q-learning method for a transmit power of 23Dbm while the other methods exhibit a data rate of around 150Mbps, 130Mbps,100Mbps and 80Mbps respectively.

Fig. 12 presents the output of the average D2D throughput for the varying D2D operating radius. The simulation results show that the throughput is maximum for a lesser radius and it starts to decrease as the radius increases. The proposed Q-learning method depicts that its throughput is a maximum of 50 Mbps for a radius of 10 m and is a minimum of around 35 Mbps for a radius of 50 m.

5.2. Simulation analysis in terms of energy and spectral efficiency

According to Fig. 13, the Q-learning method presents an energy efficiency of around 3Mbps/dBm for 10 users and it gradually increases for an increasing number of devices and shows a value of around 7 Mbps/dBm for 130 D2D devices. The proposed method is proven to provide an optimal assignment of power compared to the other

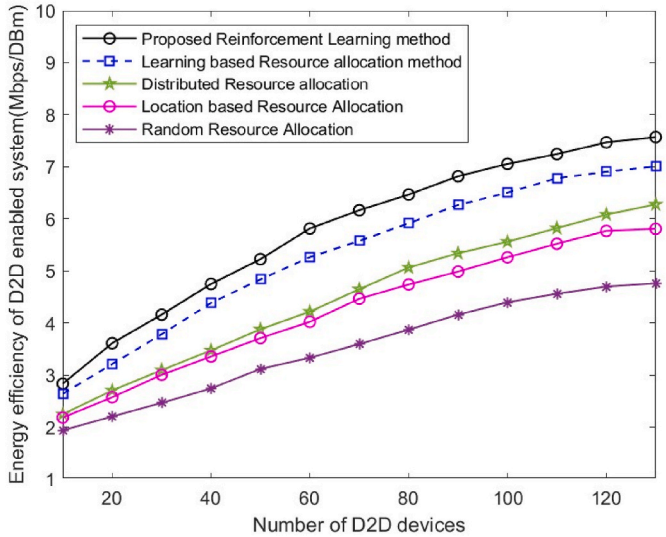


Fig. 13. Number of D2D devices vs Energy efficiency.

considered methods due to its higher energy efficiency value as depicted in Table 5.

According to Fig. 14, the simulation results show that the proposed Q-learning method presents a spectral efficiency of around 25 Mbps for 10 users and it gradually increases for an increasing number of devices and shows a value of around 40 Mbps for 130 D2D users. Same as this the spectral efficiency of the existing methods considered also increases for the increasing number of devices but not more than that of the proposed Q-learning method. Therefore, the proposed Q-learning displays a higher spectral efficiency compared to the considered existing methods like SARSA, distributed, location-based and random RA as depicted in Table 5 which show a spectral efficiency of 23.9 Mbps, 22.4 Mbps,19.7 Mbps,18.5 Mbps and 16.7 Mbps for 10 users and 37.15 Mbps,33.38 Mbps,31.24 Mbps and 26.54 Mbps for 130 users, respectively.

5.3. Analysis of fairness index: D2D-enabled wireless communication system

Fairness is one of the key advantages of D2D communication. The communication quality is consistent between the nearby and far away devices with good signal quality. The system fairness is ensured by the fair assignment of resources like spectrum and power. The fairness of the communication system is assessed by Jain's fairness index. The fairness of the system is greater with a higher fairness index. The fairness value of the proposed Q-learning technique, as shown in Fig. 15 ranges from 0.998 to 0.98. With an increasing number of devices, the system fairness slightly decreases. The fairness plot indicates the fair distribution of network resources among D2D users which is one of the major requirements in resource allocation problems.

Table 5 summarizes the performance of the various resource allocation techniques. A maximum of 208.32 Mbps of date rate is shown by the Q-learning method, while SARSA-based RL is 192.32 Mbps. Thus, the Q-learning RL algorithm gives an output of about 6 %–8 % greater than that of the SARSA method for the considered D2D system model. Jain's fairness index model is used in the determination of the system fairness. The Q-learning strategy shows an output of ~0.999–0.987 while the SARSA method exhibits ~0.998-0.982 which is almost nearer in values. The spectrum and energy efficiency of the Q-learning integrated D2D communication is greater than the state-of-the-art RL techniques like SARSA considered for 10–130 devices. The performances are degraded comparatively in the case of the other existing methods. The average fairness index of the proposed Q-learning method in a D2D environment is 0.937.

5.4. Latency assessment

Fig. 16 shows that the Q-learning technique consistently achieves lower latency compared to the random resource allocation technique and SARSA. The effectiveness of Q-learning is because of its exploration and exploitation balance, leading to a more optimized policy. This highlights the effectiveness of Q-learning in optimizing resource allocation and improving the performance of delay-sensitive applications in a D2D integrated cellular network. Q-learning performs better because it explores and exploits the environment to find the optimal policy that minimizes latency. SARSA on the other hand performs slightly worse than Q-learning because it updates based on the action taken rather than the best possible action.

6. Conclusion and future work

As a significant technology in Beyond 5G wireless networks, D2D offers numerous advantages and plays a pivotal role in various applications across diverse domains. Effective allocation of resources in a cellular communication system assisted by D2D is crucial to fully harness the benefits of D2D communication. Despite significant research efforts in this field, the joint channel and power allocations integrated

Table 5
Simulation result analysis.

S.No	D2D resource allocation techniques	Fairness	System throughput (Mbps)	Spectrum Efficiency- SE (Mbps/Hz)			Energy Efficiency-EE (Mbps/dBm)	
				Number of UEs			Number of UEs	
				10	100	130	10	130
1	Proposed Q learning-based RL method	0.999–0.987	208.32	23.9	38.75	39.87	2.83	7.57
2	SARSA based RL	0.998–0.982	192.32	22.4	35.98	37.15	2.639	7.006
3	Distributed RA	0.99–0.968	168.58	19.7	31.98	33.38	2.18	6.20
4	Location-based RA	0.996–0.96	161.4	18.5	29.88	31.24	1.98	5.808
5	Random RA	0.976–0.85	138.07	16.7	25.55	26.54	1.93	4.76

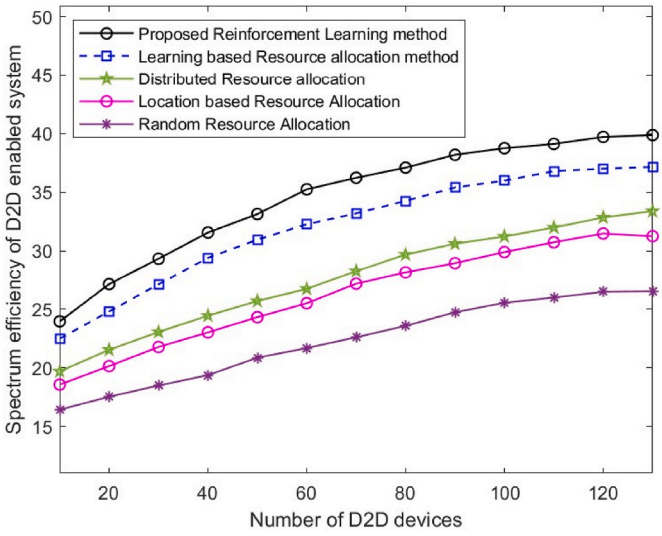


Fig. 14. Number of D2D devices vs Spectral Efficiency.

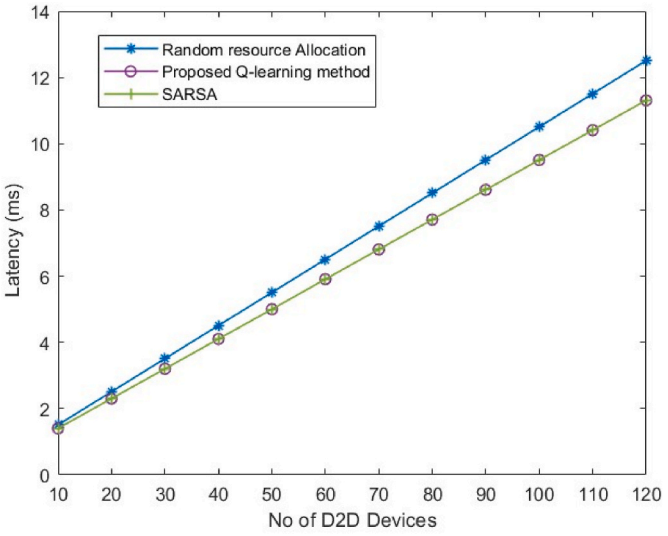


Fig. 16. No. of D2D devices vs Latency.

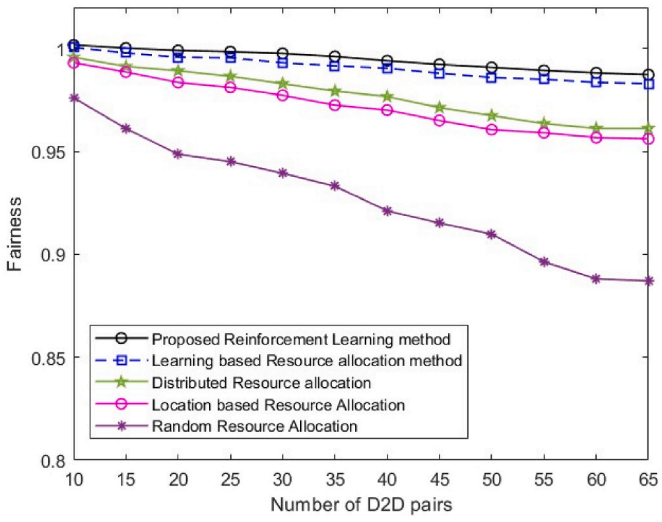


Fig. 15. No. of D2D devices vs Fairness.

with constrained learning techniques are very limited. Therefore, we propose a joint channel and power allocation approach, leveraging the Q-learning technique that addresses the constraint optimization problem. Q-learning is well-suited for exploration and demonstrates a higher potential for discovering more optimal policies compared to state-of-the-art techniques, including SARSA, distributed resource allocation, and other conventional methodologies. Our simulation results highlight the efficiency of the proposed Q-learning-based joint spectrum and power allocation, leading to a substantial throughput improvement in the range of approximately 6 %–8 % when compared to the state-of-the-art SARSA technique. Furthermore, the system exhibits a high level of fairness (0.998), indicating a well-balanced allocation that ensures nearly equal treatment of all user devices, thereby minimizing disparities in resource allocation. In conclusion, our proposed resource allocation based on Q-learning improves the system performance based on the considered performance metrics. The throughput does not drastically fall as the number of devices increases in the network, indicating that Q-learning demonstrates a scalability of 1.69 which is good.

Q-learning can optimize the allocation of resources to enhance security and minimize interference and eavesdropping risks. Q-learning also detects network traffic or behavioral patterns, which might indicate security attacks. By continuously learning, Q-learning can also help in developing an adaptive defense mechanism that ensures security. Based on the working mechanisms, it is evident that the proposed Q-learning algorithm assures better security compared to that of conventional methods. Q-learning, while a powerful RL algorithm, has inherent limitations in ensuring security and privacy by itself. With appropriate enhancements and integrations, it can contribute to more privacy and security, which will be our future research focus. In our future work, we also propose to further explore and develop an AI-driven deep learning solution to optimize resource allocations and to minimize the interference of the system.

CRediT authorship contribution statement

Steffi Jayakumar: Formal analysis, Conceptualization. **S. Nandakumar:** Supervision, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] F. Salahdine, T. Han, N. Zhang, 5G, 6G, and Beyond: recent advances and future challenges, Ann des Telecommun Telecommun (2023), https://doi.org/10.1007/s12243-022-00938-3.

[2] S.P. V, U.G. S, A. D, et al., Revolutionizing connectivity: unleashing the power of 5G wireless networks enhanced by artificial intelligence for a smarter future, Results Eng 22 (2024) 102334, https://doi.org/10.1016/j.rineng.2024.102334.

[3] C.X. Wang, X. You, X. Gao, et al., On the road to 6G: visions, requirements, key technologies, and testbeds, IEEE Commun Surv Tutorials 25 (2023) 905–974, https://doi.org/10.1109/COMST.2023.3249835.

[4] S.P. V, A.J. Albert, K.N.K. Thapa, R. Krishnaprasanna, A novel enhanced security architecture for sixth generation (6G) cellular networks using authentication and acknowledgement (AA) approach, Results Eng 21 (2024) 101669, https://doi.org/10.1016/j.rineng.2023.101669.

[5] M.S.M. Gismalla, A.I. Azmi, Salim MR. Bin, et al., Survey on device to device (D2D) communication for 5GB/6G networks: concept, applications, challenges, and future directions, IEEE Access 10 (2022) 30792–30821, https://doi.org/10.1109/ACCESS.2022.3160215.

[6] Y. Xu, G. Gui, H. Gacanin, F. Adachi, A survey on resource allocation for 5G heterogeneous networks: current research, future trends, and challenges, IEEE Commun Surv Tutorials 23 (2021) 668–695, https://doi.org/10.1109/COMST.2021.3059896.

[7] T. Islam, C. Kwon, Survey on the state-of-the-art in device-to-device communication: a resource allocation perspective, Ad Hoc Netw. 136 (2022) 102978, https://doi.org/10.1016/j.adhoc.2022.102978.

[8] N. Shingari, B. Mago, A framework for application-centric Internet of Things authentication, Results Eng 22 (2024) 102109, https://doi.org/10.1016/j.rineng.2024.102109.

[9] L. Qian, P. Yang, M. Xiao, et al., Distributed learning for wireless communications: methods, applications and challenges, IEEE J Sel Top Signal Process 16 (2022) 326–342, https://doi.org/10.1109/JSTSP.2022.3156756.

[10] S. Jayakumar, S. Nandakumar, Reinforcement learning based distributed resource allocation technique in device-to-device (D2D) communication, Wirel Networks 0123456789 (2023), https://doi.org/10.1007/s11276-023-03230-x.

[11] R. Li, P. Hong, K. Xue, et al., Resource allocation for uplink NOMA-based D2D communication in energy harvesting scenario: a two-stage game approach, IEEE Trans. Wireless Commun. 21 (2022) 976–990, https://doi.org/10.1109/TWC.2021.3100567.

[12] M. Le, Q.V. Pham, H.C. Kim, W.J. Hwang, Enhanced resource allocation in D2D communications with NOMA and unlicensed spectrum, IEEE Syst. J. 16 (2022) 2856–2866, https://doi.org/10.1109/JSYST.2021.3136208.

[13] R.K. Jha, M.K. Pedhadiya, A. Dogra, et al., Joint resource and power allocation for 5G enabled D2D networking with NOMA, Comput. Network. 222 (2023) 109536, https://doi.org/10.1016/j.comnet.2022.109536.

[14] M. Hmila, M. Fernandez-Veiga, M. Rodriguez-Perez, S. Herreria-Alonso, Distributed energy efficient channel allocation in underlay multicast D2D communications, IEEE Trans. Mobile Comput. 21 (2022) 514–529, https://doi.org/10.1109/TMC.2020.3012451.

[15] P. Pawar, A. Trivedi, Joint uplink-downlink resource allocation for D2D underlaying cellular network, IEEE Trans. Commun. 69 (2021) 8352–8362, https://doi.org/10.1109/TCOMM.2021.3116947.

[16] W. Wu, R. Liu, Q. Yang, T.Q.S. Quek, Learning-based robust resource allocation for D2D underlaying cellular network, IEEE Trans. Wireless Commun. 21 (2022) 6731–6745, https://doi.org/10.1109/TWC.2022.3152260.

[17] M. Elnourani, S. Deshmukh, B. Beferull-Lozano, Resource allocation for underlay interfering D2D networks with multiantenna and imperfect CSI, IEEE Trans. Commun. 70 (2022) 6066–6082, https://doi.org/10.1109/TCOMM.2022.3194193.

[18] M.H. Zafar, I. Khan, M.O. Alassafi, An efficient resource optimization scheme for D2D communication, Digit Commun Networks 8 (2022) 1122–1129, https://doi.org/10.1016/j.dcan.2022.03.002.

[19] X. Wang, H. Pan, Y. Shi, Distributed resource allocation for D2D communications underlaying cellular network based on Stackelberg game, EURASIP J. Wirel. Commun. Netw. (2022) 2022, https://doi.org/10.1186/s13638-021-02055-6.

[20] R.M. Rizk-Allah, A.E. Hassanien, A. Marafie, An improved equilibrium optimizer for numerical optimization: a case study on engineering design of the shell and tube heat exchanger, J Eng Res (2024), https://doi.org/10.1016/j.jer.2023.08.019.

[21] R.M. Rizk-Allah, E. Elsodany, An improved rough set strategy-based sine cosine algorithm for engineering optimization problems, Soft Comput. 28 (2024) 1157–1178, https://doi.org/10.1007/s00500-023-09155-z.

[22] R.M. Rizk-Allah, A quantum-based sine cosine algorithm for solving general systems of nonlinear equations, Artif. Intell. Rev. 54 (2021) 3939–3990, https://doi.org/10.1007/s10462-020-09944-0.

[23] R.M. Rizk-Allah, A.E. Hassanien, V. Snášel, A hybrid chameleon swarm algorithm with superiority of feasible solutions for optimal combined heat and power economic dispatch problem, Energy 254 (2022), https://doi.org/10.1016/j.energy.2022.124340.

[24] M. Al-Raeei, Applying fractional quantum mechanics to systems with electrical screening effects, Chaos, Solit. Fractals 150 (2021) 111209, https://doi.org/10.1016/j.chaos.2021.111209.

[25] C. Di, Q. Zhou, J. Shen, et al., The coupling effect between the environment and strategies drives the emergence of group cooperation, Chaos, Solit. Fractals 176 (2023) 114138, https://doi.org/10.1016/j.chaos.2023.114138.

[26] H. Jiang, S. Cao, Reinforcement learning-based active flow control of oscillating cylinder for drag reduction, Phys. Fluids 35 (2023), https://doi.org/10.1063/5.0172081.

[27] A.H. Alquhali, M. Roslee, M.Y. Alias, K.S. Mohamed, D2D communication for spectral efficiency improvement and interference reduction: a survey, Bull Electr Eng Informatics 9 (2020) 1085–1094, https://doi.org/10.11591/eei.v9i3.2171.

[28] S. Jayakumar, A review on resource allocation techniques in D2D communication for 5G and B5G technology, Peer-to-Peer Networking and Applications 14 (1) (2021) 243–269, https://doi.org/10.1007/s12083-020-00962-x.

[29] Y. Liu, X. Yuan, Z. Xiong, et al., Federated learning for 6G communications: challenges, methods, and future directions, China Commun 17 (2020) 105–118, https://doi.org/10.23919/JCC.2020.09.009.

[30] F. Hussain, S.A. Hassan, R. Hussain, E. Hossain, Machine learning for resource management in cellular and IoT networks: potentials, current solutions, and open challenges, IEEE Commun Surv Tutorials 22 (2020) 1251–1275, https://doi.org/10.1109/COMST.2020.2964534.

[31] A. Omidkar, A. Khalili, H.H. Nguyen, H. Shafiei, Reinforcement-learning-based resource allocation for energy-harvesting-aided D2D communications in IoT networks, IEEE Internet Things J. 9 (2022) 16521–16531, https://doi.org/10.1109/JIOT.2022.3151001.

[32] S.H. Lee, X.P. Shi, T.H. Tan, et al., Performance of Q-learning based resource allocation for D2D communications in heterogeneous networks, ICT Express (2023), https://doi.org/10.1016/j.icte.2023.02.003.