future internet

MDPI

Article

Machine Learning-Based Resource Allocation Algorithm to Mitigate Interference in D2D-Enabled Cellular Networks

Md Kamruzzaman, Nurul I. Sarkar * and Jairo Gutierrez

Computer Science and Software Engineering, Auckland University of Technology, Auckland 1010, New Zealand; zaman.kamruzzaman@aut.ac.nz (M.K.); jairo.gutierrez@aut.ac.nz (J.G.)
* Correspondence: nurul.sarkar@aut.ac.nz

Abstract: Mobile communications have experienced exponential growth both in connectivity and multimedia traffic in recent years. To support this tremendous growth, device-to-device (D2D) communications play a significant role in 5G and beyond 5G networks. However, enabling D2D communications in an underlay, heterogeneous cellular network poses two major challenges. First, interference management between D2D and cellular users directly affects a system's performance. Second, achieving an acceptable level of link quality for both D2D and cellular networks is necessary. An optimum resource allocation is required to mitigate the interference and improve a system's performance. In this paper, we provide a solution to interference management with an acceptable quality of services (QoS). To this end, we propose a machine learning-based resource allocation method to maximize throughput and achieve minimum QoS requirements for all active D2D pairs and cellular users. We first solve a resource optimization problem by allocating spectrum resources and controlling power transmission on demand. As resource optimization is an integer nonlinear programming problem, we address this problem by proposing a deep Q-network-based reinforcement learning algorithm (DRL) to optimize the resource allocation issue. The proposed DRL algorithm is trained with a decision-making policy to obtain the best solution in terms of spectrum efficiency, computational time, and throughput. The system performance is validated by simulation. The results show that the proposed method outperforms the existing ones.

Keywords: machine learning (ML); device-to-device (D2D); resource allocation (RA); reinforcement learning (RL); Markov Decision Process (MDP); deep reinforcement learning (DRL); ultra-dense network (UDN); heterogeneous networks (HetNets)

Citation: Kamruzzaman, M.; Sarkar, N.I.; Gutierrez, J. Machine Learning-Based Resource Allocation Algorithm to Mitigate Interference in D2D-Enabled Cellular Networks. *Future Internet* **2024**, *16*, 408. https://doi.org/10.3390/fi16110408

Academic Editor: Sachin Sharma

Received: 22 September 2024
Revised: 20 October 2024
Accepted: 30 October 2024
Published: 6 November 2024

1. Introduction

Wireless communication technologies have come a long way from the first generation (voice-only) to 5G and beyond 5G (B5G) [1]. The data transmission speed in each generation has increased significantly. Through revolutionary changes in mobile networks, 5G and B5G can achieve the best performance for throughput, latency, coverage, energy consumption, and spectral efficiency.

Due to various innovative applications, the volume of mobile and broadband traffic, and end-users' demand for faster data access, cellular networks have experienced unprecedented growth in the last decade [2].

Globally, the total number of cellular subscribers has grown from 7.5 billion in 2017 to 8.4 billion in 2022 and is projected to be 9.2 billion by 2028. According to a recent Ericsson mobility report and Cisco visual networking index [3,4], global mobile data and connected mobile devices to the networks increased significantly in 2017. Similarly, mobile data traffic increased from 686 petabytes to 11.6 exabytes per month from 2012 to 2017, and it reached 90 EB/month in 2022. This trend will continue, and it is expected that there will be a further 4-fold increment between 2022 and 2028, reaching 90 exabytes per month to 325 EB/month by 2028 [4]. The number of mobile devices is drastically increasing, with a demand for

higher data rates and faster data-accessing applications in recent years. Cutting-edge applications such as 3D holography, artificial intelligence (AI), machine-to-machine (M2M) communication, Internet-of-Things (IoT), virtual reality (VR), e-learning, video-based applications, augmented reality (AR), and ultra-broadband demand more bandwidth than can be fulfilled with the existing 4G or even 5G networks. Only unconventional thinking in 5G and B5G cellular networks can achieve this exponential demand for higher data rates and capacity requirements [5]. Spectral efficiency and bandwidth determine wireless communication's capacity and data rates. The combined requirement of hyper-connected devices and new applications will trigger a major evolution in cellular networks, which will improve data rates, latency, energy consumption, and coverage. According to ETSI [6], the target peak data rates for downlink and uplink in 5G are 20 Gbps and 10 Gbps, respectively. The user-centric KPI for 4G, 5G, and 6G is shown in Table 1.

Table 1. Performance comparison of 4G, 5G, and 6G mobile technologies.

KPI	4G	5G	6G (Possible)
User data rate	150 Mbps	1 Gbps	100 Gbps
DL data rate	300 Mbps	20 Gbps	>1 Tbps
Latency for U-plane	60 ms	0.5 ms	0.1 ms
Latency for C-plane	200 ms	10 ms	<1 ms
Spectral efficiency	6 bps/Hz	30 bps/Hz	100 bps/Hz
Frequency	600 MHz–2.5 GHz	3–300 GHz	upto 1 THz

To overcome the limitations of incredibly low latency and extremely high bandwidth requirements, 5G is designed with various promising technologies like D2D communications, millimeter-wave (mm), ultra-dense networks, massive MIMO, and cognitive radio (CR) networks [1]. As mentioned in Ericsson's recent mobility report and Cisco VNI [3,4], future wireless traffic explosion is unavoidable, and to manage the wireless traffic explosion, large numbers of small cell (SCeNBs), especially femtocells (HeNBs), deployments are inevitable, which will not only increase the network capacity but also eliminate the coverage holes [1]. In addition, massive machine type communications (mMTC), ultra-reliable and low latency communication (uRLLC), and enhanced mobile broadband (eMBB) services will be more dominant in 5G and B5G networks. So, cellular network operators will face a challenge in managing the higher user density and resulting immense data volumes [1].

To fulfill the ever-increasing data demands, D2D technology is prominent not only in 5G but also in 6G. By adopting D2D communications, cellular operators can have several advantages in spectrum efficiency, latency, throughput, and energy conservation through sharing radio resources. In addition, public safety, proximity-based services (ProSec), group communication system enablers (GCSE), traffic offloading, and content sharing will be enabled by D2D communication [7].

However, to unlock the full potential of D2D technology, we need to address research challenges such as mode selection, neighbor discovery, mobility management, network security, interference and radio resource management, energy consumption, and the coexistence of small cells with D2D communications [8]. Among various challenges, interference management in cellular HetNets is one of the key factors influencing the system performance [7].

Overlay and underlay are the two main categories of D2D communication in terms of spectrum usage. In overlay D2D communication, dedicated radio resources are used for D2D communication, while in underlay D2D communication, the same radio resources are shared among conventional cellular users and D2D communications. Therefore, better spectrum efficiency can be achieved in underlay D2D communication at the cost of complex interference scenarios [1].

In Figure 1 D2D pairs, eNB and SCeNB share cellular resources to communicate with each other, introducing mutual interference among different tiers. Therefore, it is essential to manage interference to unlock the benefits of D2D communication in cellular networks. And resource allocation is one of the key factors for achieving this.

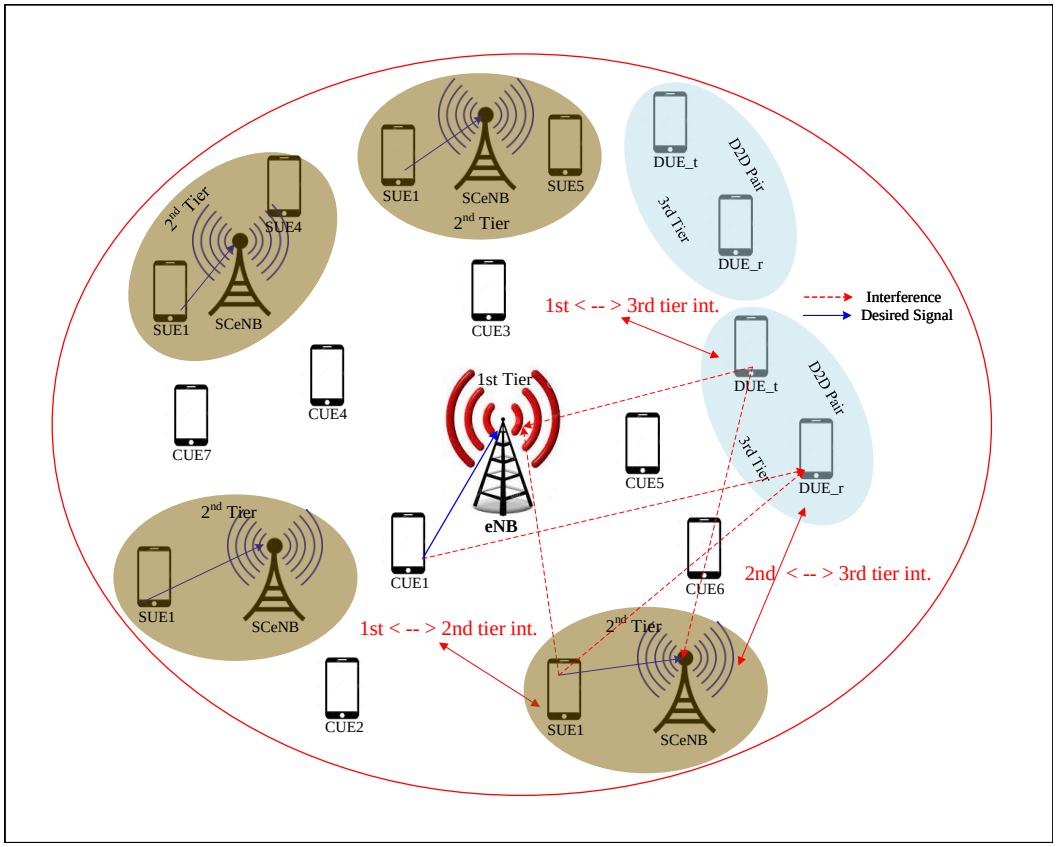

Figure 1. Interference scenario in a D2D-enabled three-tier cellular network.

In such an underlay D2D-enabled cellular HetNet, the cellular system architecture will be moved from tier-2 to tier-3; this involves technologies such as eNBs, SCeNB, and Femtocells, in addition to D2D layers, as shown in Figure 1. In that situation, the total interference between different tiers can be categorized into the following three scenarios:

1. The first tier and the second tier;
2. The first tier and the third tier;
3. The second tier and the third tier.

The high mobility of cellular and spectrum sharing among DUEs between cellular users and DUEs will cause severe interference in the networks and consequently affect the performance of D2D communications seriously. Therefore, for all the issues mentioned for the D2D-enabled cellular HetNets, interference management is the most important.

In any wireless communication, the resultant signal of a receiver device is always the combination of the desired signal along with unwanted interference. To cancel this interference, an interference model is created, and the estimated interference is subtracted from the resultant signal to derive the desired signal. However, using proper resource allocation interference can be reduced as well as avoided to a significant degree. In addition to interference mitigation, resource allocation also helps improve the data rate, throughput, and system sum rate of the wireless communication system.

The sum-rate maximization, latency minimization, and computation energy minimization are the different optimization problem approaches for resource allocation in wireless communications. In conventional methods, most of the formulated optimization problems for resource allocation are non-convex and non-deterministic polynomial time

(NP-hard). Moreover, to obtain optimal solutions with an increasing number of users, the time complexity becomes exponentially higher. Therefore, with traditional optimization methods, it is difficult or impossible to optimize effectively the resource allocation problem, which is modeled as a combination of an optimization problem with nonlinear constraints. Fortunately, to solve decision-making problems under uncertainty, deep reinforcement learning (DRL) is highly effective [9].

In this paper, we optimize resource assignment, mode selection, and transmit power control to maximize the throughput and to ensure minimum QoS requirements for all users in D2D-enabled cellular HetNets. The main contributions of the paper are highlighted below.

- A deep Q-network-based deep reinforcement learning (DRL) algorithm is proposed to minimize interference for achieving maximum throughput with guaranteed quality of services (QoS) for all active users on the network.
- A mathematical model is developed to optimize resource assignment, mode selection, and power control in a D2D-enabled HetNet. The proposed DRL algorithm is analyzed and compared with the existing methods. To validate the system's performance, a simulation approach is used.
- We establish the relationship between the agent's reward and system objectives in a D2D-enabled HetNet for achieving optimum benefit in a Q-network-based DRL environment.

The structure of this paper is organized as follows. A literature review for interference management in D2D-enabled cellular HetNets using both conventional and machine learning approaches is shown in Section 2. System models, including the communication mode, transmission rate, and problem formation, are presented in Section 3. The proposed ML-based resource allocation model is discussed using RL and DL-based algorithms in Section 4. The analytical model, along with resource allocation algorithms, is also presented in this section. The system performance is analyzed and evaluated in Section 5, and the paper is concluded in Section 6. Table 2 illustrates the key notations and abbreviations used in this paper.

Table 2. Key notations and abbreviations.

Notation	Definition
Φ	PPP for eNB
Φ_s	PPP for SCeNB
Φ_d	PPP for DUEs
λ_s	SCeNBs density
λ_d	DUEs density
P_c	Transmit power for CUEs
P_d	Transmit power for DUEs
P_s	Transmission power for SUEs
N_C, N_D, N_S	Set of CUEs, DUEs and SUEs
α	Path loss exponent
γ_{th}	SINR requirements for cellular and D2D links
N_0	Thermal noise
$h_{x,y}$	Channel co-efficient between nodes x and y
ψ	Binary variable

Table 2. *Cont.*

Notation	Definition
η	Discount factor for RL
μ	Learning factor rate for RL
τ	Learning factor for DRL
π	Strategy policy
ρ	Positive constant
S	Set of network environment states
A	Set of actions
SINR	Signal-to-noise-plus-interference ratio
KPI	Key Performance Indicator
PPP	Poison point process
eNB	Evolved Node B
SCeNB	Small Cell evolved Node B
QoS	Quality of Service
UDN	Ultra-Dense Network (UDN)
MIMI	Multiple-Input Multiple-Output
DRL	Deep Reinforcement Learning
DQN	Deep Q-Network
DDQN	Double Deep Q-Network
D3QN	Double-dueling deep Q-Network
DNN	Deep Neural Network
CSI	Channel Stahe Information

2. Related Work

The state-of-the-art researchers propose different methods to manage interference in D2D-enabled heterogeneous cellular networks. Depending on interference behavior, interference management is mainly categorized as interference avoidance, interference cancellation, and interference reduction. Interference avoidance techniques are used in [10] where DUEs coverage area is defined based on pre-defined SINR and the specific transmission power of UEs, and the resource allocation problem is solved using graph coloring techniques.

To improve spectral efficiency, system throughput, and link outage probability, in [11] authors proposed an interference-aware power allocation (IAPA) solution. To minimize interference for a 3-tier network, in [12] the authors initially derived the outage probabilities of all links and then defined the lower distance boundaries from DUEs to base stations and vice-versa. Finally, the authors proposed an optimal small cell deployment density algorithm to ensure the minimum QoS for both D2D and cellular communications [1].

Using DRL methods in [13], the authors proposed a knowledge plane (KP)-based management and orchestration system for network slicing (NS) in 5G and beyond. Finally, the authors compared the outcomes of the proposed solutions in terms of energy consumption, CPU utilization, and time efficiency. Reinforcement learning is becoming popular for solving resource allocation problems in 5G networks. In [14], the authors proposed a Q-learning-based resource allocation technique for solving power allocation and spectrum problems for a single-tier D2D network. Here, Q-learning is used to assign the proper channel and appropriate transmit power for a D2D user. In [15], Kamran et al. propose a

new solution for their earlier resource allocation problem in a D2D multi-tier HetNet by using an ML algorithm. To maximize system throughput by minimizing the interference of cellular users, an ML-based distributed resource allocation scheme is proposed where D2D users interact with the environment and autonomously select the proper cellular resources. Whereas in [16], the authors propose a resource allocation scheme based on Q-learning where D2D users and small cell users select resource block (RB) cooperatively to meet their QoS requirements.

To maximize the overall system throughput for a 2-tier network, the authors formulate a mixed integer non-linear programming resource optimization problem in [17]. Initially, they assign resources to the cellular users using an orthogonal technique, and for D2D users, they adopted a low complexity-based ML algorithm where D2D pairs reuse different sub-carriers to further enhance the system throughput without affecting the CUE's performance. A coalition game-based resource allocation for D2D communication in underlying cellular HetNets with mm-wave communication is proposed in [18]. Here, the authors first formulate a resource allocation scheme for D2D pairs using both mm-wave and cellular spectrum and, based on their characteristics, propose a coalition game theory to maximize the overall sum rate. The joint resource allocation and power control scheme for D2D communication in cellular HetNets is proposed in [19]. For achieving an optimum solution with higher overall system throughput and lower computational complexity, the mixed-integer non-linear optimization problem is subdivided into smaller pieces. In [20], the authors proposed ML and distributed Q-learning decision tree PC algorithms to manage interference in D2D-enabled cellular networks and achieved better throughput and energy efficiency with lower computational complexity. In [21], the authors proposed a hybrid resource allocation scheme where deep neural network (DNN) and heuristic-based PC and resource allocation approaches are used for D2D communication in D2D-enabled underlay cellular networks for two different QoS requirements.

To achieve optimal resource allocation, a DRL-based distributed resource matching scheme is proposed in [22]. A random non-cooperative game approach is used by the authors to formulate the resource allocation problem where each player (D2D pairs) is a learning agent, and the best strategy is learned by interacting with the environment by using a double Deep Q-network (DDQN) [22]. By using a DRL-based double-dueling deep Q-network (D3QN) algorithm, the authors investigate efficient resource allocation and joint user association for D2D communication in ultra-dense cellular networks in [23]. By simulation, they prove that their proposed scheme can achieve better sum throughput even at higher distances between D2D pairs. In [24], the authors proposed a DRL-based double deep Q-network (DDQN) algorithm that enables D2D pairs to learn an optimal spectrum access strategy autonomously for maximizing the sum throughput without complete prior network information. To achieve faster convergence and smart power control for efficient resource allocation for D2D-enabled cellular networks, the authors proposed a Stackelberg game theory-based multi-agent DRL algorithm in [25]. In this proposal, Stackelberg game theory is used to guide the learning direction efficiently for achieving a faster optimal strategy to maximize sum throughput. In [26], a DL-based algorithm is proposed to solve the resource allocation problem where a deep neural network (DNN) model is used to select an optimal policy randomly. A multi-agent DQN-based DRL algorithm is proposed to approximate the loss function for optimal policy, and simulation results are used to prove the higher system throughput in [26]. To improve spectrum efficiency and maximize system capacity, a DQN-based resource allocation and power control algorithm for a 2-tier D2D-enabled cellular network is proposed in [27], where minimum QoS requirements are ensured.

In [28], the authors formulate a non-convex and NP-hard joint power control and resource allocation optimization problem where multiple D2D users share the same cellular resource. To solve this problem, authors initially use fractional programming (FP) techniques. But this technique requires instantaneous CSI which is not feasible for a large and dynamic cellular network. Therefore, they propose a distributed DRL-based scheme

where an agent (D2D user) interacts with the environment repeatedly to select an optimal strategy for maximizing system throughput and spectrum efficiency. Finally, the authors prove how their DRL-based scheme is better compared to the FP technique even without instantaneous CSI. ML/DRL-based resource allocation works are summarized in Table 3.

Table 3. Summary of related work on ML-based resource allocation.

Ref-Year	Main Contribution	Method	Limitation
[29], 2024	Proposed multi-agent DRL-based joint RA and transmit power scheme to maximize energy efficiency.	Deep Q-network model with a graph attention network (GAT).	Unlicensed and optimally assigned licensed spectrum are used.
[30], 2023	Proposed a multi-agent RL for resource and power allocation to maximize throughput with satisfied QoS.	Multi-agent RL method.	Limited to a single cell and few D2D pairs.
[31], 2023	Proposed an energy-efficient DRL-based RA scheme.	Multi-agent DRL framework.	Limited to macro and D2D tiers for energy efficiency.
[32], 2023	Improving network performance by efficient resource allocation.	Co-operative distributed RA algorithm.	Limited to co-channel interference.
[15], 2022	RL-based resource allocation scheme for multi-tier cellular HetNets.	RL-based approach	Not suitable for large network; high computational time.
[26], 2022	Resource allocation in multi-channel cellular systems using DL framework.	Deep neural network (DNN) approach	Depends on CSI availability
[21], 2022	Resource allocation using Deep neural network (DNN) in underlay cellular networks	DNN and heuristic approaches	QoS constraints are violated.
[27], 2021	Joint resource allocation and power control to maximize system performance.	DRL-based ML algorithm	Limited scenarios considered.
[22], 2021	Double Deep Q-Network (DDQN)-based resource allocation sum rate maximization and energy efficiency.	DDQN approach	Focus on one-time slot only.
[23], 2021	Joint user association and resource allocation investigation for D2D-enabled ultra-dense cellular networks.	Double-dueling-deep Q-network	Central controller in the network considered as an agent.
[24], 2021	DRL-based spectrum access scheme for D2D communication in underlay networks.	DDQN approach	Limited scenarios for resource re-usages.
Our work: The proposed DRL method provides improved system performance over heterogeneous cellular networks by optimizing resource allocation and adjusting transmit power dynamically for all users on the network.			

Research Gap: Table 3 reveals that DQN-based joint resource allocation and power control have not been explored yet in a D2D-enabled HetNet considering mutual interference among macro cells, micro cells, and a large number of users simultaneously. In this paper, we formulate a DQN-based joint resource allocation and power control algorithm for D2D-enabled cellular HetNets, which consider mutual interference with all 3-tiers simultaneously to minimize interference with higher user density.

As discussed, most of the state-of-the-art papers are focused on interference management for single or 2-tier cellular networks only. In most cases, researchers consider only one D2D pair, and the simultaneous impact of eNB and multiple small cells is neglected. But in ultra-dense networks (UDNs), an enormous number of small cells will be deployed, and multiple CUEs will be associated with different cellular layers. Hence, interference management will be most challenging with added D2D layers [1]. Despite having very promising prospects for D2D communications, very few research addressed mutual interference problems in mode selection and resource allocation in 3-tier cellular networks considering all tiers simultaneously. In a 3-tiers D2D-enabled cellular HetNet, the most and worst challenging interference scenario is when all 3-tiers (eNB, SCeNB, and D2D layers) will interfere with each other mutually, and none of the authors consider it [1]. Considering this, our proposal for managing interference in a 3-tier cellular HetNet demonstrates the main difference with existing ones.

3. System Model and Problem Formulation

3.1. System Model

For our analysis, we have considered a 3-tier D2D-enabled cellular HetNet where a macro cell (eNB) is placed at the center of the coverage area, and it is surrounded by multiple small cells (SCeNBs) and DUEs. The radius of eNB is 500 m, and small cells, as well as DUEs, are distributed randomly. As small cell locations are unpredictable and random, we have modeled the small cell spatial locations of small cells by a homogeneous poison point process PPP, ϕ_s with density, λ_s. Similar to small cells, DUEs are also distributed by using another independent homogeneous PPP, ϕ_d, with density λ_d. Here, $i \in \{1, 2, \ldots, N_C\}$, $j \in \{1, 2, \ldots, N_D\}$, and $k \in \{1, 2, \ldots, N_S\}$ represent the set of CUEs, DUEs, and SUEs respectively. Figure 1 illustrates a 3-tier heterogeneous cellular network where DUEs, SUEs, and CUEs are communicated simultaneously by sharing the same radio resources. By using the D2D communication mode, DUEs can communicate with each other directly, and CUEs and SUEs can communicate by using cellular and small-cell communication modes. In this system model, the interference scenarios can be defined as follows: (1) eNB receives signals from the transmitter of DUE and SUE; (2) DUE receiver receives signals from the transmitter of CUE and SUE; (3) DUE receiver receives signals from the transmitters of other D2D pairs; (4) SCeNB receives signals from the transmitter of DUE and CUE.

We further assume that there are F orthogonal channels in the network, and all channels are used by CUEs. In other words, all the cellular resources are used by CUEs, and to enable communication, DUEs must reuse the CUE's channels. To mitigate the intra-cell interference, each cellular user uses separate RBs, and cellular resources are assigned orthogonally. In cellular communication mode, only one cellular resource will be shared with a D2D user at a time to limit the co-channel interference. For ease of analysis, here we have considered only the Rayleigh fading environment and exponentially distributed channel coefficients. In such a D2D-enabled HetNet, the received signal can be defined as follows:

$$P_r = P_t h_{xy} D^{-\alpha} \tag{1}$$

where P_t, α, D and h_{xy} are transmission power, path loss exponent, the distance between link nodes, and channel coefficient, respectively, for that link.

With the above assumptions, the SINR, γ, at the receiver, y, can be expressed as follows:

$$\gamma_y = \frac{P_t h_{xy} d_{xy}^{-\alpha}}{I + N_0} = \frac{Received\ Power}{Total\ Interference} \tag{2}$$

Here, I and N_0 are the total interference experienced by the receiver, y, and the spectral noise density, respectively.

To decode a message successfully by a receiver, the SINR at the receiver has to be higher than the threshold SINR γ_{th}, otherwise this link will experience an outage. So the outage probability of any x, y link can be defined as follows:

$$P_{out} = P_r\{\gamma_y \leq \gamma_{th}\} \tag{3}$$

Let us assume that uplink cellular radio resources are shared among DUEs and small cells. Therefore, we can express the mutual interference for the above-defined network as follows:

$$I_i = \sum_{j=1}^{N_D} \zeta_{j,f} P_j h_{j,e} d_{j,e}^{-\alpha} + \sum_{k=1}^{N_S} \zeta_{k,f} P_k h_{k,e} d_{k,e}^{-\alpha} + N_0 \tag{4}$$

$$I_j = P_i h_{i,r} d_{i,r}^{-\alpha} + \sum_{j'=1, j' \neq j}^{N_D} \zeta_{j',f} P_{j'} h_{j',r} d_{j',r}^{-\alpha} + \sum_{k=1}^{N_S} \zeta_{k,f} P_k h_{k,r} d_{k,r}^{-\alpha} + N_0 \tag{5}$$

$$I_k = P_i h_{i,s} d_{i,s}^{-\alpha} + \sum_{j=1}^{N_D} \zeta_{j,f} P_j h_{j,s} d_{j,s}^{-\alpha} + \sum_{k'=1, k' \neq k}^{N_S} \zeta_{k',f} P_{k'} h_{k',s} d_{k',s}^{-\alpha} + N_0 \tag{6}$$

Here, I_i, I_j, and I_k are the total interference received by eNB, all j_{th} D2D pairs except the j_{th} transmitter, and all SUEs except the k_{th} with regard to the SCeNB, respectively. Binary variable $\psi_j^f \in \{0,1\}$ indicates whether D2D users share the sub-channel f with cellular users. If a D2D user reuses the spectrum resource of i_{th} CUE, then $\psi_j^f = 1$; otherwise, $\psi_j^f = 0$. That is,

$$\psi_j^f = \begin{cases} 1, & \text{if sub-channel } f \text{ is shared } i_{th} \text{ CUE,} \\ 0, & \text{otherwise.} \end{cases}$$

Similarly,

$$\psi_k^f = \begin{cases} 1, & \text{if sub-channel } f \text{ is shared } i_{th} \text{ CUE,} \\ 0, & \text{otherwise.} \end{cases}$$

where $\psi_k^f \in \{0,1\}$ shows whether small cell users share the sub-channel f with cellular users or not. Here, the impact analysis of various factors on interference in a D2D-enabled cellular network is limited to performance matrices of small cell and D2D pairs density, SINR, and distances.

3.2. Cellular and D2D Communication Mode

In our defined system model, communications can happen in any mentioned scenarios (eNB, SCeNB, and D2D links). According to Equations (2) and (3), the SINR in uplink for the j_{th} D2D receiver and outage probability of the j_{th} D2D link can be expressed by Equations (7) and (8) respectively [1]:

$$\begin{aligned} \gamma_j^D &= \frac{P_j h_{t,r} d_{t,r}^{-\alpha}}{I_j} \\ &= \frac{P_j h_{t,r} d_{t,r}^{-\alpha}}{P_i h_{i,r} d_{i,r}^{-\alpha} + \sum_{j'=1, j' \neq j}^{N_D} \zeta_{j',f} P_{j'} h_{j',r} d_{j',r}^{-\alpha} + \sum_{k=1}^{N_S} \zeta_{k,f} P_k h_{k,r} d_{k,r}^{-\alpha} + N_0} \end{aligned} \tag{7}$$

$$P_{out,j}^D = 1 - \delta_D \exp(-\beta_D ({P_{j'}}^m \lambda_d + {P_k}^m \lambda_s)) \tag{8}$$

where for effective D2D communication, the required SINR threshold is γ_{th}. Moreover, $P_{out,j}^D$ is the outage probability of the j_{th} D2D link, $\delta_D = exp\left(\frac{-N_0 \gamma_{th} d_{t,r}^{\alpha})}{P_j}\right)\left(\frac{P_j d_{i,r}^{\alpha}}{P_i \gamma_{th} d_{t,r}^{\alpha} + P_j d_{i,r}^{\alpha}}\right)$, $\kappa = \pi m \Gamma(m)\Gamma(1-m)$, $\beta_D = \kappa {\gamma_{th}}^m d_{t,r}^2 / {P_j}^m$. λ_d, and λ_s represent the density of DUEs and small cells (SCeNBs). Appendix A in [33] is referred to as a proof for the above equation.

For the D2D communication mode, Equation (8) infers that small cell and DUEs density, transmit power, path loss coefficient, required SINR, distances between UEs, etc., have a significant impact on the outage probability of D2D links. It is also visible that outage probability increases with higher SINR requirements but decreases while SUEs and D2D receivers are closer to the respective transmitters.

Similarly, for the macro cellular communication mode, the SINR at the receiver and outage probability of the macro cell link can be expressed as [1]:

$$\begin{aligned} \gamma_i^M &= \frac{P_i h_{i,e} d_{i,e}^{-\alpha}}{I_i} \\ &= \frac{P_i h_{i,e} d_{i,e}^{-\alpha}}{P_i h_{i,r} d_{i,r}^{-\alpha} + \sum_{j'=1, j' \neq j}^{N_D} \zeta_{j',f} P_{j'} h_{j',r} d_{j',r}^{-\alpha} + \sum_{k=1}^{N_S} \zeta_{k,f} P_k h_{k,r} d_{k,r}^{-\alpha} + N_0} \end{aligned} \tag{9}$$

$$P_{out,i}^{M} = 1 - \delta_M \exp(-\beta_M (P_j{}^m \lambda_d + P_k{}^m \lambda_s)) \tag{10}$$

where γ_{th} is the required SINR for i_{th} CUE for effective cellular communication, $\delta_M = \exp(-\frac{N_0 \gamma_{th}^m d_{i,e}^{\alpha}}{P_i})$, $\kappa = \pi m \Gamma(m)\Gamma(1-m)$ and $\beta_M = \frac{\kappa \gamma_{th}{}^m d_{i,e}^2}{P_i{}^m}$. So, from the above expressions, it is clear that the outage probability decreases with higher distances between UEs and SUEs as well as DUEs.

For small cell mode communications, the link between SUE and SCeNB will suffer from the interference of other SCeNBs, D2D pairs, and eNB. Therefore, the required SINR and corresponding link outage probability of a small cell can be expressed as follows [1]:

$$\begin{aligned} \gamma_k{}^S &= \frac{(P_k h_{k,s} d_{k,k}^{-\alpha})}{I_k} \\ &= \frac{(P_k h_{k,s} d_{k,k}^{-\alpha})}{P_i h_{i,s} d_{i,s}^{-\alpha} + \sum_{j=1}^{N_D} \zeta_{j,f} P_j h_{j,s} d_{j,s}^{-\alpha} + \sum_{k'=1,k' \neq k}^{N_S} \zeta_{k',f} P_{k'} h_{k',s} d_{k',s}^{-\alpha} + N_0} \end{aligned} \tag{11}$$

$$P_{out,k}^{S} = 1 - \delta_S \exp(-\beta_S (P_j{}^m \lambda_d + P_{k'}{}^m \lambda_s)) \tag{12}$$

where $\delta_S = \left(\frac{P_k d_{k,e}^{\alpha}}{P_i \gamma_{th} d_{k,s}{}^{\alpha} + P_k d_{k,e}{}^{\alpha}}\right) \exp\left(-\frac{N_0 \gamma_{th} d_{k,s}^{\alpha}}{P_k}\right)$, $\kappa = \pi m \Gamma(m)\Gamma(1-m)$ and $\beta_S = \frac{\kappa \gamma_{th}{}^m d_{k,s}^2}{P_k{}^m}$, and $I_{s'}$ is the interference from all SCeNBs except for their respective small cell links.

So, from the above expressions, we can infer that the intensity of interference (i.e., outage probabilities of a SCeNB link) depends on small cells and DUEs density, distances between the receiver and transmitter of those links, SINR requirements, and transmit powers. By increasing DUE receiver distances from CUEs or SUEs will increase the success probability of D2D links. Similarly, the outage probability of SCeNB links will decrease by increasing distances between the D2D transmitter and CUEs or SUEs [1].

3.3. Transmission Rate

Signal-to-interference-plus-noise ratio (SINR) directly affects the achievable data rate. According to the Shannon capacity formula, the achievable data rate of i_{th} CUE, J_{th} D2D pairs, and k_{th} SUE link over the sub-carrier, f can be expressed as follows:

$$R_i = B\log_2(1 + \gamma_i^M). \tag{13}$$

$$R_j = B\log_2(1 + \gamma_j^D). \tag{14}$$

$$R_k = B\log_2(1 + \gamma_k^S). \tag{15}$$

Therefore, the total throughput of the entire network will be as follows:

$$R_{total} = \sum_{i=1}^{C} R_i + \sum_{j=1}^{D} R_j + \sum_{k=1}^{S} R_k$$

3.4. Problem Formulation

As shown in a D2D-enabled HetNet in Figure 1, when multiple DUEs and SUEs share the same radio resources, they will cause mutual interference. Therefore, our goal is to maximize the overall system throughput under the minimum QoS requirements of all users. Therefore, the objective function and constraints for mode selection, resource allocation, and power control of resource management issues among cellular, small cell, and D2D users can be formulated mathematically as follows:

$$\mathcal{P}1: \max_{\psi_j^f, \psi_k^f, \psi_j^k, P_i^C, P_j^D, P_k^S} R_{\text{overall}} = \sum_{i \in C} [B\log_2(1+\gamma_i^M) + \sum_{j \in D} \psi_j^f B\log_2(1+\gamma_j^D) + \sum_{k \in S} \psi_j^f B\log_2(1+\gamma_i^S)] \quad (16)$$

s.t.

$$C1: \gamma_i^M \geq \gamma_{i,th}^M, \quad \forall i \in C \quad (16a)$$

$$C2: \gamma_j^D \geq \gamma_{j,th}^D, \quad \forall j \in D \quad (16b)$$

$$C3: \gamma_k^S \geq \gamma_{k,th}^S, \quad \forall k \in S \quad (16c)$$

$$C4: 0 \leq P_i^C \leq P_{\max}^C, \quad \forall i \in C \quad (16d)$$

$$C5: 0 \leq P_j^D \leq P_{\max}^D, \quad \forall j \in D \quad (16e)$$

$$C6: 0 \leq P_k^S \leq P_{\max}^S, \quad \forall k \in S \quad (16f)$$

$$C7: \sum_{i \in C} \psi_j^f \leq 1, \quad \psi_j^f \in \{0,1\}, \quad \forall j \in D \quad (16g)$$

$$C8: \sum_{i \in C} \psi_k^f \leq 1, \quad \psi_k^f \in \{0,1\}, \quad \forall k \in S \quad (16h)$$

where $\gamma_{i,th}^C$, $\gamma_{j,th}^D$, and $\gamma_{k,th}^S$ are the minimum SINR requirements for the CUE, D2D pair, and SUE communication links, respectively. ψ_i^f, ψ_j^f, and ψ_k^f are mode selection indicators for cellular users, D2D pairs, and small cell communication, respectively. $\psi_j^f = \psi_k^f = 1$ when D2D pairs or SUE users reuse the i_{th} CUE resource; otherwise, $\psi_j^f = \psi_k^f = 0$. Constraints (16a), (16b), and (16c) represent QoS requirements of CUEs, DUEs, and SUEs, respectively. However, constraints (16g) and (16h) ensure that DUES or SUES can reuse at most one existing CUE resource. Finally, (16d), (16e), and (16f) constraints make sure that the transmit power of CUEs, DUEs and SUEs will be within the maximum limit.

4. ML-Based Resource Allocation

Wireless resource allocation techniques can be classified as traditional and learning-based algorithms. The above-formulated problem $\mathcal{P}1$ is integer nonlinear programming and NP-hard with polynomial time complexity. To address such optimization problems by using conventional methods, the original problem is divided into multiple sub-problems and solved individually. In addition, exact environmental information, such as channel state information (CSI), is required for this method, which is not feasible in a dynamic wireless network. To solve the above problem $\mathcal{P}1$, we propose a deep reinforcement learning (DRL) algorithm, which is a combination of reinforcement learning (RL) and deep learning (DL) techniques.

In DRL, sequential decision-making is addressed by minimizing a reward function by interacting with the unknown environment in each layer of applications. Combining DNN with Q-learning, also known as DQN, improves the learning performance as well as speed significantly. The most recent advancement of DRL can achieve the near-optimal solution of a sophisticated network optimization problem without knowing the complete and exact network information. Interacting with the environment, a DRL agent can obtain the best policy locally with or without exchanging information with other agents which significantly reduces communication overheads. Hence, DRL can be highly effective in a large, complex, and dynamic environment like 5G and B5G networks.

4.1. Machine Learning Approach

To overcome the limitations of traditional resource allocation methods and autonomous decision-making processes, ML and DL techniques can play a significant role [34]. with-

out being explicitly programmed, devices can learn, implement, and improve results automatically with the help of Machine Learning techniques [35]. Model training and decision-making are the two phases of an ML approach. In the training phase, a system model is trained by using training datasets, and later, this trained model is used to estimate each new input [36]. Depending on how the learning is achieved, Machine Learning algorithms are simply classified as supervised, unsupervised, and reinforcement learning [37]. In supervised learning, labeled training datasets are used to train the model, while in unsupervised learning, raw and unlabeled training datasets are used for the same, and in the reinforcement learning method, agents are used to interact with its environment and learn from the actions to achieve an optimal solution. By using historical data, supervised learning is used to predict or estimate future events. Regression, Decision Trees (DT), Support Vector Machines (SVM), Neural Networks (NN), and K-Nearest Neighbor (K-NN) are the most common supervised learning techniques. Whereas to discover the patterns and characteristics of raw, unlabeled data unsupervised learning is more suitable. Self Organizing Maps (SOM), K-means clustering, and Gaussian mixture models (GMMs) are the most important techniques in unsupervised learning [34,36].

In reinforcement learning, a learning entity known as an agent interacts with the environment and obtains rewards or penalties for each action; by this means, it creates a policy to set up its learning scheme and decides which action to choose in a certain situation. Maximizing the rewards over time is the main aim of the RL algorithm [38]. Unlike other ML algorithms, in RL agents are not advised what actions need to be performed to achieve optimal strategies, rather based on each interaction's output with the environment it makes the decision autonomously. Action-state, rewards and penalties, policy, value function, and environment model are the main components of an RL system.

4.2. Deep Learning Approach

Deep learning (DL), also known as deep network, is a part of machine learning where several layers of a neural network are used to mimic the behavior of the human brain. Artificial neural networks (ANN), recurrent neural networks (RNN), convolutional neural networks (CNN), deep neural networks (DNN), etc., are the different existing DL algorithms. One of the common and most widely used algorithms in the field of deep learning is DNN which is a collection of computational functions and an activation matrix that significantly improves the computational efficiency and provides better accuracy [39]. In DNN, each neuron generates its output when input meets certain conditions, and one neuron's output feeds into the next neuron's input to complete the chain [40]. In this way, the output of each layer is determined by the multiplication of input, weights, and the addition of a bias.

Deep learning has emerged as a powerful tool for addressing complex optimization problems in wireless communication systems, particularly in resource allocation. Traditional methods often require solving complicated optimization problems, which can be computationally expensive and difficult to implement in real-time environments. Due to flexibility, lower computational time, and the capability to achieve near-optimal performance without solving the complicated optimization problem explicitly [41–43], we propose a DQN-based DL algorithm.

4.3. Reinforcement Learning Approach

Reinforcement learning (RL) is an adaptable machine learning algorithm that can obtain the optimal decision-making policy in a dynamic environment without a training dataset [44] by using a trial-and-error process. Thus, it automatically adapts to the new environment. In the RL, the decision maker is called an agent.

Markov Decision Processes (MDP) are widely used as optimization tools for wireless communication systems due to the stochastic and dynamic nature of networks like fluctuating signal quality, user mobility and variable traffic loads. In the RL framework, MDP provides a mathematical model in the decision-making process, which is used to interact

between the agent and the environment for optimal strategies. Based on the outcome of each action, the agent either receives a reward or penalty, and the process randomly moves into a new state based on state transition probabilities. Figure 2 is a basic RL framework for a D2D-enabled cellular network where basic elements are illustrated. The different components of the reinforcement learning algorithm are described as follows:

AGENT: In the RL algorithm, D2D pairs act as agents. For each communication link, an agent learns and makes decisions by interacting with the environment. In MDP, an agent, under different network states, decides the proper communication modes and optimum resource blocks (RB) from the available radio resources. each D2D user's transmitter functions as an agent in D2D-enabled cellular HetNets.

STATE: To determine the states for a particular time slot, t a learning agent relies on the environmental conditions. The state observed by any agent, 'j' can be given as:

$$s_j = (A_{RB}, A_{CSI}, A_{QoS}) \in S \tag{17}$$

where A_{CSI}, A_{RB}, and A_{QoS} represent the observed channel status information, RB occupancy status among users, and QoS requirements for all users.

ACTION: In a D2D-enabled HetNet, the agent has various modes and a certain number of resource blocks available for communication. Based on current state conditions, an agent will decide on appropriate mode selection, transmit power, and RB allocation. So, the action of an agent a_j can be defined as:

$$a_j = (A_{MS}, A_{PC}, A_{RB}) \in A. \tag{18}$$

where A_{MS}, A_{PC}, and A_{RB} represent mode selection, transmit power control, and RB assignment for a specific action $a_j \in A$ under the current state s_j.

TRANSITION PROBABILITY:

Transition probability describes the changes in the environmental status during the interactions with the agent. For example, an agent executes an action, a, in a time slot, t, and state, s, and gains a reward, r. The probability of that agent will move to the next state, s', at time $t + 1$ is defined by the transition probability, $P_r(s'|s, a)$. So, the transition probability, $P_r(s'|s, a)$, is the probability that an agent executes the action $a \in A$ under the state $s \in S$ and moves into a new state $s' \in S$ [45], and it can be expressed as:

$$P_r(s'|s, a) = \begin{cases} 1, & s' = state(a), \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

REWARD:

In RL, the reward function is constructed to achieve the goal of the optimization problem. To maximize the overall system sum rate, each agent makes a joint decision for mode selection, power control, and resource allocation, and therefore, the reward of the agent, j, can be defined as follows:

$$R_j(s, a) = \begin{cases} R_j^D, & \text{if conditions C2, C5 and C7 are met,} \\ -1, & \text{otherwise.} \end{cases}$$

POLICY: Simply put, the policy is a decision-making rule for an agent. In any state, s, the strategy of the agent to execute an action, a, is represented by policy $\pi(s, a)$, where we have $a \in A$, $s \in S$, and $\sum_{a \in A} \pi(s, a) = 1, \forall s \in S$.

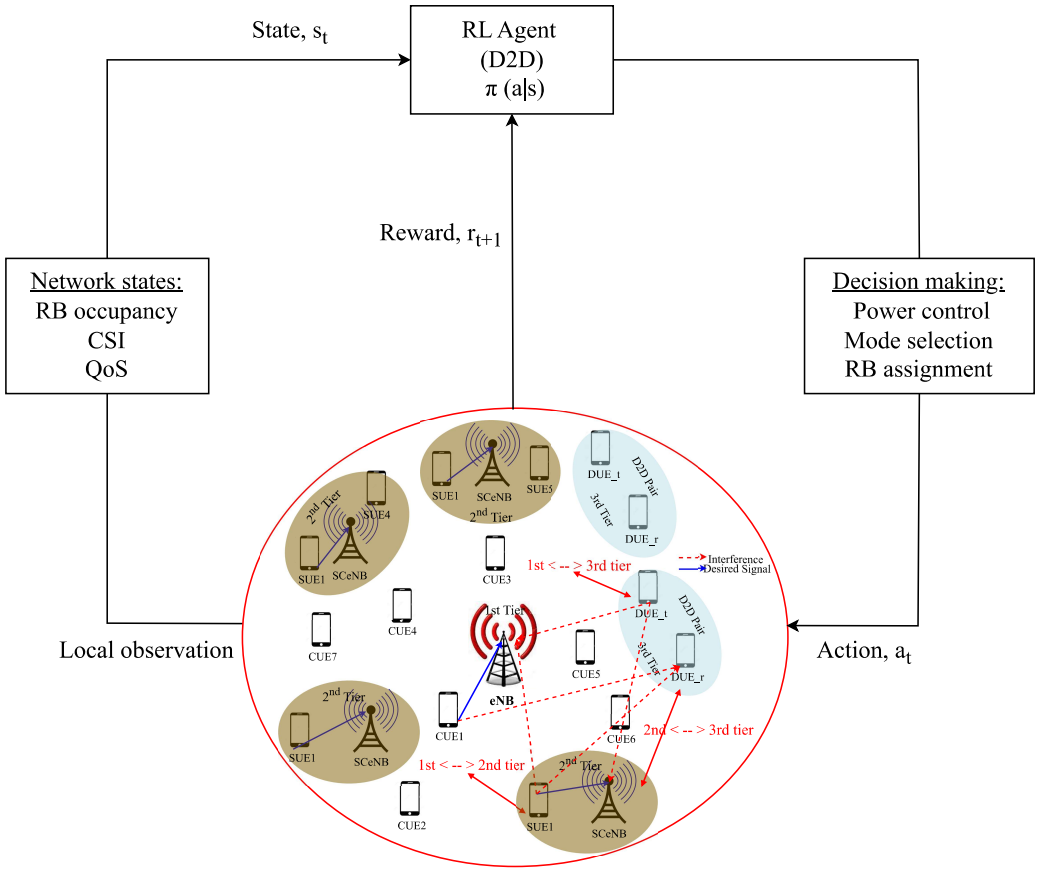

Figure 2. Agent–environment interaction in RL D2D-enabled cellular networks.

In an RL-based scheme for the given network, the D2D pair acts as an agent that interacts with the environment, generates trajectory, and moves into a new state $s \rightarrow s'$ by executing an action a. For each action executed by the agent, it will receive either a penalty or reward, r from the environment. In other words, in an RL-based D2D-enabled network, each j^{th} agent will act as follows: (i) Monitor its current status using the environment state $s \in S$; (ii) Perform one of the valuable action $a \in A$; (iii) Receive an instant penalty or reward, r from the environment; (iv) Move into a new state $s' \in S$. If the reward is based on the current state–action and not affected by any earlier state–action, then the j^{th} agent will satisfy the Markov property. In that case, MDP can be defined as follows: (i) A discrete set of environment states S_j; (ii) A discrete set of possible actions A_j; (iii) The probability of state transition for an environment-specific time slot, $P_r(s'|s) = P_r(s', a, s)$, where $\forall a_j \in A_j$ and $\forall s_j \in S_j$.

Q-learning is an extremely popular algorithm in RL to determine the optimal action of a system. Figure 2 represents the agent–environment interactions in an RL-based network where a reward is represented by a Q-function and the main goal is to maximize it. The agent explores various action states by interacting with the environment and updates the Q-value given by [46,47].

To solve the above MDPs problem smoothly, the state-value function $V^{\pi}(s)$ and the action-value function $Q^{\pi}(s, a)$ for $a \in A$ and $s \in S$ are defined as follows:

$$\begin{aligned} V^{\pi}(s) =& \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \eta^t r(t) | s_0 = s\right] \\ =& \mathbb{E}^{\pi}\left[r(t) + \eta V(s') | s' = s\right] \end{aligned} \tag{20a}$$

$$\begin{aligned} Q^{\pi}(s,\ a) =& \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\eta^{t}r(t)|s_0 = s,, a_0 = a\right] \\ =& \mathbb{E}^{\pi}[r(t) + \eta Q(s', a'] \\ =& r(s,a) + \eta \sum_{s'\in S} p(s'|s,a)\sum_{a'}\pi(a'|s')Q(s',a') \end{aligned} \tag{20b}$$

where $\eta \in (0,1)$ represents the discount factor and $r(s,a)$ is the immediate reward. $p(s'|s,a)$ and π represent the transition probability and policy for action, a, in environment state, s, respectively. Since expected future rewards are determined by the state–action value functions, they can be obtained for each state–action on every strategy. Therefore, at equilibrium, there will always be a maximum of state–action value functions and the optimal policy π^*, and hence Equations (20a) and (20b) can be converted as

$$\begin{aligned} V^{*}(s) =& \max_{\pi} V^{\pi}(s) \\ =& \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\eta^{t}r(t)(s,a,s')|s_0 = s,\pi\right] \end{aligned} \tag{21a}$$

$$\begin{aligned} Q^{*}(s,\ a) =& \max_{\pi} Q^{\pi}(s,a) \\ =& r(t) + \eta \arg\max_{a\prime} Q^{*}(s\prime, a\prime) \\ =& r(s,a) + \eta \sum_{s'\in S} p(s'|s,a)\max_{a'} Q^{*}(s',a') \end{aligned} \tag{21b}$$

Thus, π^* is obtained by $\pi^*(s) = arg\max_a Q^*(s,a)$. By using Equation (21b) we cannot derive the value of Q^* and π^* due to unknown transition probabilities. But to obtain them we can use a Q-learning algorithm where the RL agent will continuously update the Q-values by executing new actions to interact with the environment iteratively. Therefore, the updated Q-value functions can be expressed as

$$Q^{*}(s,\ a) \leftarrow Q(s,a) + \mu\left[(r + \eta \max_{a'} Q(s',a') - Q(s,a)\right] \tag{22}$$

where $Q^*(s,a)$ and $Q(s,a)$ represent the new and old Q value, respectively, μ is the learning rate, and the term $\max_{a'} Q(s',a')$ is the target value to the new state–action.

As the objective of the agent is to learn the strategy for maximizing the rewards, the action is selected based on QoS, i.e., SINR threshold, RB availability, and data rates. The actions that meet the constraints are rewarded otherwise negatively rewarded. To meet the objectives of C2, C5, and C7 constraints and maximize the throughput, the agent automatically learns and adapts to the newer situation under dynamic network conditions. Algorithm 1 represents the resource allocation and power control in a D2D-enabled cellular network using an RL-based algorithm.

Q-learning is a classical algorithm to learn state–action value functions in RL. However, for large state–action spaces, the environment is complex and dynamic. In such a situation, a Q-learning algorithm cannot use the Q-table to execute and store all state–action values. DRL is one of the algorithms which is used to overcome this issue.

Algorithm 1: RL-based resource allocation

Input: D2D-enabled HetNet cellular network environment parameters and minimum SINR requirements for D2D pairs, SUEs, and CUEs;
Output: Effective resource allocation with power control using RL algorithm in D2D-enabled HetNets according to optimal allocation policy $\pi^*(s_t) = \arg\max_a Q(s_t, a_t)$;
Initialization: Initialize the storing Q-value function for the state–action pairs;
Initialization: For all available actions in the selected current state s_t, select the action a_t for maximizing the Q-value function according to the $\epsilon - greedy$ strategy to obtain r_t;
for *each* $s \in S$, *each* $a \in A$ **do**
    for *episode* e = 1,2,3,. . . ,E **do**
        Initialize the current state s_t at time t in the D2D-enable HetNets Cellular network environment;
        for *steps* t = 1,2,3,. . . ,T **do**
            The agent (D2D pairs) takes action a_t based on $\epsilon - greedy$ strategy to obtain to obtain r_t;
            Execute action a_t and move into the next state s_{t+1} of the environment;
            In memory D, store the experience (s_t, a_t, r_t, s_{t+1}).
            Update the target network by (Equation (22))
            Update state–action as $s_t \leftarrow s_{t+1}$ and $a_t \leftarrow a_{t+1}$ until all $Q(s_t, a_t)$ converge
        End for
    End for
End for

4.4. Deep Reinforcement Learning Approach

Deep reinforcement learning (DRL) is a computational approach to learning from action, and the most common algorithm used for that is Deep Q-learning. DRL is the fusion of DL and RL algorithms where the agent learns the best actions from the environment to reach its goal. To overcome the issues in a large, complex, and dynamic environment, DRL uses a deep Q-network (DQN) algorithm that utilizes neural network models to predict the next appropriate action. Hence, RL can further be classified as model-based methods like deep-Q-Networks and model-free methods such as Q-Learning and state action reward state action (SARSA). In the Q-learning method, for each state–action pair a DNN is used by DQN algorithm to estimate Q(s, a) instead of calculating them. The DQN flowchart is described in Figure 3 where a DNN is used by DQN with weight θ for an action-value function network model $Q(s, a; \theta)$, and for the target network model $Q(s, a; \theta')$ another DNN with weight θ' is used. With each interaction with the environment, the action-state function is updated with the present iteration parameter θ'. Here DQN algorithm objective is to minimize the loss function $L(\theta)$, which is given by [46] as follows:

$$L(\theta) = \mathbb{E}\left[\left(y_t^{tar} - Q(s, a; \theta)\right)^2\right] \tag{23a}$$

where the target Q-value, y_t^{tar}, can be estimated by the following equation:

$$y_t^{tar} = r_{t+1} + \eta \max_a Q(s', a; \theta') \tag{23b}$$

where $a = argmax_a Q(s', a; \theta)$ and θ' will be updated by θ after every training step.

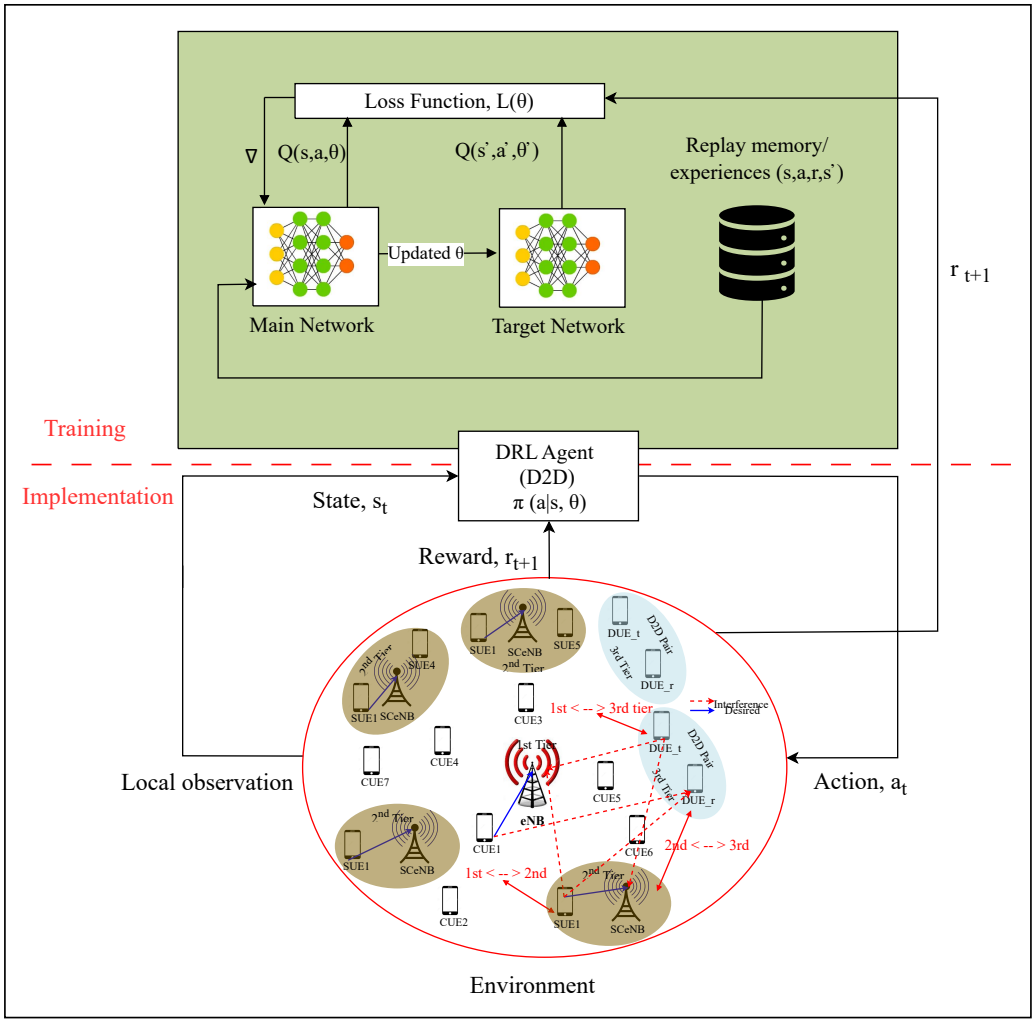

Figure 3. DQN framework for D2D-enabled cellular networks.

Like the Q-learning algorithm, to obtain the optimal value of θ^* and corresponding Q-values, an interactive process is adopted in DRL. In DQN, the agent uses an ϵ-greedy strategy to avoid local optimal solutions and receive a new experience when it interacts with the environment. During DQN training period, the agent will use the ϵ-greedy strategy to select the next action, a, from $Q(s, a; \theta)$, obtain reward r_t immediately and move into the new state s_{t+1} at time, $t + 1$ and finally store the quadruple information (s_t, a_t, r_t, s_{t+1}) in replay memory. During network training, the agent reuses these sample experiences data stored in the replay memory to improve the convergence speed and stability of the algorithm. After every iteration, the main DNN parameters θ of the network model in the action-value function are synchronized to the target network, and it is represented by

$$\theta \leftarrow \theta' + \tau \nabla L(\theta) \tag{24}$$

where $\tau \in (0, 1)$ and ∇ are the learning rate and gradient descent respectively [47]. Algorithm 2 represents the resource allocation and power control in a D2D-enabled cellular network using a DQN-based DRL algorithm.

The optimal policy can be obtained through the DRL agent by repeating the above procedure until the convergence of the weights θ. Converged DQN is well-trained to learn the patterns of the environment and is capable of managing unknown environmental states. Hence, by using the well-trained DQN model, each agent can make decisions autonomously to keep high performance without further training.

Algorithm 2: DRL-based resource allocation in D2D-enabled cellular network

Input: D2D-enabled HetNets cellular network environment parameters, and minimum SINR requirements for D2D pairs, SUEs, and CUEs;
Output: DQN/RL model training and decision results in the testing phase.;
Output: Effective resource allocation and power control;
Initialization: Set $\theta = \theta'$, a random DNN parameters initialization for the main network and target network as θ and θ' respectively.;
Initialize action functions $Q(s, a : \theta)$ and $Q(s, a : \theta')$ for the main and target DNN networks, respectively; target policy, π^* and Experience replay, D.
for $(cue = 1; cue \leq N_C; cue++)$ **do**
    Resource allocation for Cellular users, i.e., CUEs ;
for $(sue = 1; sue \leq N_S; sue++)$ **do**
    if *Check cellular resource reuse status for SCeNB, i.e., SUEs* **then**
        Reused for SCeNB
    else
    for $(due = 1; due \leq N_D; due++)$ **do**
        for *episode e = 1,2,3,. . . ,E* **do**
            Initialize the current state s_t at time t in the D2D-enable HetNets Cellular network environment;
            for *steps t = 1,2,3,. . . ,T* **do**
                The agent (D2D pairs) executes an action a_t based on $\epsilon - greedy$ strategy to obtain reward r_t;
                Compute the SINR using (7)
                Compute transmission power using (8)
                Take action a_t and move into the next state s_{t+1} of the environment;
                In memory D, store the experience (s_t, a_t, r_t, s_{t+1})
                Samples of historical experiences from the memory, D Uniformly and randomly
                Minimize loss function using gradient descent $L(\theta) = \sum_{e \in E}[y_t - Q(s, a; \theta)]^2$;
                Compute and update θ using (24) ; Update the target network by (23a).
            End for
        End for
    End for
End for

4.5. Computational Complexity

For any algorithm, complexity is defined as the amount of time taken to run it as a function of the size of the input. From Algorithm 2, we can see that the DNN parameters of the DQN algorithm converge into a stable state after E iterations and T time slots. The time complexity of our DNN model is determined by the number of input state actions, the number of layers in the neural network, and the number of neurons used in each layer. Let us consider that the deep neural network is fully connected, and there are L layers in our DNN, including input and out layers. In our analysis, we have considered one eNB and multiple SCeNB as well as D2D pairs. Let us assume the number of neurons for s^{th} layer for small cell communication is m_s. Therefore, the computational complexity for S layers in small cells is $\mathcal{O}\left(\sum_{s=0}^{S-1}(m_{s-1}m_s + m_s m_{s+1})\right)$. Similarly, for D2D communication in l_{th} layer, let us assume the number of neurons is n_d. The computational complexity for

D2D layers is $\mathcal{O}\left(\sum_{d=0}^{L-2}(n_{d-1}n_d + n_d n_{d+1})\right)$. So, the computation complexity for the overall training process considering all 3-tiers simultaneously is

$$\mathcal{O}\left(ET \cdot (\sum_{s=0}^{S-1}(m_{s-1}m_s + m_s m_{s+1}) \cdot \sum_{d=0}^{L-2}(n_{d-1}n_d + n_d n_{d+1})\right)$$

Therefore, the computational time complexity of our proposed algorithm increases in a multiplicative manner with the increasing number of D2D pairs as well as small cells, which makes it suitable for a large and dynamic D2D-enabled cellular HetNet.

5. Performance Evaluation

5.1. Simulation Environment and Parameters

We validated our proposed system model using MATLAB-based simulation. In the proposed model, the DRL-based learning algorithm is trained by its agent, i.e., DUEs, instead of training eNBs/SCeNBs, which helps to better coordinate between network elements. During simulation, we considered a single eNB located at the center of a D2D-enabled HetNet with a 500 m cell radius and CUEs are randomly distributed. DUEs and SUEs are realized according to two independent PPPs with λ_d and λ_s densities, respectively. We have assumed that one RB has been reused by multiple D2D pairs and SUEs which are already occupied by one CUE. Therefore, the number of CUEs is selected equal to the number of RBs.

A DQN-based small neural network (single input, three hidden and single output layers) is considered for quicker decision-making of each agent. The number of neurons in the hidden layers is 200, 100, and 50 respectively. DQN is trained with stochastic gradient descent with a batch size of 32. In addition, θ' is updated with θ in every 20-time steps. Batch size and epochs have significant effects on optimal performance like convergence, accuracy, and stability of the DQN model. To maintain the reasonable training time, memory requirements, and computation cost, we have chosen these values as shown in Table 4. As in our system model, micro cells are deployed with 10^{-5} density. we have chosen a moderate shadowing standard deviation of 8 dB with path loss coefficients 3 and 4 to make the test scenario as realistic as possible. Rest of the parameters like mobile maximum transmit power (23 dBm), RB Bandwidth (180 KHz) and Noise power (-118 dBm) are chosen as per the standard industry specifications or used by other similar research so that we can compare our model outcomes with their results effectively. Table 4 represents the list of parameters used in the network as well as DRL simulations for a D2D-enabled cellular HetNet.

Table 4. Parameters used in the simulation.

Simulation Parameter	Value
Cell Radius	500 m
SCeNBs density, λ_s	10^{-5}
DUEs density, λ_d	10^{-3}
CUEs maximum transmission power, P_i	23 (dBm)
DUEs maximum transmission power, P_j	10–20 (dBm)
SUEs maximum transmission power, P_k	10–20 (dBm)
Path loss exponent, α	3 and 4
SINR threshold for CUE, SUE and DUE, γ_{th}	8 dB

Table 4. *Cont.*

Simulation Parameter	Value
Noise power, N_0	−118 dBm
Shadowing standard deviation	8 dB
RB Bandwidth, B	180 KHz
Learning rate, τ	0.01
Discount factor, η	0.9
ϵ-greedy	0.1
Replay memory capacity, D	2000
Mini-batch size	32
Number of steps in each epoch	20

5.2. Results and Discussion

We measure the sum throughput of all DUEs for various conditions and outputs by simulation. Figure 4 shows the relationship between the sum throughput of DUEs/CUEs and the number of active users in the system. With increasing the number of DUEs, the overall system throughput increases. However, the increases in DUEs also introduce more interference between DUEs and CUEs link, which decreases the sum throughput of CUEs, as seen in Figure 5. From the output results, we can see that the proposed DRL scheme is much better compared to other schemes even maximizing their performance by setting power flexibility in optimized power schemes. The proposed DRL scheme can mitigate interference efficiently and achieve a higher sum throughput of DUEs. For instance, for the number of DUEs 50, the proposed DRL scheme achieves 10.2% and 44.1% higher sum throughput than D3QN DRL and Only Power Optimize, respectively.

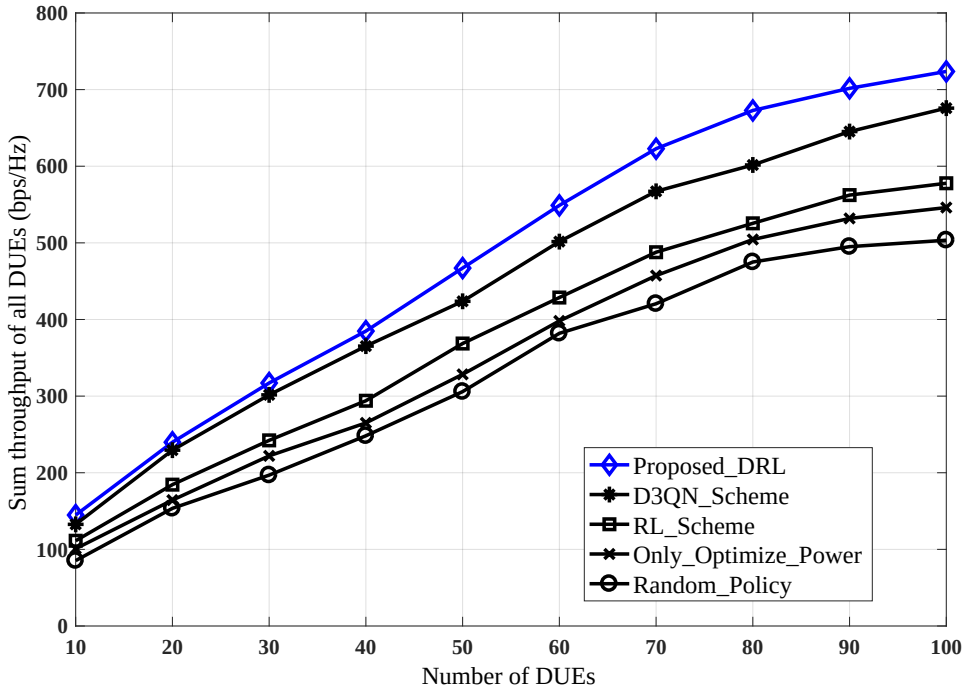

Figure 4. Sum throughput of all DUEs vs. no. of D2D pairs.

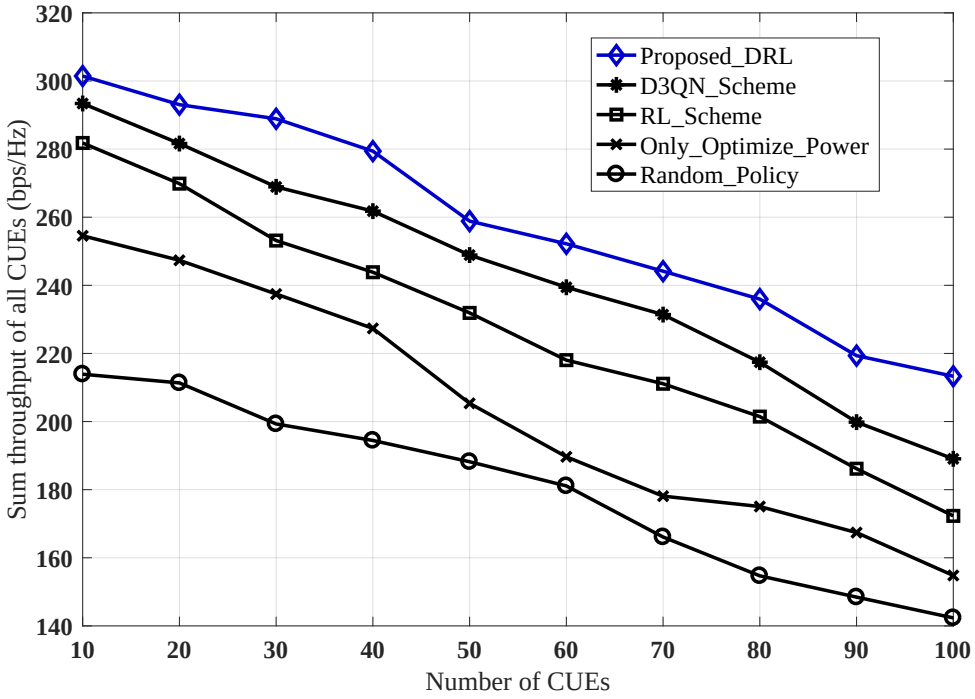

Figure 5. Sum throughput of all CUEs vs. no. of CUEs.

Figure 6 shows the sum throughput of DUEs as a function of the distance between D2D pairs in the system. With the increasing of distance between D2D pairs, the achieved sum throughput of DUEs for all schemes gradually decreases. With increasing the distances between D2D pairs from 10 m to 20 m, the sum throughput of all DUEs decreases by 14.7%, 19.4%, and 20.7% in proposed DRL, D3QN DRL, and Only Power Optimize schemes, respectively. Moreover, a 14.7% reduction indicates the proposed DRL scheme has a better sum throughout at higher distances between D2D pairs compared to others. This is happening due to two major reasons. First, the average channel gain between the Tx and Rx of D2D pairs decreases with increasing distances, and second, to ensure the minimum QoS requirements of D2D pairs, DUE Tx needs to increase its transmit power, which increases interference between other DUEs and CUEs.

Figure 7 represents the influence of the cell radius of eNB/SCeNB on the sum throughput for D2D pairs. Due to the increase in cellular cell radius, the mutual interference between users, and the interference severity impact of eNBs to D2D pairs reduces. As a result, with increasing the cell radius, the achieved sum throughput of DUEs by all schemes increases gradually, and our proposed DRL algorithm outperforms all the schemes, which can be seen in Figure 7. Increasing the cell radius from 200m to 400m, sum throughput increases by 40.2%, 38.9%, and 33.8% in the proposed DRL, D3QN DRL, and Only Optimize Power schemes, respectively.

Figure 8 shows the relation of the sum throughput of all DUEs with their Tx power. For all schemes, an optimal maximum transmission power results in a significant improvement in the sum throughput of DUEs. Increasing D2D pairs Tx power from 10dBm to 15dBm, sum throughput improves 10.5%, 10.2%, and 15.1% in our proposed DRL, D3QN DRL, and Only Power Optimize schemes, respectively. Only Optimize Power scheme is showing better slop at this point as it has focused power optimization only for DUEs without ensuring minimum QoS requirements for all users.

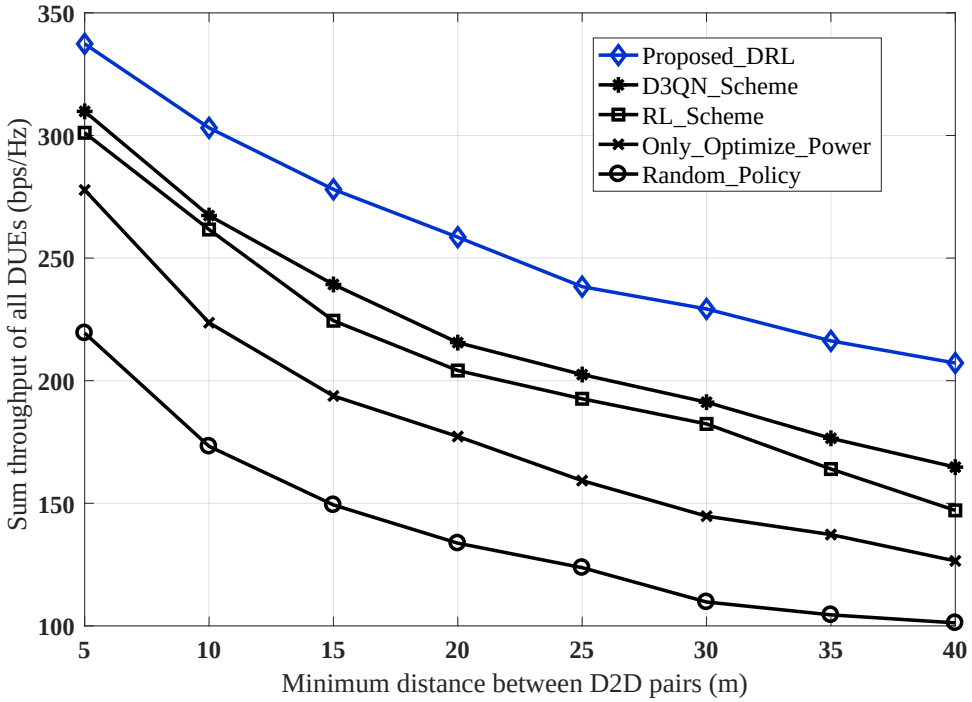

Figure 6. Sum throughput of all DUEs vs. distance between D2D pairs.

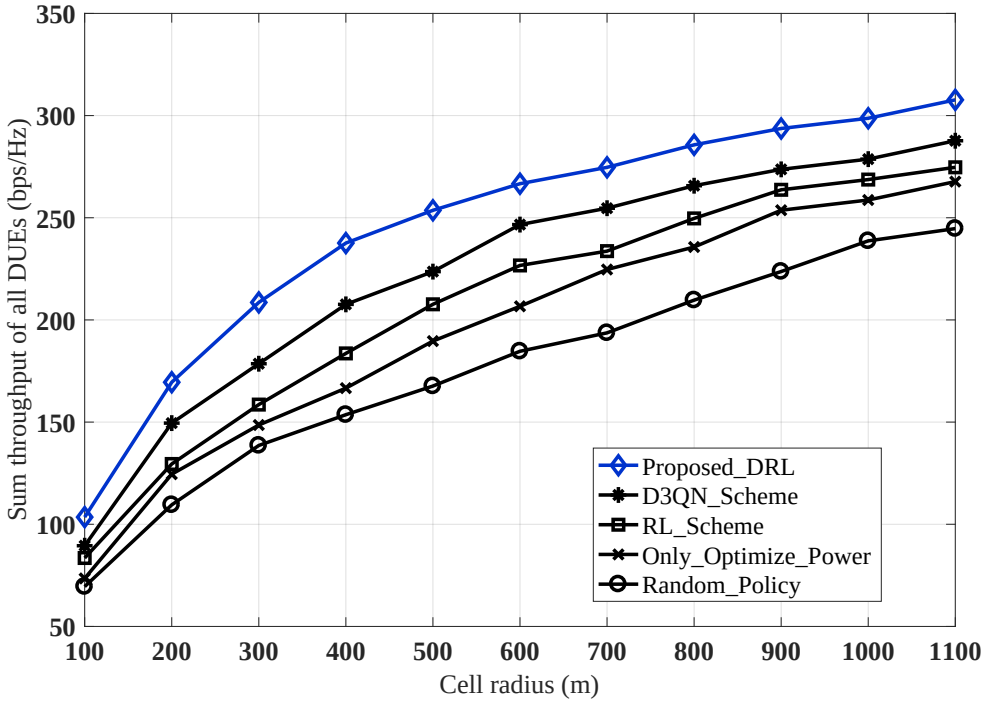

Figure 7. Sum throughput of all DUEs vs. cell radius.

Figure 9 shows the sum throughput of all D2D pairs against the minimum QoS requirements, i.e., minimum throughput requirements, to establish a successful link between the users. By increasing the minimum throughput requirement from 2 bps/Hz to 9 bps/Hz, the sum throughput drops from 300 bps/Hz to 148 bps/Hz for the proposed DRL, 107 bps/Hz for D3QN DRL, 85 bps/Hz for Only Optimize Power and 53 bps/Hz for random policy schemes respectively. The admission constraints become tighter with increasing the QoS requirements, and hence, the sum throughput of DUEs is degraded if a suitable power

control, mode selection, and resource allocation approach is not applied, which is reflected in our simulation results.

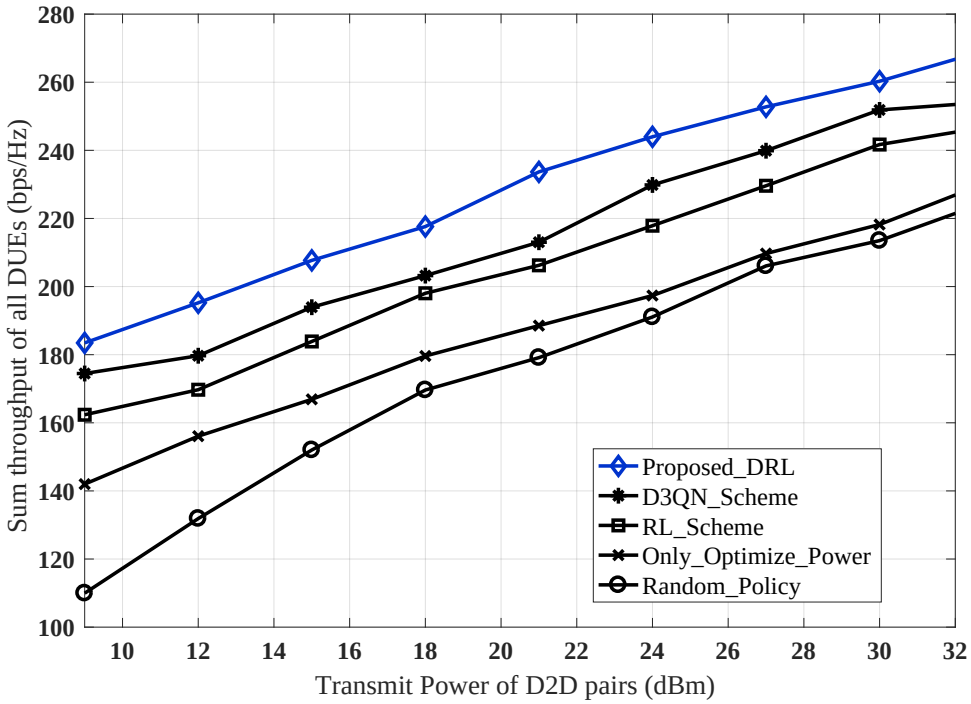

Figure 8. Sum throughput of all DUEs vs. transmit power of D2D pairs.

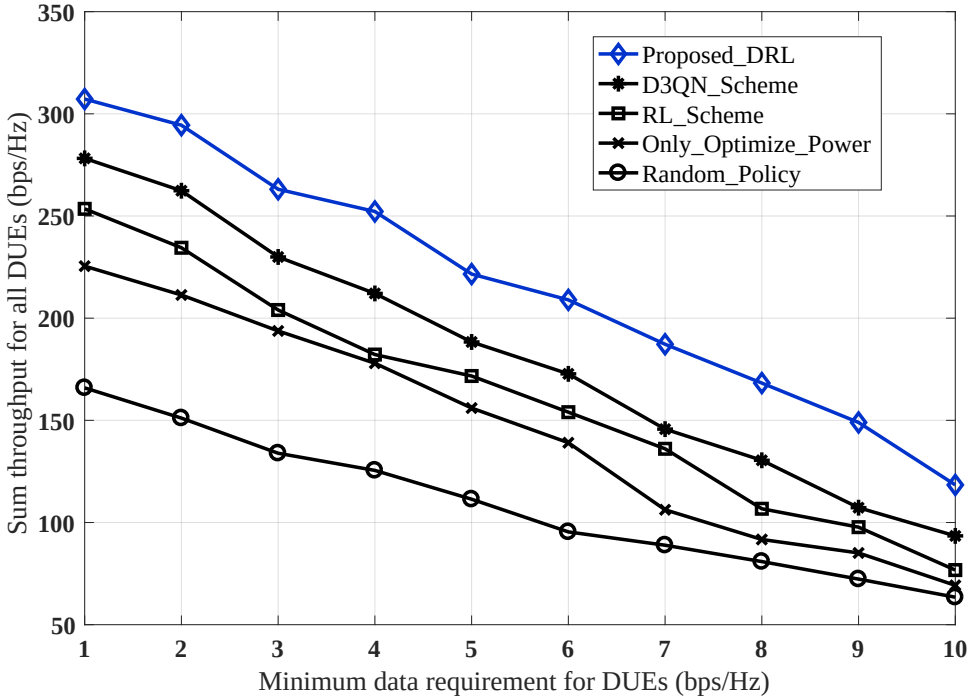

Figure 9. Sum throughput of all DUEs vs. minimum QoS requirement.

5.3. Model Validation and System Implication

The analytical model is validated by MATLAB-based simulation. This simulation tool is selected due to its availability and credibility. In simulation, we consider large-scale path loss model, path loss component, distance, and small-scale Rayleigh fading. To obtain good samples of data, the simulation was run longer to reach the steady state condition to avoid

biased results. For instance, we consider 1000 different seed values to have results with a confidence level of 95%. We create a more realistic simulation scenario with −174 dBm/Hz as noise power spectral density and 7 dB as noise figure. In the proposed DRL model, we consider DUEs instead of training eNBs/SCeNBs which help to make better coordination between network elements. A learning rate of 0.001, a discount rate of 0.9, a greedy rate of 0.1, a training memory size of 2000 MB, and a mini-batch size of 32 are used in the deep reinforcement learning algorithm. The DQN model is trained based on Algorithms 1 and 2.

Implementing D2D technology in a cellular network is challenging due to its several known issues. First, for a small-sized network where overheads are comparatively low, a centralized approach can be used. No matter what strategies are used, network performance degrades significantly with the increasing number of users as the network grows. In a D2D-enabled cellular HetNet where interference management is a major challenge; the proposed distributed DRL method can be useful to network designers to implement the system in a large-scale network scenario. So, our DRL approach can be used to design the network with desired throughput and QoS requirements for both cellular and D2D users with proper mode selection and controlled interference manner.

6. Conclusions

In this paper, we proposed a Q-network-based deep reinforcement learning (DRL) algorithm to mitigate interference in D2D-enabled heterogeneous cellular networks. We developed an analytical model by incorporating resource block assignment, mode selection, and transmit power control in a D2D-enabled HetNets for system modeling and analysis. The proposed DRL algorithm is trained using an agent-based decision-making policy to achieve the optimal solution for computational time, spectrum efficiency, and throughput. The system performance is validated by a MATLAB-based simulation. The simulation results obtained have shown that the proposed DRL method has achieved up to 52.7% higher sum throughput for all users on the network than the existing methods. The higher throughput is achieved as a result of the proposed interference mitigation algorithm in which transmitting power is adjusted dynamically. The proposed technique offers better system performance by keeping the minimum QoS requirements for all users in the network. To contribute to the development of next-generation cellular networks, the findings reported in this paper provide some insights into the machine learning-based resource allocation approaches in achieving guaranteed throughput and QoS. However, a test-bed measurement approach to validate the system's performance more realistically is suggested as a future research direction.

Author Contributions: Conceptualization, M.K.; investigation, M.K. and N.I.S.; methodology, M.K. and N.I.S.; project administration, N.I.S.; resources, N.I.S. and J.G.; supervision, N.I.S. and J.G.; validation, M.K. and N.I.S.; writing—original draft M.K.; writing—review and editing, N.I.S. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kamruzzaman, M.; Sarkar, N.I.; Gutierrez, J. A dynamic algorithm for interference management in D2D-enabled heterogeneous cellular networks: Modeling and analysis. *Sensors* **2022**, *22*, 1063. [CrossRef] [PubMed]
2. Jayakumar, S.; S, N. A review on resource allocation techniques in D2D communication for 5G and B5G technology. *Peer-to-Peer Netw. Appl.* **2021**, *14*, 243–269. [CrossRef]
3. Forecast, G. Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022. *Update* **2019**, *2017*, 2022.
4. Ericsson. *Ericsson Mobility Report: Global 5G Growth Amid Macroeconomic Challenges*; Ericsson: Hongkong, China, 2022.
5. Tehrani, M.N.; Uysal, M.; Yanikomeroglu, H. Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions. *IEEE Commun. Mag.* **2014**, *52*, 86–92. [CrossRef]

6. ETSI. *5G: Study on Scenarios and Requirements for Next Generation Access Technologies, 3GPP TR 38.913 Version 16.0.0 Release 16*; ETSI: Sophia-Antipolis, France, 2020.
7. Kamruzzaman, M.; Sarkar, N.I.; Gutierrez, J.; Ray, S.K. A mode selection algorithm for mitigating interference in D2D enabled next-generation heterogeneous cellular networks. In Proceedings of the 2019 International Conference on Information Networking (ICOIN), Kuala Lumpur, Malaysia, 9–11 January 2019; pp. 131–135.
8. Lin, X.; Andrews, J.G.; Ghosh, A.; Ratasuk, R. An overview of 3GPP device-to-device proximity services. *IEEE Commun. Mag.* **2014**, *52*, 40–48. [CrossRef]
9. Gu, B.; Zhang, X.; Lin, Z.; Alazab, M. Deep multiagent reinforcement-learning-based resource allocation for internet of controllable things. *IEEE Internet Things J.* **2020**, *8*, 3066–3074. [CrossRef]
10. Cai, X.; Zheng, J.; Zhang, Y. A graph-coloring based resource allocation algorithm for D2D communication in cellular networks. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 5429–5434.
11. Kuruvatti, N.P.; Hernandez, R.; Schotten, H.D. Interference Aware Power Management in D2D Underlay Cellular Networks. In Proceedings of the 2019 IEEE AFRICON, Accra, Ghana, 25–27 September 2019; pp. 1–5.
12. Xu, Y.; Liu, F.; Wu, P. Interference management for D2D communications in heterogeneous cellular networks. *Pervasive Mob. Comput.* **2018**, *51*, 138–149. [CrossRef]
13. Rezazadeh, F.; Chergui, H.; Christofi, L.; Verikoukis, C. Actor-critic-based learning for zero-touch joint resource and energy control in network slicing. In Proceedings of the ICC 2021-IEEE International Conference on Communications, Montreal, QC, Canada, 14–23 June 2021; pp. 1–6.
14. Luo, Y.; Shi, Z.; Zhou, X.; Liu, Q.; Yi, Q. Dynamic resource allocations based on Q-learning for D2D communication in cellular networks. In Proceedings of the 2014 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 19–21 December 2014; pp. 385–388.
15. Zia, K.; Javed, N.; Sial, M.N.; Ahmed, S.; Pirzada, A.A.; Pervez, F. Distributed Multi-Agent RL-Based Autonomous Spectrum Allocation in D2D-Enabled Multi-Tier HetNets. In *Interference Mitigation in Device-to-Device Communications*; John Wiley & Sons: Hoboken, NJ, USA, 2022; pp. 109–132.
16. AlQerm, I.; Shihada, B. A cooperative online learning scheme for resource allocation in 5G systems. In Proceedings of the 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 22–27 May 2016; pp. 1–7.
17. Zhu, L.; Liu, C.; Yuan, J.; Yu, G. Machine learning-based resource optimization for d2d communication underlaying networks. In Proceedings of the 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), Virtual, 18 November–16 December 2020; pp. 1–6.
18. Chen, Y.; Ai, B.; Niu, Y.; Guan, K.; Han, Z. Resource allocation for device-to-device communications underlaying heterogeneous cellular networks using coalitional games. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 4163–4176. [CrossRef]
19. Saied, A.; Okaf, A.; Qiu, D. An efficient resource allocation for d2d communications underlaying in hetnets. In Proceedings of the 2021 International Symposium on Networks, Computers and Communications (ISNCC), Dubai, United Arab Emirates, 31 October–2 November 2021; pp. 1–6.
20. Fan, Z.; Gu, X.; Nie, S.; Chen, M. D2D power control based on supervised and unsupervised learning. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; pp. 558–563.
21. Lee, W.; Lee, K. Resource allocation scheme for guarantee of QoS in D2D communications using deep neural network. *IEEE Commun. Lett.* **2020**, *25*, 887–891. [CrossRef]
22. Yuan, Y.; Li, Z.; Liu, Z.; Yang, Y.; Guan, X. Double deep q-network based distributed resource matching algorithm for d2d communication. *IEEE Trans. Veh. Technol.* **2021**, *71*, 984–993. [CrossRef]
23. Kai, C.; Meng, X.; Mei, L.; Huang, W. Deep reinforcement learning based user association and resource allocation for d2d-enabled wireless networks. In Proceedings of the 2021 IEEE/CIC International Conference on Communications in China (ICCC), Xiamen, China, 28–30 July 2021; pp. 1172–1177.
24. Huang, J.; Yang, Y.; He, G.; Xiao, Y.; Liu, J. Deep reinforcement learning-based dynamic spectrum access for D2D communication underlay cellular networks. *IEEE Commun. Lett.* **2021**, *25*, 2614–2618. [CrossRef]
25. Shi, D.; Li, L.; Ohtsuki, T.; Pan, M.; Han, Z.; Poor, H.V. Make smart decisions faster: Deciding d2d resource allocation via stackelberg game guided multi-agent deep reinforcement learning. *IEEE Trans. Mob. Comput.* **2021**, *21*, 4426–4438. [CrossRef]
26. Lee, W.; Schober, R. Deep learning-based resource allocation for device-to-device communication. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 5235–5250. [CrossRef]
27. Wang, D.; Qin, H.; Song, B.; Xu, K.; Du, X.; Guizani, M. Joint resource allocation and power control for D2D communication with deep reinforcement learning in MCC. *Phys. Commun.* **2021**, *45*, 101262. [CrossRef]
28. Tan, J.; Liang, Y.C.; Zhang, L.; Feng, G. Deep reinforcement learning for joint channel selection and power control in D2D networks. *IEEE Trans. Wirel. Commun.* **2020**, *20*, 1363–1378. [CrossRef]
29. Pan, Z.; Yang, J. Deep Reinforcement Learning-based Optimization Method for D2D Communication Energy Efficiency in Heterogeneous Cellular Networks. *IEEE Access* **2024**, *12*, 140439–140455. [CrossRef]
30. Guo, J.; Chen, J. Hybrid Action Space D2D Resource Allocation Algorithm Based on Multi-Agent Reinforcement Learning. In Proceedings of the 2023 3rd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Wuhan, China, 15–17 December 2023; pp. 375–380.

31. Jeong, Y.J.; Yu, S.; Lee, J.W. DRL-Based Resource Allocation for NOMA-Enabled D2D Communications Underlay Cellular Networks. *IEEE Access* **2023**, *11*, 140270–140286. [CrossRef]
32. Cai, Y.; Jin, S.; Yu, W.; Nie, X.; Liu, H. Cooperative Distributed Resource Allocation in Heterogeneous Networks with D2D Communication. *IEEE Trans. Veh. Technol.* **2023**, *72*, 16426–16440. [CrossRef]
33. Xu, Y. On the performance of device-to-device communications with delay constraint. *IEEE Trans. Veh. Technol.* **2016**, *65*, 9330–9344. [CrossRef]
34. Morocho-Cayamcela, M.E.; Lee, H.; Lim, W. Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions. *IEEE Access* **2019**, *7*, 137184–137206. [CrossRef]
35. Fourati, H.; Maaloul, R.; Chaari, L. A survey of 5G network systems: Challenges and machine learning approaches. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 385–431. [CrossRef]
36. Gures, E.; Shayea, I.; Ergen, M.; Azmi, M.H.; El-Saleh, A.A. Machine learning based load balancing algorithms in future heterogeneous networks: A survey. *IEEE Access* **2022**, *10*, 37689–37717. [CrossRef]
37. El Amine, A. Radio Resource Allocation in 5G Cellular Networks Powered by the Smart Grid and Renewable Energies. Ph.D. Thesis, Ecole Nationale supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, Nantes, France, 2019.
38. Cayamcela, M.E.M.; Lim, W. Artificial intelligence in 5G technology: A survey. In Proceedings of the 2018 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 17–19 October 2018; pp. 860–865.
39. Sun, H.; Chen, X.; Shi, Q.; Hong, M.; Fu, X.; Sidiropoulos, N.D. Learning to optimize: Training deep neural networks for wireless resource management. In Proceedings of the 2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Sapporo, Japan, 3–6 July 2017; pp. 1–6.
40. Song, Y.; Khandaker, M.R.; Tariq, F.; Wong, K.K.; Toding, A. Truly intelligent reflecting surface-aided secure communication using deep learning. In Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Virtual Event, 25–28 April 2021; pp. 1–6.
41. Sun, H.; Chen, X.; Shi, Q.; Hong, M.; Fu, X.; Sidiropoulos, N.D. Learning to optimize: Training deep neural networks for interference management. *IEEE Trans. Signal Process.* **2018**, *66*, 5438–5453. [CrossRef]
42. Lee, W.; Jo, O.; Kim, M. Intelligent resource allocation in wireless communications systems. *IEEE Commun. Mag.* **2020**, *58*, 100–105. [CrossRef]
43. Lee, W.; Kim, M.; Cho, D.H. Deep power control: Transmit power control scheme based on convolutional neural network. *IEEE Commun. Lett.* **2018**, *22*, 1276–1279. [CrossRef]
44. Zeng, Y.; Wang, G.; Xu, B. A basal ganglia network centric reinforcement learning model and its application in unmanned aerial vehicle. *IEEE Trans. Cogn. Dev. Syst.* **2017**, *10*, 290–303. [CrossRef]
45. Saied, A.; Qiu, D.; Swessi, M. Resource management based on reinforcement learning for D2D communication in cellular networks. In Proceedings of the 2020 International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada, 20–22 October 2020; pp. 1–6.
46. Hausknecht, M.; Stone, P. Deep recurrent q-learning for partially observable mdps. In Proceedings of the 2015 AAAI Fall Symposium Series, Arlington, VA, USA, 12–14 November 2015.
47. Kai, C.; Meng, X.; Mei, L.; Huang, W. Multi-agent reinforcement learning based joint uplink–downlink subcarrier assignment and power allocation for D2D underlay networks. *Wirel. Netw.* **2023**, *29*, 891–907. [CrossRef]