Base Station Assisted (BSA) Reinforcement Learning for Resource Allocation in Wireless Industrial Environments

Idayat O. Sanusi and Karim M. Nasr
Faculty of Engineering and Science, University of Greenwich, Kent, ME4 4TB, UK
{i.o.sanusi, k.m.nasr}@gre.ac.uk

Abstract— **Device-to-Device (D2D) enabled cellular networks are a promising solution for Ultra-Reliable Low-Latency Communication (URLLC) systems. Integrating D2D into future wireless industrial networks and next-generation manufacturing can support the creation of massive machine-type wireless connections. In this paper, we present a Base Station Assisted (BSA) reinforcement learning approach for resource allocation in a D2D-enabled cellular network targeting smart manufacturing and Industry 4.0 applications. A distributed local Q-table is used for the D2D agents to prevent global information gathering and a stateless Q-learning approach is adopted to reduce the complexity of learning and the dimension of the Q-table. The Q-tables of the D2D agents are then uploaded to the Base Station (BS) for the resource allocation to be implemented centrally. Simulation case studies show that the presented semi distributed BSA technique results in reduced signalling overheads and a good Quality of Service (QoS) across the network compared to other conventional schemes**.

Keywords—Fifth Generation (5G) and beyond networks; Radio Resource Management (RRM); Distributed Algorithms; Device-to-Device Communication (D2D); Reinforcement Learning.

I. INTRODUCTION

The increasing growth in the number of wireless smart devices and applications necessitates novel and efficient Radio Resource Management (RRM) schemes to address the different challenges faced. Device to Device (D2D) communication is considered one of the key technologies for 5G and beyond networks aiming to provide improvements in performance metrics such as throughput, spectrum, and energy efficiency especially for new verticals such as smart manufacturing and Ultra Reliable Low Latency Communication (URLLC) use cases, e.g., in wireless industrial applications [1]. Machine learning and artificial intelligence techniques are some of the main techniques currently gaining increased interest to realise the expectations of future generation wireless systems [2]-[3].

Spectrum access where a cellular and D2D users share the same resources can potentially result in improved spectrum efficiency. However, if shared resource allocation is not properly coordinated, mutual interference between cellular and D2D links may degrade the Quality of Service/Quality of Experience (QoS/QoE) of end-users.

Future wireless networks are characterised by a high density of devices and dynamic environments with rapidly changing Channel State Information (CSI). Centralised and distributed schemes are two RRM approaches used to allocate resources to users. In a centralised scheme, the global acquisition of CSI by a centralised controller (e.g., a Base Station (BS)) often incurs high signaling overheads and computation complexity which, tend to increase with the number of users, therefore making it impractical to deploy. Furthermore, RRM problems are often formulated as optimisation problems where the QoS requirements are modelled as the constraints. These optimisation problems are often complex and difficult to solve directly. A distributed approach does not need a central entity. Resource allocation is executed by users, therefore reducing the amount of information exchange, computations, and processing by the base station, and resulting in improved QoS across the network.

Game theory and machine learning are important techniques that can be used to realise a distributed RRM scheme. Matching theory, which, has been used to solve assignment or pairing problems between two distinct sets of players with diverse QoS objectives [4], may get complex in a multiuser scenario with rapidly changing channel conditions using full CSI, as in [5].

Reinforcement Learning (RL) has been explored to address RRM problems in dynamic environments [6]-[7]. RL is a machine learning approach, well-suited to support decision making in 5G-and-beyond networks with uncertainties, for example, in distributed resource allocation with unknown or partial information of network conditions. Q-learning is a reinforcement learning technique that uses a look-up table, known as Q-table, to determine an optimal strategy to adopt, by storing the values used to compute the maximum expected future rewards for actions taken at each state. A large number of agents, states and actions can lead to a high-dimension Q-table which, may result in slow convergence and limit the practical applications due to the high memory requirements [8]. These challenges can be addressed by using Deep Reinforcement Learning (DRL) which, uses deep neural networks to approximate the tables [9]. However, DRL is associated with high complexity and large learning data [10]-[11].

RL has been widely investigated to study intelligent power level and spectrum channel allocations for D2D-enabled cellular networks in a multi-agent environment. The work in [12] formulated the resource allocation problem as a stochastic non-cooperative game among D2D users. However, the QoS requirements of cellular users sharing the same frequency bands with D2D users were not considered

in the reward model. In [13], a multi-agent actor-critic framework was proposed which, involves cooperation between users and sharing of all historical information (states, actions, and policies) in a centralised training scheme to ensure stability. This will consequently increase the amount of signalling overheads and information exchange. In [12]-[15], the reward function captured the QoS metric of cellular users in a centralised Q-learning approach, which, also leads to increased signalling overheads.

In this paper, we present a semi-distributed reinforcement learning scheme for spectrum resource sharing of D2D Users (DUEs) and Cellular Users (CUEs) targeting smart manufacturing environments and URLLC networks. This semi distributed approach relies on two phases. First, a decentralised training of agents is implemented. This is followed by Q-tables being uploaded to the base station for final resource allocation. The reward function is modeled in such a way that there is no information exchange related to other agents' actions or rewards. To address the problem of the 'curse of dimensionality' associated with Q-learning, a stateless Q-learning approach is adopted to reduce the dimension of the Q-table, nonetheless capturing the QoS demands of the D2D users. The main contributions of this work are summarised below:

- A hybrid RRM scheme with distributed D2D training and a centralised channel allocation is presented with an advantage of reduced signalling overheads compared with conventional centralised approaches. This hybrid RRM scheme relies on stateless reinforcement learning algorithm is presented, where there is no state transition, to ensure a reduced dimension of the state-action mapping, nevertheless capturing the key performance metrics of the DUEs. With this technique, there is a decrease in complexity and signalling overheads making scheme adaptable to high-density networks.
- In previous works [16]-[18], the QoS of cellular users is captured by integrating it in the state space or reward function of the D2D users. Rather than the BS exchange the QoS estimation of the CUE with the DUE at each time slot, a Q-table for the CUEs is maintained and updated.
- Numerical simulations are used verify the performance of the presented algorithm in comparison to other approaches in terms of achieved throughput, signalling overheads and complexity.

The paper is organised as follows: The system model and problem formulation are presented in Section II. In Section III, a stateless reinforcement learning algorithm for base station-assisted resource allocation is presented. Section IV presents simulation case studies and results. The main conclusions and directions for future work are summarised in Section V.

II. System Model and Problem Formulation

We consider D2D and cellular users coexisting within a cellular network for uplink spectrum-sharing as illustrated in Fig. 1. There are N Cellular Users (CUEs) represented by a set $C = \{c_1, \dots, c_i, \dots c_N\}$ and M D2D Users (DUEs) denoted by a set $D = \{d_1, \dots, d_j, \dots d_M\}$ deployed randomly within the coverage of the base station in a single cell system.

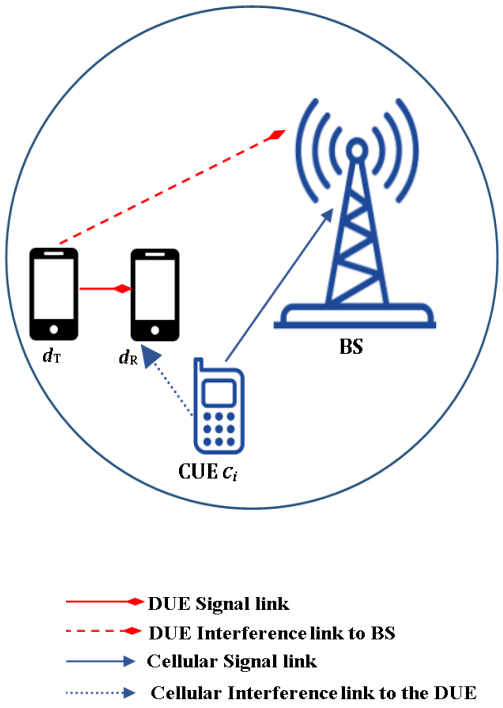

Figure 1. An illustration of a D2D enabled cellular network

The DUEs can autonomously select a Resource Block (RB) denoted by a set $K = \{k_1, \dots, k_i, \dots k_N\}$, from a pool of radio resources [18]19, which, can overlap with that of the CUEs for the benefit of reuse gain. The cellular users have strict performance requirements in the form of minimum Signal-to-Interference Plus-Noise-Ratio (SINR) values to guarantee their throughput. The D2D links also have minimum SINR thresholds to guarantee their throughput demands, in addition, to the reliability and delay requirements.

We assume that each CUE has been pre-allocated a resource block. The transmit power of the CUEs and DUEs are denoted by P_{c_i} and P_{d_j} respectively. $g_{c,B}$, $g_{d_T,B}$, g_{d_T,d_R} and g_{c,d_R} are the channel gains of cellular communication from the CUE c_i to the BS, the interference link from the DUE transmitter d_T to the BS, the D2D communication link

from the DUE transmitter d_T to the receiver d_R and the interference link form the CUE transmitter to the DUE receiver d_R, respectively. The channel gain comprises small-scale fading which, is assumed to be exponentially distributed with a unit mean and large-scale fading which, includes pathloss and shadowing with log-normal distribution.

The instantaneous received SINR at the BS from ith CUE and jth DUE over ith sub-channel at time slot t is given as [1]:

$$\Gamma_{c_i}(t) = \frac{P_{c_i} g_{c,B}(t)}{\sigma^2 + \sum_{d_j \in D} \lambda_j^i(t) P_{d_j} g_{d_T,B}(t)} \quad (1)$$

$$\Gamma_{d_j}(t) = \frac{P_{d_j} g_{d_T,d_R}(t)}{\sigma^2 + \sum_{c_i \in C} \lambda_j^i(t) P_{c_i} g_{c,d_R}(t)} \quad (2)$$

$\lambda_j^i \in \{0,1\}$ denotes the binary resource reuse indicator, $\lambda_j^i = 1$ implying that the jth DUE selects ith CUE sub-channel at time slot t and $\lambda_j^i(t) = 0$ otherwise. We assume that each DUE can access only one CUE sub-channel i.e., $\sum_N \lambda_j^i \leq 1$ and each CUE sub-channel is accessed by only one DUE i.e., $\sum_M \lambda_j^i \leq 1$. The data rates of the ith CUE and jth DUE is at time slot t given by:

$$\mathrm{T}_{c_i}(t) = W_i \log_2\big(1 + \Gamma_{c_i}(t)\big), \quad (3)$$

$$\mathrm{T}_{d_j}(t) = W_i \log_2\Big(1 + \Gamma_{d_j}(t)\Big), \quad (4)$$

where W_i is the bandwidth of each resource block. The variance of the Additive White Gaussian Noise (AWGN) is denoted by σ^2.

The channel gains for the different links (q, r) be expressed as follows:

$$\begin{cases} g_{c,B} = G_1 \gamma_{c,B} \chi_{c,B} L_{c,B}^{-\alpha_1} \triangleq \zeta_{c,B} L_{c,B}^{-\alpha_1} \\ g_{d_T,B} = G_2 \gamma_{d_T,B} \chi_{d_T,B} L_{d_T,B}^{-\alpha_2} \triangleq \zeta_{d_T,B} L_{d_T,B}^{-\alpha_2} \\ g_{d_T,d_R} = G_3 \gamma_{d_T,d_R} \chi_{d_T,d_R} L_{d_T,d_R}^{-\alpha_3} \triangleq \zeta_{d_T,d_R} L_{d_T,d_R}^{-\alpha_3} \\ g_{c,d_R} = G_4 \gamma_{c,d_R} \chi_{c,d_R} L_{c,d_R}^{-\alpha_4} \triangleq \zeta_{c,d_R} L_{c,d_R}^{-\alpha_4} \end{cases} \quad (5)$$

where G_r is the pathloss constant, $\gamma_{q,r}$ is the small-scale fading gain due to multipath propagation and assumed to have an exponential distribution with unit mean. The large-scale fading comprises pathloss with exponent α_r and shadowing which, has a slow fading gain $\chi_{q,r}$ with a log-normal distribution. $L_{q,r}$ is the distance from terminal q to terminal r [20].

The channel gain g_{d_T,d_R} and g_{c,d_R} can be estimated at the DUE receiver, d_R and made available at its transmitter, d_T instantaneously [19]. Similarly, $g_{c,B}$ and $g_{d_T,B}$ can be obtained at BS through local information since uplink transmission is considered.

The reliability of the DUE $d_j \in D$, $\xi_{d_j}(t)$, is defined as the probability of packet delay exceeding a predefined delay bound, $l_{d_j,\max}$, on channel i at slot t is less than a threshold [21]. The objective of the system is to maximise the total throughput, T_R, of paired CUEs and DUEs while satisfying the QoS demands. The optimisation problem and constraints are described in (6).

$$\mathbf{Max}_{\lambda_j^i} \; T_R = W_i\big(\lambda_j^i\big(\textstyle\sum_{c_i \in C} \log_2\big(1 + \Gamma_{c_i}\big) + \sum_{d_j \in D} \log_2(1 + \Gamma_{d_j})\big)\big) \quad (6)$$

subject to

$$\lambda_j^i \Gamma_{c_i} - \Gamma_{c_i,\min} \geq 0 \qquad \forall c_i \in C \quad (6a)$$

$$\Pr\Big(l_{d_j} > l_{d_j,\max}\Big) < 1 - \xi_{d_j}^* \qquad \forall\, d_j \in D \quad (6b)$$

$$\textstyle\sum_{c_i \in C} \lambda_j^i \leq 1 \qquad \forall\, d_j \in D \quad (6c)$$

$$\textstyle\sum_{d_j \in D} \lambda_j^i \leq 1 \qquad \forall\, c_i \in C \quad (6d)$$

The minimum SINR, $\Gamma_{c_i,\min}$, to guarantee the throughput requirement of the CUEs is defined in constraint (6a). Constraint (6b) takes into account reliability and delay, where l_{d_j} is the packet delay constraint for packet transmission of DUE d_j. The expression captures the fact that the end-to-end delay should be less than $l_{d_j,\max}$ with a probability of at least $1 - \xi_{d_j}^*$. Constraints (6c) and (6d) are channel association criteria. The reliability of the DUE links in (6c) is evaluated using an empirical estimation of number of packets transmitted similar to [21], from d_T to d_R whose delay is within the budget $l_{d_j,\max}$ over the total number of packets sent to d_R at time slot t i.e.,

$$\xi_{d_j}(t) = 1 - \Pr\Big(l_{d_j} > l_{d_j,\max}\Big) \approx 1 - \frac{L_{d_j}(t)}{B_{d_j}(t)} \cong \frac{L'_{d_j}(t)}{B_{d_j}(t)}, \quad (7)$$

where $L_{d_j}(t)$ is the number of packets for which, $l_{d_j} > l_{d_j,\max}$ and $L'_{d_j}(t)$ is the number of packets transmitted with $l_{d_j} \leq l_{d_j,\max}$ (or number of packets delivered within the delay bound). $B_{d_j}(t)$ is total packet transmitted by DUE d_j at time slot t. Reliability can also be measured in terms of the outage probability, which, is the probability that the measured SINR is lower than a minimum is less than a predefined threshold. The expression of the outage probability of jth DUE conditioned on the selected ith channel at time slot t is given below [22].

$$p_R(t) = \Pr\Big(\Gamma_{d_j} \leq \Gamma_{d_j,\min}\Big) = 1 - \frac{P_{d_j} g_{d_T,d_R} \exp\left(-\frac{\Gamma_{d_j,\min}\sigma^2}{P_{d_j} g_{d_T,d_R}}\right)}{P_{d_j} g_{d_T,d_R} + \Gamma_{d_j,\min} P_{c_i} g_{c,,d_R}} \leq p_{R_0}, \quad (8)$$

where $p_R(t)$ is the measured outage probability of DUE d_j at time slot t and p_{R_0} is the maximum tolerable outage probability of d_j.

The reliability of the DUE in terms of outage probability is expressed as [21]:

$$\xi_{d_j}(t) = 1 - p_R(t). \quad (9)$$

Transmission delay is given as the ratio of packet size transmitted within delay bound to the transmission rate [23]. From (7), (8) and (9) the transmission delay of jth DUE using the ith RB is formulated as:

$$l_{d_j}(t) = \frac{L'_{d_j}(t)}{W_i \log_2(1 + \Gamma_{d_j})}. \quad (10)$$

At each time slot t, the resource allocation system implements two functions, namely:
i) determining the SINR, Γ_{c_i} for the ith CUE and the SINR Γ_{d_j} that the jth DUE to ensure that the minimum SINR and target reliability $\xi^*_{d_j}$ thresholds are achieved and
ii) allocating RBs to jth DUE so that T_R is maximised.

The resource allocation optimisation problem for D2D communication in a cellular network is NP hard and a direct solution is not feasible. We present a base station-assisted resource allocation scheme which, adopts a semi-distributive RRM approach.

III. Stateless Reinforcement Learning for Base Station-Assisted Resource Allocation

The goal of the agents is to maximise throughput in a D2D-enabled cellular network. At each time slot t, a DUE observes a state s^t and takes an action a^t from the action space (i.e., select an RB k_i), according to a policy π. Q-learning enables an agent to determine the optimal strategy that maximises its long term expected cumulative reward. The Q-value is updated as follows [23]:

$$Q^{t+1} = \begin{cases} Q^t(s^t, a^t) + \sigma\left[r^t + \eta \max_a Q^t(s^{t+1}, a^{t+1}) - Q^t(s^t, a^t)\right] \\ \text{if } s = s^t,\ a = a^t \\ Q^t(s^t, a^t), \quad \text{otherwise} \end{cases}, \quad (11)$$

where $\sigma \in [0,1]$ is the learning rate. With $\sigma = 0$, the Q-values are never updated, hence no learning has taken place; setting σ to a high value such as means that learning can occur quickly and $0 \leq \eta \leq 1$ is the discount factor used to balance immediate and future reward [24].

The state space, action space and rewards function in the learning environment are defined as follows:

1) State Space: The state observed by DUE $d_j \in D$, $S^i_{d_j}(t)$, using resource block RB k_i at time slot t is defined by three variables, resulting in eight possible states as defined in Table I.

$$S^i_{d_j}(t) = \left\{S^i_{\Gamma_{d_j}}, S^i_{\xi_{d_j}}, S^i_{l_{d_j}}\right\}, \quad (12)$$

where $S \in S^i_{d_j} = \{0,1\}$. $S^i_{\Gamma_{d_j}}(t)$ indicates the interference level and is defined as:

$$S^i_{\Gamma_{d_j}}(t) = \begin{cases} 1 & \Gamma_{d_j}(t) \geq \Gamma_{d_j,\min} \\ 0 & \text{otherwise} \end{cases}, \quad (12a)$$

$S^i_{\xi_{d_j}}(t)$ indicates the level of reliability and is defined as:

$$S^i_{\xi_{d_j}}(t) = \begin{cases} 1 & \xi_{d_j}(t) \geq \xi^*_{d_j} \\ 0 & \text{otherwise} \end{cases}, \quad (12b)$$

$S^i_{l_{d_j}}(t)$ indicates the packet transmission time and is defined as:

$$S^i_{l_{d_j}}(t) = \begin{cases} 1 & l_{d_j}(t) \leq l_{d_j,\max} \\ 0 & \text{otherwise} \end{cases}, \quad (12c)$$

TABLE I. State Space for DUEs

$S^i_{\Gamma_{d_j}}$	$S^i_{\xi_{d_j}}$	$S^i_{l_{d_j}}$	$S^i_{d_j}$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

The state-action dimension is reduced by adopting a stateless learning approach. For the considered scenario, an action $a_i \in A$ taken by an agent will result in the end of an episode i.e., states 0 and 1 are terminal states, where $S^i_{d_j}(t) = 1$ is the goal state of the DUEs. Therefore, the learning environment can be modelled entirely using a stateless Q-learning i.e., action-reward only since the state transition is not required. An agent can choose its action based solely on its Q-value. The updated Q-value of the chosen action is based on the current Q-value and the

immediate reward from selecting that action. The update function in (11) is re-formulated as follows:

$$Q^{t+1}(a^t) = \begin{cases} Q^t(a^t) + \sigma[r(a^t) - Q^t(a^t)], \text{if } a = a^t \\ Q^t(a^t), \quad \text{otherwise} \end{cases} \quad (13)$$

where $r(a^t)$ is the immediate reward of selecting a.

In contrast to the standard Q-value update function in (11), it can be seen in (13) that not only the state-action formation (s, a) is not necessary, but also the information of the next state s^{t+1} is not required because the actions lead to a terminal state. Therefore, the Q-table is defined in terms of the actions only and updated using the immediate reward. This results in $1 \times |N|$ dimension Q-table for jth DUE. This method reduces the learning complexity and the Q-table dimension.

The traditional cellular users in the network need to be protected from the interference caused by the DUEs for their minimum SINR to be satisfied. This may be achieved by integrating the SINR of the CUE, Γ_{c_i} in the state space or by reward function modelling. This way, the DUEs can obtain the information from the BS at time slot t as in [17]-[18], [25]; hence, the DUEs get a reward if the CUE SINR $\Gamma_{c_i} \geq \Gamma_{c_i,\min}$, and a penalty otherwise. Rather than the BS exchange the measured CUE SINR, Γ_{c_i}, with the DUEs for every action a^t taken at each time slot, we adopt a scheme in which, the BS keeps a look-up table of the ith CUE based on the actions on the DUEs. Therefore, the Q-table for the ith CUE is $1 \times |M|$ considering a stateless Q-learning structure.

2) Action Space: The action space of DUE $d_j \in D$ is a set of all actions denoted by $A = \{a_1^t, \dots, a_i^t, \dots a_N^t\}$, where a_i^t is the action taken by $d_j \in D$ at time slot t and defined as the selection of an RB k_i.

3) Action-Selection Strategy: There are methods to select an action based on the current evaluation of the Q-value at every time slot t using a policy denoted by $p_{d_j}^t$. These methods are used to balance the exploration and exploitation of actions taken by the agents [26]. Epsilon greedy (ε-greedy) is one of the methods of choosing an optimal Q-value and described as follows:

$$p_{d_j}^t = \begin{cases} \underset{a \in A}{\text{argmax}}\, Q(a) & \text{probability } 1 - \varepsilon \text{ (exploitation)} \\ \text{Random action} & \text{probability } \varepsilon \text{ (exploration)} \end{cases} \quad (14)$$

where ε is the exploration rate with $0 \leq \varepsilon \leq 1$. The exploration rate is the probability that the agents will explore the environment rather than exploit it. $\varepsilon \to 1$ results in greater exploration whereas $\varepsilon \to 0$ means greater exploitation.

4) Reward Function: The reward function is modelled such that it relies only on local observations and can be implemented in a distributive manner. The rewards of the jth DUE and ith CUE for taking an action a_i^t is expressed in terms of the achievable throughput using the Shannon capacity formula. Thus, the reward is directly related to the objective function of the optimisation problem.

Equation (15) shows that jth DUE only gets a reward when all the state variables are 1 (i.e., the minimum QoS demands are met), while ith CUE gets a reward if its minimum SINR is satisfied at each time slot for the action taken by jth DUE. From the reward function defined above, learning can be implemented independently in a decentralised manner such that each agent maintains a local Q-table. There is no information exchange relating to other agents' actions or rewards and no cooperation is needed between the agents, which, results in reduced signalling overheads and reduced complexity compared with a centralised Q-learning approach.

$$r_{d_j}(a^t) = \begin{cases} T_{d_j}^k(t) & S_{d_j}^i(t) = 1 \\ 0, & S_{d_j}^i(t) = 0 \end{cases}, \quad (15a)$$

$$r_{c_i}(a^t) = \begin{cases} T_{c_i}^k(t) & \Gamma_{c_i} \geq \Gamma_{c_i,\min} \\ 0, & \text{otherwise} \end{cases}. \quad (15b)$$

The Q-value of the jth DUE for selecting ith RB at time slot t is updated as follows:

$$Q_{d_j}^i(a^t) = \begin{cases} Q_{d_j}^i(a^t) + \sigma\left[r_{d_j}(a^t) - Q_{d_j}^i(a^t)\right], \text{if } a = a^t \\ Q_{d_j}^i(a^t), \quad \text{otherwise} \end{cases}. \quad (16a)$$

Similarly, the Q-value of the ith CUE for action taken by the jth DUE is updated as follows:

$$Q_{c_i}^j(a^t) = \begin{cases} Q_{c_i}^j(a^t) + \sigma\left[r_{c_i}(a^t) - Q_{c_i}^j(a^t)\right], \text{if } a = a^t \\ Q_{c_i}^j(a^t), \quad \text{otherwise} \end{cases}. \quad (16b)$$

From (16), it can be seen that after the training, the Q-table of the jth DUE, $Q_{d_j}(a)$, will return $Q_{d_j}^i(a) = 0$ for its action on ith RB that do not meet its QoS requirements. Similarly, the Q-table of the ith CUE, $Q_{c_i}(a)$, will return $Q_{c_i}^j(a) = 0$ for the action of $\boldsymbol{j}$th DUE on $\boldsymbol{i}$th RB that do not meet its QoS requirements.

The BSA algorithm summarised in Algorithm I, aims to optimise the achieved system throughput. After the training phase, each DUE loads its Q-value table, $Q_{d_j}(a)$, to the BS for centralised matching. The BS will then allocate cellular resource blocks to D2D links in such a way that spectrum sharing is optimised, network throughput is maximised and there is no need for information exchange between the UEs to find a suitable candidate.

Algorithm I: The BSA Reinforcement Learning Algorithm

1: Initialise the action-value function for the DUEs
$Q_{d_j}(a) = 0 | Q_{d_j}(a) \equiv Q^i_{d_j}(a^t), i = 1,2,\dots,N \quad \forall\, d_j \in D$

2: Initialise the action-value function for the BS for the actions of the jth DUE on the ith RB
$\left[Q_{c_i}(a) = 0 | Q_{c_i}(a) \equiv Q^j_{c_i}(a^t),\ j = 1,2,\dots,M\right] \quad \forall\, c_i \in C$

3: **for** $d_j \in D$ $1 \le j \le M$ do
4: **while** not converge **do**
5: generate a random number $x \in \{0,1\}$
6: **if** $x < \varepsilon$ **then**
7: Select action a_i^t randomly
8: **else**
9: Select action a_i^t=argmax$_{a\in A}\, Q_{d_j}(a^t)$
10: **end**

11: Evaluate ξ_{d_j}, Γ_{d_j} and l_{d_j} of $d_j \in D$ for the action a^t

12: Measure the SINR, ξ_{c_i}, of CUE $c_i \in C$ for the action a^t taken by $d_j \in D$
13: Observe immediate reward of $d_j \in D$ and $c_i \in C$,

14: Update action-value for action of $d_j \in D$ on the ith RB $Q^i_{d_j}(a) = Q^i_{d_j}(a) + \sigma\left[r_{d_j}(a^t) + Q^i_{d_j}(a)\right]$

15: Update action-value for $c_i \in C$ for action a^t of jth DUE $Q^j_{c_i}(a) = Q^j_{c_i}(a) + \sigma\left[r_{c_i}(a^t) + Q^j_{c_i}(a)\right]$

16: **end while**
17: **end for**

18: Load $Q_{d_j}(a)$ to the BS $\quad \forall\, d_j \in D$

19: **for** $d_j \in D$ $1 \le j \le M$ **do**
20: Obtain $Q(a) = \left\{Q^i_{d_j}(a), Q^j_{c_i}(a)\right\}$ $i = 1,2,\dots,N$
21: $\bar{Q}(a) \subseteq Q(a) | \left\{Q^i_{d_j}(a), Q^j_{c_i}(a)\right\} \in \mathbb{R}^+$, where $\mathbb{R}^+$ positive real number
22: $Q_{\text{TOT}} = Q^i_{d_j}(a) + Q^j_{c_i}(a) \quad \forall\, q \in \bar{Q}(a)$
23: **end for**

24: Set up a list for unmatched DUE $D_u = \{d_j : \forall\, d_j \in D_u\}$
25: **while** $D_u \neq \emptyset$ **do**
26: Rank D_u in increasing order of $|0\ \bar{Q}(a)|$
27: Start DUE $d_j \in D_u$: $\bar{Q}(a) \neq \emptyset$ with the least $|\ \bar{Q}(a)|$
28: $c_i^* = \max_{r_i \in R} Q_{\text{TOT}}$
29: $D_u = D_u - d_j$
30: $\bar{Q}(a) = \bar{Q}(a) \backslash c_i^* \quad \forall\, d_{j'} \in D_u | \, j' \neq j$
31: **end while**

IV. Simulation Case Study and Performance Evaluation

The performance of the BSA approach described in Section III, is verified by considering a single-cell network in an industrial scenario. The simulation set-up and channel models are as described in [1] and summarised in Tables II and III. The network dynamics is captured by generating the channel fading effects randomly. The throughput is the main metric used to evaluate the performances of the algorithms. The performance of BSA is compared with centralised optimisation and the game theoretic Deferred Acceptance (DA) techniques [1][20].

A. Throughput Performance

The throughput performance of matched DUEs as a function of the number of DUEs in the system M, is shown in Fig. 2. It can be concluded that the sum throughput of the DUEs increases with the number of cellular users M for all the considered algorithms. As expected, the number of admitted DUEs increases with the introduction of new DUEs to the system, but unchanged if a valid cellular resource-sharing partner cannot be found because the minimum QoS requirements are not satisfied. The performances of centralised and BSA approaches are comparable, while the DA method shows the least performance. The BSA algorithm outperforms the DA algorithm by up to 9.69% increase in the DUE throughput performance. However, it is semi-distributive as the final resource allocation is implemented by the BS whereas the DA approach is decentralised (the channel selection is user-centric, and no BS intervention is necessary to achieve autonomy). Players can make their resource allocations choices to maximise their individual throughput and ultimately achieve system stability. The performance of the sum throughput of the matched UEs (that is valid pairings between CUEs and DUEs) with respect to the number of cellular users M is shown in Fig. 3. The sum throughput increases with M. The BSA approach indicates better performance at $M \le 35$ with up to 12.05% increase in sum throughput compared to the centralised approach, while the centralised approach performed better at $M > 35$ with up to 9.39% increase in throughput. The DA algorithm again shows the least performance compared to the BSA technique.

The effects of the outage probability of the DUE, p_{R_0}, and delay threshold of the DUEs, $l_{d_j,\max}$ on the sum rate of the matched UEs for all algorithms are shown in Fig. 4 and Fig. 5, respectively. The sum throughput of the matched UEs increases with p_{R_0} and $l_{d_j,\max}$. This is because higher p_{R_0} causes the interference from the CUEs to be more tolerable by the DUEs, therefore making potential CUE-DUE pairing possible. Similarly, higher $l_{d_j,\max}$ increases the sum throughput at fixed outage probability and payload since the delay requirement is less stringent. More DUEs are able to

satisfy the delay constraint and the number of admitted DUEs is increased.

TABLE II. MAIN SIMULATION PARAMETERS [1][20][27]

Parameter	Value
Carrier frequency, f_c	2GHz
System bandwidth	10MHz
Number of resource blocks (RB), K	50
RB bandwidth	180 kHz
Maximum CUE transmit power, $P_{c_i,\max}$	23dBm
Maximum DUE transmit power, $P_{d_j,\max}$	13dBm
D2D distance, L_{d_T,d_R}	$10\text{m} \leq L_{d_T,d_R} \leq 20\text{m}$
CUE SINR Threshold, $\Gamma_{c_i,\min}$	7 dB
DUE SINR Threshold, $\Gamma_{d_j,\min}$	3 dB
Noise power density	-174 dBm/Hz
Number of CUEs, N	50
Number of DUEs, M	50
Reliability for DUE, p_{R_0}	10^{-5}
Exploration rate, ε	0.7
Learning rate, σ	0.9
DUE Maximum Delay, $l_{d_j,\max}$	50ms
DUE Message Size, B_{d_j}	15kB

TABLE III. CHANNEL MODEL FOR LINKS [28]-[30]

Parameter	In-factory DUE link	UE-UE link	BS-UE link
Pathloss model	$36.8\log_{10}(d[\text{m}] + 35.8$	$40\log_{10}(d[\text{m}]) + 28$	$37.6\log_{10}(d[\text{m}]) + 15.3$
Shadowing	4dB	6dB	8dB
Fast fading	Rayleigh	Rayleigh	Rayleigh

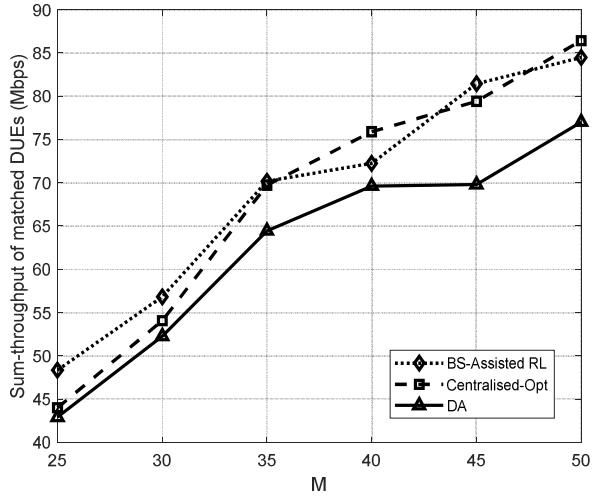

Figure 2. Sum-rate of matched DUEs with varying number of DUEs, M in the System, for $N = 50$

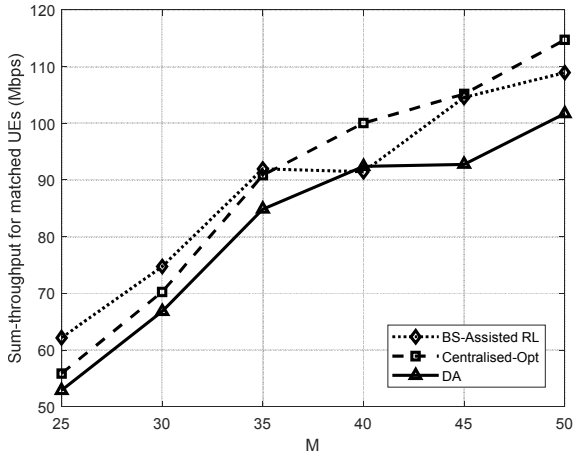

Figure 3. Sum Throughput of matched UEs as a function of the number of DUEs M, in the system, for $N = 50$

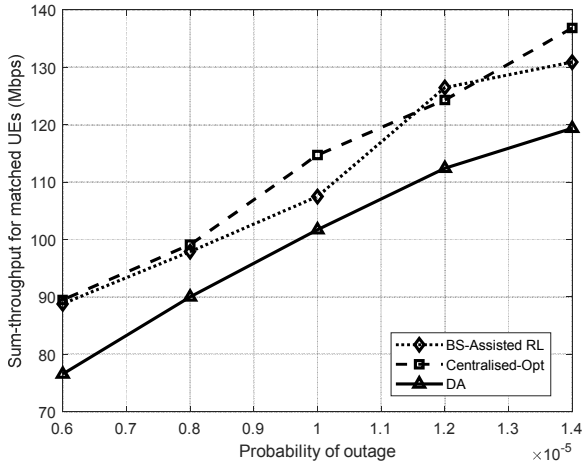

Figure 4. Effect of the DUE outage ratio p_{R_0}, on the sum throughput of matched CUE-DUE pair for $N = M = 50$, $l_{d_j,\max} = 50$ms

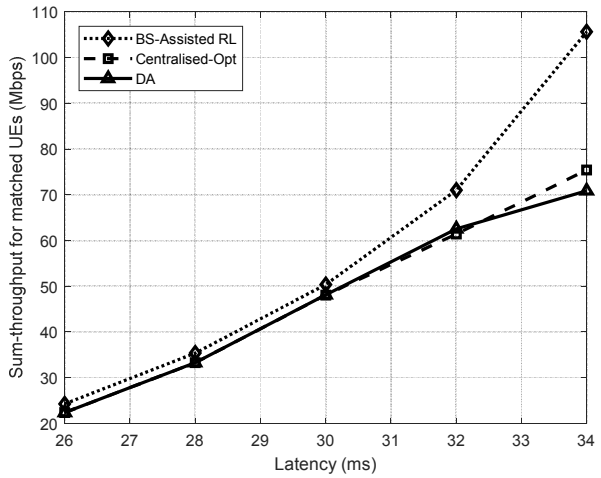

Figure 5. Effect of the delay bound, $l_{d_j,\max}$ on the sum throughput of matched CUE-DUE pair for $N = M = 50$, $p_{R_0} = 10^{-5}$

B. Signalling Overheads and Complexity Analysis

We now evaluate and compare the signalling overheads and computation complexity of the investigated algorithms. Signalling overheads are evaluated in terms of the level of involvement of the BS and User Equipment (UE), i.e., BS-UE communication. The signalling overhead evaluated is an aggregation of contributions of channel information acquisition and information exchange by the BS-UE links. The number of iterations T depends on the network dynamics. A summary of the signalling overhead estimation is presented in Table IV. The different approaches are also evaluated in terms of their computation complexity. The run time for the algorithm also depends on the number of iterations and on the number of users. It can be concluded that the centralised algorithm has the highest complexity, while the DA scheme has the least complexity, with a 10.38% reduction in processing time compared with the centralised approach for the studied scenario.

An overall comparison for the studied techniques based on throughput, signalling and complexity metrics is shown in Fig. 6 for different numbers of users. It can be seen that the centralised approach has the best throughput performance, however it has higher signalling overheads and computation complexity in comparison to the other approaches. DA has the lowest throughput performance and complexity, while BSA achieves the lowest signalling overheads. BSA achieves a 49.81% reduction in signalling overheads and 0.94% reduction in complexity with less than 9% lower throughput performance compared to the centralised approach which, is a good tradeoff of throughput and signalling overheads.

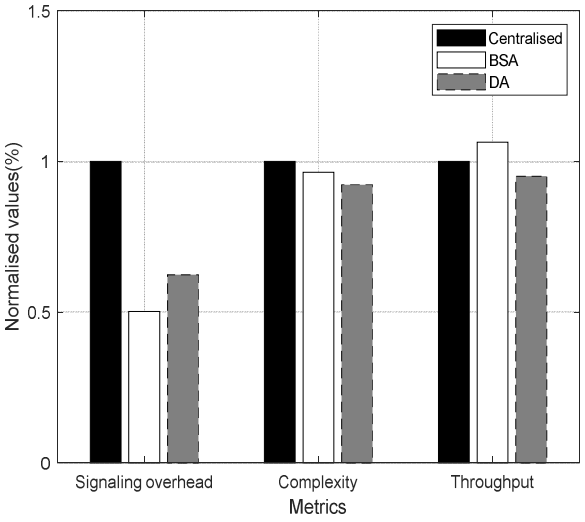

Figure 6a. Use-case 1: $M = 30, N = 50$

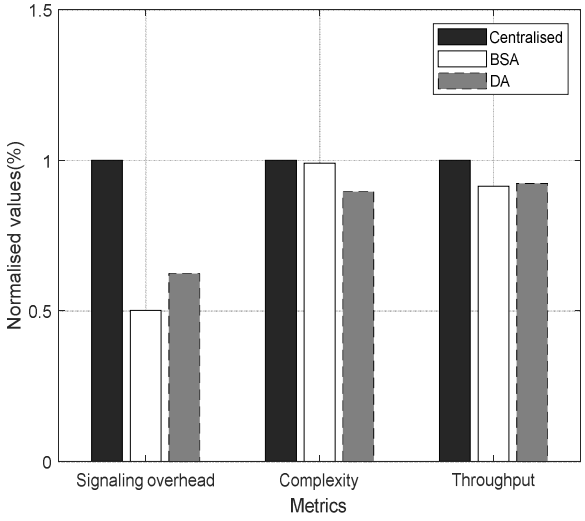

Figure 6b. Use-case 2: $M = 40,\ N = 50$

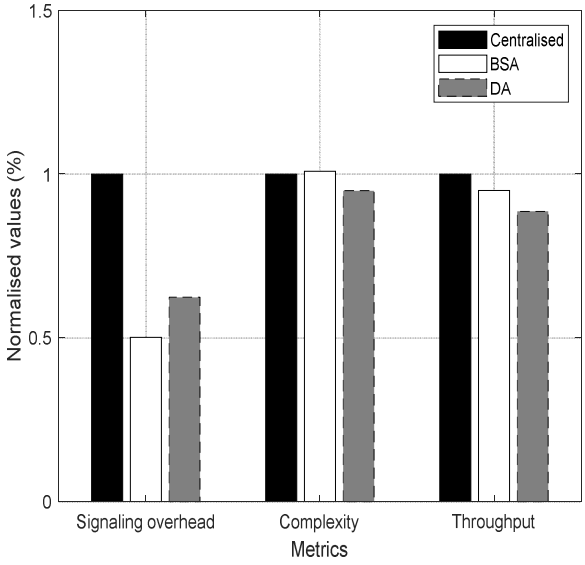

Figure 6c. Use-case 3: $M = N = 50$

Figure 6. Overall performance comparison with the centralised approach as a reference

TABLE IV. SIGNALLING OVERHEAD ESTIMATION

Estimation of the Signalling Overhead by the BS	
Centralised	M(1+4T)
DA	2M(N+T)
BS-A	2M(1+T)

In summary, the results indicate that for throughput maximisation in a low-density network in which, self-organisation is not important, the centralised scheme is the best to adopt at the cost of signalling overheads. The DA is a promising technique to achieve good throughput performance at lower signalling overheads and complexity if device autonomy and network stability are essential. On the other hand, BSA is a semi-distributive approach which, offers a good trade-off of throughput, complexity and signalling overheads trade-off compared to DA and centralised optimisation schemes.

Regardless of the limitations of the investigated and developed RRM techniques presented in this paper, the results from adopting these methodologies, suggest the possibility of developing a conceptual qualitative evaluation framework to assist in the selection of an appropriate scheme to achieve specific priorities for the target industrial scenarios, as presented in Table V.

TABLE V. QUALITATIVE COMPARISON OF THE DIFFERENT METHODOLOGIES

Scheme	BSA	Centralised Optimisation	DA
RRM Approach	Semi distributed	Centralised	Distributed
RRM Technique	Reinforcement learning	Mathematical optimisation	Matching theory
Throughput	Average	Best	Worst
Complexity	Average	Worst	Best
Signalling Overheads	Best	Worst	Average

V. CONCLUSIONS

We presented a semi-distributed BSA scheme for RRM of a D2D enabled cellular network targeting wireless industrial applications. The BSA scheme is an RL based approach which relies on distributed training of the D2D agents. Subsequently, the look-up tables for the D2D agents are loaded to the BS for centralised channel allocation.

The performance of the BSA scheme was compared with centralised optimisation and the game theoretic DA approaches in terms of throughput, signalling overheads and computation complexity. It is concluded that BSA offers a good trade-off of throughput, complexity and signalling overheads compared to DA and the centralised optimisation schemes. However, the BSA scheme is semi distributed. The future work aims at exploring optimised fully distributed techniques with the aim of facilitating an increased DUE autonomy through the combination of game theory and machine learning techniques.

REFERENCES

[1] I. O. Sanusi and K. M. Nasr, "A Machine Learning Approach for Resource Allocation in Wireless Industrial Environments," in Proc. of the Eighteenth Advanced International Conference on Telecommunications (AICT), pp. 18-23, Jun. 2022.

[2] J. Kaur, M. Arif Khan, M. Iftikhar, M, Imran and Q. E. Ul Haq, "Machine learning techniques for 5G and beyond networks," IEEE Access, Vol. 9, pp. 23742-23488, Jan. 2021.

[3] F. Tariq, M. R. Khandaker, K. K. Wong, M. A. Imran, M. Bennis and M. Debbah. "A speculative study on 6G," IEEE Wireless Communications, Vol. 27, no. 4, pp. 118-125, Aug. 2020.

[4] Y. Gu, W. Saad, M. Bennis, M. Debbah and Z. Han, "Matching theory for future wireless networks: fundamentals and applications," IEEE Communication Magazine, Vol. 53, no. 5, pp. 52-59, May 2015.

[5] B. Tian, L. Wang, Y. Ai and A. Fei, "Reinforcement learning based matching for computation offloading in D2D communications," in Proc. of 2019 IEEE/CIC International Conference on Communications in China (ICCC), pp. 984-988, Aug. 2019.

[6] D. L. Van and C. K. Tham, "A deep reinforcement learning based offloading scheme in ad-hoc mobile clouds," in Proc. of IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 760-765, Apr. 2018.

[7] H. Ye and Y. L. Geoffrey, "Deep reinforcement learning for resource allocation in V2V communications," in Proc. of IEEE International Conference on Communications (ICC), pp. 1-6, May 2018.

[8] B. Fernandez-Gauna, I. Etxeberria-Agiriano and M. Graña, "Learning multirobot hose transportation and deployment by distributed round-robin Q-learning," PloS one, Vol. 10, no. 7, Jul. 2015.

[9] K.K. Nguyen, T.Q. Duong, N.A. Vien, N.A. Le-Khac and M.N. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," IEEE Access, Vol. 7, pp. 100480-100490, Jul. 2019.

[10] S. De Bast, R. Torrea-Duran, A. Chiumento, S. Pollin and H. Gacanin, "Deep reinforcement learning for dynamic network slicing in IEEE 802.11 networks," in Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 264-269, Apr. 2019.

[11] H. Wu, X. Li and Y. Deng, "Deep learning-driven wireless communication for edge-cloud computing: opportunities and challenges," Springer, Journal of Cloud Computing, Vol. 9, no. 21, Dec. 2020.

[12] A. Asheralieva and Y. Miyanaga, "An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in cellular networks," IEEE transactions on communications, Vol. 64, no. 9, pp. 3996-4012, Jul. 2016.

[13] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," IEEE Transactions on Vehicular Technology, Vol. 69, no. 2, pp. 1828-1840, Feb. 2020

[14] S. Nie, Z. Fan, M. Zhao, X. Gu and L. Zhang, "Q-learning based power control algorithm for D2D communication," in Proc. of IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1-6, Sep. 2016.

[15] K. Zia, N. Javed, M. N. Sial, S. Ahmed, A. A. Pirzada and F. Pervez, "A distributed multi-agent RL-based autonomous spectrum allocation scheme in D2D enabled multi-tier HetNets," IEEE Access, Vol. 15, no. 7, pp. 6733-6745, Jan. 2019.

[16] Y. F. Huang, T. H. Tan, Y. L. Li and S.C. Huang, "Performance of resource allocation for D2D communications in Q-Learning based heterogeneous networks," in Proc. of 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), pp. 1-5, May 2019.

[17] I. Budhiraja, N. Kumar and S. Tyagi, "Deep-Reinforcement-Learning-Based Proportional Fair Scheduling Control Scheme for Underlay D2D Communication, " IEEE Internet of Things Journal, Vol. 8, no. 5, pp. 3143-3156, Mar. 2021.

[18] S. Nie, Z. Fan, M. Zhao, X. Gu and L. Zhang, "Q-learning based power control algorithm for D2D communication," in Proc. of IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1-5, Sep. 2016.

[19] L. Liang, H. Ye and G.Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," IEEE Journal on Selected Areas in Communications, Vol. 37, no. 10, pp. 2282- 2292, Aug. 2019.

[20] I. O. Sanusi, K. M. Nasr and K. Moessner, "Radio resource management approaches for reliable Device-to-Device (D2D) communication in wireless industrial applications," IEEE Transactions of Cognitive Communication and Networking, Vol. 7, no. 3, pp. 905-916, Oct. 2021.

[21] A. T. Kasgari and W. Saad, "Model-free ultra-reliable low delay communication (URLLC): A deep reinforcement learning framework," in Proc. of IEEE International Conference on Communications (ICC), pp. 1-6, May 2019.

[22] H. Wang and X. Chu, "Distance-constrained resource-sharing criteria for device-to-device communications underlaying cellular networks," Electronics letters, Vol. 48, no. 9, pp. 528-530, Apr. 2012.

[23] H. Yang, X. Xie and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-delay IoV communication networks," IEEE Transactions on Vehicular Technology, Vol. 68, no. 5, pp. 4157-4169, Jan. 2019.

[24] F. E. Souhir, A. Belghith and F. Zarai, "A reinforcement learning-based radio resource management algorithm for D2D-based V2V communication," in Proc. of 15th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 1367-1372, Jun. 24, 2019.

[25] Y. Wei, Y. Qu, M. Zhao, L. Zhang and F.R. Yu, "Resource allocation and power control policy for Device-to-Device communication using multi-Agent reinforcement learning," Computers, Materials & Continua, Vol. 63, no. 3, pp. 1515-1532, May 2020.

[26] J. Kim, J. Park, J. Noh and S. Cho, "Autonomous power allocation based on distributed deep learning for device-to-device communication underlaying cellular network," IEEE Access, Vol. 8, pp. 107853-107864, Jun. 2020.

[27] G. Brown, "Ultra-Reliable Low-Latency 5G for Industrial Automation", Qualcomm white paper, 2018

[28] WINNER II Channel Models, Standard IST-4-027756 WINNER II D1.1.2 V1.2, Sep. 2007.

[29] H. Xing and S. Hakola, "The investigation of power control schemes for a device-to-device communication integrated into ofdma cellular system," in Proc. of IEEE Personal Indoor and Mobile Radio Communication (PIMRC), pp. 1775-1780, Sep. 2010.

[30] Evolved Universal Terrestrial radio Access (E-UTRA), "Further Advancements for E-UTRA Physical Layer Aspects (Release 9)," 3GPP TR 36.814, Tech. Rep., 2010.