Deep Learning-based Algorithm for Optimizing Relay User Equipment Activation in 5G Cellular Networks

Juan Jesús Hernández-Carlón, Jordi Pérez-Romero, Oriol Sallent, Irene Vilà and Ferran Casadevall

***Abstract*— This paper addresses the problem of optimally using the relay capabilities of user equipment (UE) to augment the radio access network (RAN) in 5G deployments and beyond. This can be particularly useful in coverage constrained scenarios, such as those using millimeter waves, due to the difficulty radio signals penetrate some structures. This can lead to signal blockages and high penetration losses when providing outdoor-to-indoor coverage. To overcome these limitations, the use of relay UEs (RUEs) is seen as a possible solution to effectively extend the coverage of a cellular network. In this context, this paper proposes a deep learning-based algorithm to optimize the decision regarding when RUEs should be activated and deactivated in accordance with the benefits they can provide for increasing the spectral efficiency and decreasing outage probability for the network users. The obtained results reveal a promising capability of the proposed solution to activate the most beneficial RUEs given the network conditions being experienced, leading to improvements of average spectral efficiency of 12.3% and reductions of outage probability of 89% with respect to the case without relays.**

Index Terms*— *Radio Access Network, Beyond 5G, Deep-Q Network, Deep Learning, User Equipment, UE-to-network Relaying

I. INTRODUCTION

OVER the past years, there has been continued and substantial growth in the traffic generated by users of mobile networks. According to the Ericsson Mobility Report [1], the average usage per smartphone was expected to surpass 15 GB in 2022, whereas the 5G-associated mobile traffic was expected to grow by up to 60% by 2027. Additionally, the report states that the traffic associated with video services currently represents 69% and is estimated to increase to 79% by 2027. Overall, all mobile traffic is projected to grow in the coming years. In this context, mobile network operators (MNOs) need to scale the deployed capacity in their radio access network (RAN) infrastructure to face the enormous demand for traffic to come. This implies a large capital expenditure (CAPEX) on network improvements (e.g., deploying 5G RAN infrastructure). These investments can be particularly important when considering the operation of 5G at high frequencies, such as millimeter waves (mmWaves), which are much more sensitive to signal blockages due to obstacles. They introduce higher penetration losses when providing outdoor-to-indoor coverage, thus requiring denser base station deployments.

In line with the increasing demands of traffic and new services in mobile networks over the years, there has been a tremendous technological evolution not only at the network infrastructure side but also at the user equipment (UE) side, leading to the availability of UEs with powerful communication and computational capabilities. Following these trends, in our previous work [2], we described a beyond 5G (B5G) scenario where the UE actively cooperates in the provision of network services, e.g., by relaying traffic from other UEs toward the network. The obtained results revealed that the use of relay UEs (RUEs) can be beneficial for MNOs thanks to a substantial reduction in the number of base stations to be deployed and the consequent reduction in capital expenditures.

Although the idea of using relay stations for coverage and capacity extension has already been studied in the literature (see, e.g.,[3]), its practical implementation in previous technologies such as 4G has been quite limited and focused on very specific situations, such as using fixed relays for extending coverage in a tunnel. However, the idea of using relays has recently gained momentum due to the challenges in providing coverage and capacity in certain scenarios, such as in-home residential environments with mmWave frequencies, smart factories or even public safety applications. Accordingly, the Third Generation Partnership Project (3GPP) has recently introduced a new relaying technology, referred to as Integrated Access and Backhaul (IAB), which makes use of 5G New Radio (NR) technology for supporting the backhaul of a base station, thus offering an alternative to fiber backhaul [4][5].

Manuscript submitted January 10, 2023. Revised March 3, 2023, June 21, 2023 and September 4, 2023. This work was supported in part by the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreements No. 101096034 (VERGE project) and No. 101097083 (BeGREEN project) and in part by the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033 under ARTIST project (ref. PID2020-115104RB-I00). The work of J.J. Hernández-Carlón was supported by the Spanish Ministry of Science and Innovation under grant ref. PRE2018-084691. The work of I. Vilà was supported by the European Union-NextGenerationEU, Spanish Ministry of Universities and the Plan for Recovery, Transformation and Resilience, through the call for Margarita Salas Grants of the Universitat Politècnica de Catalunya (ref. 2022UPC-MSC- 94079). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the other granting authorities. Neither the European Union nor the granting authority can be held responsible for them.

The authors are with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain (e-mail: juan.jesus.hernandez@upc.edu; jordi.perez-romero@upc.edu; sallent@tsc.upc.edu; irene.vila.munoz@upc.edu; ferranc@tsc.upc.edu).

Similarly, the use of vehicle-mounted relays that consider moving vehicles with onboard relay base stations is also a subject of a recent work item in 3GPP Release 18, whose outcomes are reported in [6]. The UE-to-network relaying feature, in which a UE relays the traffic of another UE to/from the network in a two-hop link, has also been the subject of a 3GPP study item [7] and was recently incorporated as one of the connectivity models of [8], which discusses different applicability scenarios of relay UE and defines requirements and key performance indicators. The possible applications of relay UE in [8] cover a wide variety of scenarios, such as in homes, smart farming, smart factories and even public safety applications. In addition to the work in standardization, the interest in UE-to-network relaying is also reflected by different works, such as the recent survey [9] and references therein.

To successfully realize the UE-to-network relaying concept, it is necessary to design and develop new functions in B5G systems. On the one hand, these cover top-level service layer capabilities so that mobile network operators and UE holders can interact to establish the conditions when the UE can be integrated as part of the RAN, including service level agreements at both the business and technical levels. On the other hand, adequate management and control layer functionalities need to be developed to leverage the connectivity supplied by the RUEs. In this context, a critical functionality is "RUE activation", which is the focus of this paper. This functionality is in charge of deciding under which conditions a UE is eligible to be activated to act as a relay and be incorporated as an additional element of the RAN.

The RUE activation functionality was studied in [10] in relation to the type of context information to be considered by this problem. The authors in [10] analyzed seven RUE activation strategies that differ in the criteria and context information used and found that the most effective strategies for reducing outage probability were those that considered the number of UEs that an RUE could serve based on the knowledge of the spectral efficiency of these UEs. Leveraging the outcomes of this previous work, a functional framework for supporting RUE activation was presented in [11] based on the characterization of each potential RUE through a utility metric that measures the coverage enhancements brought to the network when the RUE is activated.

The main contribution of this paper is a new RUE activation strategy that makes use of deep reinforcement learning (DRL), and more specifically of the deep Q-network (DQN) technique, to optimally activate UEs as relays to enhance the coverage conditions of the deployed base station. The use of DQN is particularly convenient because it is able to learn decision-making policies for problems with high-dimensional state and action spaces and it is able to progressively update the learned policy based on the accumulated experience during a training process. Then, with the proposed approach, each base station is associated with a DQN agent that learns the RUE activation policy depending on the potential benefit brought by the RUEs to be activated in accordance with the network dynamics. Overall, the proposed solution aims at improving the spectral efficiency and reducing the outage in the scenario, but in a way that RUEs are only activated when they are beneficial for the network, thereby reducing the time that RUEs remain active.

To the best of our knowledge, no previous works in the literature have addressed the RUE activation problem by means of DRL. Only a few works have considered machine learning (ML) methods in the context of relay-assisted networks, but they address different problems. For example, in [12], a deep-learning model is used for the user-to-relay association problem, in charge of predicting the best serving relay for a UE. The algorithm made decisions in accordance with the distances between the UE, the relay and the base station. Similarly, the authors of [13] considered the problem of selecting the set of relays that enable a data packet to travel from a start node to an end node. The problem was modeled as a Markov decision process with actions associated with specific nodes, and a Q-Learning algorithm was trained to determine the best communication path. The rest of the paper is organized as follows. Section II presents the system model and formulates the considered relay-activation problem. Section III presents the proposed DQN-based solution. Then, Section IV provides the performance assessment of the proposed solution in terms of different indicators and the comparison against different benchmark solutions, including the exhaustive search strategy, genetic algorithm and a reference from the state-of-the-art. Finally, Section V summarizes the conclusions and discusses future work.

II. Problem Definition and Proposed Architecture

Let us consider a scenario with a 5G network infrastructure deployed by a mobile network operator, as depicted in the lower part of Fig. 1. It is composed of the RAN that includes the base stations (BS), the core network (CN) and the service management and orchestration (SMO) system that enables the configuration and performance management (PM) of the infrastructure. To address coverage-limited situations due to, e.g., high penetration losses or obstructions when using mmWave bands, the system has the possibility of activating some UEs to act as relays for other UEs. In this way, a UE can reach the core network through a direct link with the BS or by means of an RUE, as shown in Fig. 1.

The SMO includes an RUE activation management (RAM) function that determines when, where and under what conditions a UE served by a BS is suitable to be activated as an RUE to solve coverage problems. To this end, acquiring knowledge about the UEs in the area of a BS and their behavioral patterns will be relevant. For example, let us suppose a situation in which a UE is located inside an office building during working hours. Then it is highly likely that it remains stationary and is connected to its serving BS for long periods of time. If the signal quality of this UE is sufficiently good and its battery level is above a certain threshold, this UE can be considered to be a potential candidate relay to be activated. The decision on whether to activate a candidate RUE needs to trade-off different aspects, such as the global benefit of activating the RUE in accordance with the performance

experienced by the UEs connected to it and the cost of activating the RUE, e.g., in terms of energy consumption. Based on this, the problem considered by this paper is the optimization of the RUE activation decisions to ensure that RUEs are only activated when the network dynamics require them.

The RUE activation decision-making for a given BS is executed at the RUE activation controller (RAC) of the RAM function at the SMO, as depicted in the functional architectural framework of Fig. 1. It is supported by the candidate RUE database that contains the list of UEs capable of acting as RUEs in the BS. These are the UEs that stay within the BS for a sufficiently long period of time and whose owners have reached the appropriate agreements with the MNO so that they are authorized to act as RUEs. How the list is built is out of the scope of this paper, based on the assumption that the identification of "home UE" associated with BSs would not be difficult to achieve (e.g., by observing that a UE is always served by the same BS at certain times, hours, etc.). Similarly, the business model-related mechanisms for engaging UE owners to let their terminals act as relays are also out of the scope of this paper but could be based, e.g., on giving incentives to UE owners [8].

The candidate RUE database includes different fields such as the identifier of the relay, the position and the current status mode that specifies if the RUE is active or not at a certain time. These data, together with specific performance measurements collected from the network, constitute the inputs to the RAC so that it can decide to activate the RUEs at a certain time. To formalize the problem, let us first model the performance experienced by UEs. For this purpose, let us suppose a UE is located in the coverage area of a given BS. If the UE is directly connected with the BS, the spectral efficiency S_D can be obtained by using the Shannon formula:

$$S_D = min\left(S_{max}, log_2(1+SINR_{BS\text{-}UE})\right) \quad (1)$$

where S_{max} represents the maximum possible spectral efficiency associated with the maximum modulation and coding scheme (MSC) of the 5G NR [14] and $SINR_{BS\text{-}UE}$ is the signal-to-interference and noise ratio (SINR) in the link between the UE and the BS. In turn, when the UE is connected to the BS via an activated RUE, the spectral efficiency is bounded by the link with the worst conditions between both the BS-RUE and RUE-UE links and it is expressed as:

$$S_R = min\left(S_{max}, log_2(1+min(SINR_{BS\text{-}RUE}, SINR_{RUE\text{-}UE}))\right) \quad (2)$$

where $SINR_{BS\text{-}RUE}$ and $SINR_{RUE\text{-}UE}$ denote the SINR in the BS-RUE and RUE-UE links, respectively.

It is assumed that a UE is in outage if it is experiencing a spectral efficiency lower than a given threshold S_{min} that establishes the minimum requirement for proper service provisioning. It is also assumed that a UE will only try to connect to an activated RUE if it is in outage with its serving BS (i.e., $S_D<S_{min}$). In this case, the UE will try to connect to the activated RUE, providing the highest spectral efficiency S_R.

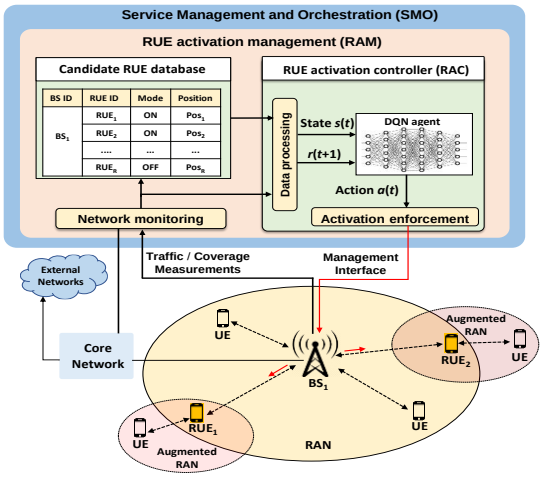

Fig. 1. Architectural components of the considered approach.

It is also assumed that RUEs with spectral efficiency less than S_{min} are not available for activation.

Focusing on the RUE activation problem, let us denote a given BS in the scenario as b. There are a total of R candidate RUEs associated with this BS b in the database. The candidate RUEs are numbered r=1,...,R. The r-th candidate RUE has a status mode denoted $a_{b,r}$ ={0,1}, where 0 means that the RUE is deactivated and 1 means it is activated. Therefore, the global status mode configuration associated with BS b can be defined as the R-length vector $\mathbf{C_b}=\{a_{b,r}\}$.

The objective of the considered problem is to find a policy that optimally activates the RUEs. This means finding the optimum configuration $\mathbf{C_b}(t)=\{a_{b,r}(t)\}$ to be applied at every time t when the RAC decides on the activation or deactivation of the existing RUEs. It is assumed that these decisions are made in discrete time instants t with a granularity of ΔT. These discrete times are denoted as t, t+1,..., t+k,...

The criterion to consider a configuration $\mathbf{C_b}(t)$ as optimum is based on the efficiency of each candidate RUE when activated or deactivated. If the r-th RUE is active, i.e., $a_{b,r}(t)$=1, the efficiency $E_{b,r}(a_{b,r}(t))$ accounts for the average number of UEs in outage $N_{b,r}$ served by this RUE until the next decision period. Specifically, $E_{b,r}$ $(a_{b,r}(t))$=1 if $N_{b,r} \geq x$, where x is a certain threshold, meaning that the activation is worthwhile as the RUE has served a significant number of UE. Instead, the efficiency will be 0 if the RUE has served less than x UEs on average. On the other hand, if the RUE is inactive, i.e., $a_{b,r}(t)$=0, the efficiency is computed based on $P_{b,r,}$ which is defined as the average number of UEs that would have been served by the RUE if it had been active during the period [t, t+1]. Then, the efficiency $E_{b,r}$ $(a_{b,r}(t))$ will be equal to 1 if $P_{b,r} < x$, meaning that in this case, it is efficient not to have the RUE active. Similarly, it will be $E_{b,r}$ $(a_{b,r}(t))$ =0 if $P_{b,r} \geq x$, meaning that in this case, keeping the RUE inactive is not efficient, as it could serve a significant number of UEs. In view of the above, the formal problem to solve is to find at every time t the configuration

$\mathbf{C_b}(t)$ that maximizes the aggregate global efficiency AGE, which is defined as follows:

$$AGE=\frac{1}{R}\left[\sum_{r=1}^{R} E_{b,r}(a_{b,r}(t))\right] \quad (3)$$

TABLE I
LIST OF ACRONYMS AND NOTATION

Acronyms	
5G NR	5G new radio.
B5G	Beyond 5G.
BS	Base Station.
CDF	Cumulative Distribution Function.
CN	Core Network.
DNN	Deep Neural Network.
DQN	Deep Q-Network.
DRL	Deep Reinforcement Learning.
KPI	Key Performance Indicator
MNO	Mobile Network Operator.
RAC	RUE Activation Controller.
RAM	RUE Activation Management.
RAN	Radio Access Network.
RL	Reinforcement Learning.
RUE	Relay User Equipment.
SINR	Signal-to-Interference and Noise Ratio.
SMO	Service Management and Orchestration.
UE	User Equipment.
Notation	
$\mathcal{A}$	Action space containing all eligible actions.
AGE	Aggregate global efficiency.
$a_{b,r}$	Status mode of RUE r in BS b.
$\boldsymbol{a}(t)$	Action selected at time t.
$\mathbf{C_b}$	Configuration associated with BS b.
D	Experience dataset associated with the agent.
E	A given experience of the agent.
$E_{b,r}$	Efficiency of the activation of relay r.
F	Efficiency of RUEs in active mode.
G	Number of genes of each individual.
$K(t_n)$	Total number of activated RUEs given an action $a(t_n)$.
$L(\theta)$	Average mean squared error.
M_T	RUE time in active mode.
N	Total number of actions taken during policy evaluation.
$N_b(t)$	Average number of UEs in the outage served by all RUE at time t.
$N_{b,r}(t)$	Number of UEs in outage that have been served by RUE r.
N_{gen}	Number of generations.
N_p	Number of individuals.
$N_{samp}(t_n)$	Total number of samples taken at time t_n.
O	Outage probability.
$o(i,t_n)$	Sample i of UE outage at time t_n.
$P_{b,r}$	Number of UEs served by the RUE if it were active.
P_{mut}	Mutation probability.
$Q(\boldsymbol{s},\boldsymbol{a},\theta)$	Q-network associated with DQN agent.
$r(t+1)$	Reward obtained at time t+1.
R	Total number of candidate RUEs.
R_w	Average reward.
S	Average spectral efficiency.
$S_{b,r}(t)$	Spectral efficiency of RUE r in BS b.
$\boldsymbol{s}(t)$	State of a given base station at time t.
S_D	Spectral efficiency in the link between a UE and a BS.
$S_{eff}(t)$	Spectral efficiency of all RUEs in BS b.
$s_{eff}(i,t_n)$	Sample i of UE spectral efficiency at time t_n.
S_{max}	Maximum possible spectral efficiency.
S_R	Spectral efficiency of a UE connected to a BS through a RUE.
t_n	Time at which action number n is made.
$U(D)$	Mini-batch of experiences.
X	Minimum UEs to be served by a RUE.
ΔT	Duration of a time step.
π	Policy learnt by the agent.

III. DQN-BASED SOLUTION

The development of an efficient solution to the RUE activation problem involves a multiplicity of variables, such as the current status mode of the RUEs, the propagation conditions of the RUEs and the nearby UEs, and the traffic dynamics. To address these multiple dimensions, this paper proposes the use of DRL.

DRL techniques combine the use of deep neural networks (DNNs) with reinforcement learning (RL) to assist a software-based agent that makes decisions in relation to a specific problem. This combination is especially interesting because of its capability to handle large state and action spaces. DRL techniques have been applied in many different fields, such as robotics, video processing, and gaming, demonstrating outstanding success, as noted in [15].

Among the different DRL techniques, this work specifically proposes a solution based on the DQN algorithm [16]. In this approach, the learning process is carried out dynamically by a DQN agent that interacts with an environment, and after observing the consequences of its actions measured in terms of a certain reward signal, it learns to modify its own decision-making behavior.

The DQN algorithm has been selected to address the RUE activation problem mainly for two reasons. The first is that the DQN algorithm has been designed to support high-dimensional states and action spaces. This is convenient for the RUE activation problem since the network dynamics implies a large amount of data that needs to be considered by the agent. The second reason is that with DQN, the policy is progressively updated by considering individual samples of experience as opposed to other methods such as Monte Carlo simulations [15] that update the policy by considering multiple samples collected during an episode. This feature is suitable for the case of the RUE-activation problem since continuous learning of the policy is desired. Moreover, DQN is a useful technique for learning how to select actions from discrete action spaces, as in the problem considered here where the actions involve activations or deactivations of RUEs.

A variety of works have approached different RAN-related problems by means of the DQN technique. For example, DQN was used for capacity sharing in [17], while in [18], it was used to address the resource allocation problem in heterogeneous networks. Similarly, in [19], DQN was applied to the multiconnectivity problem, and in [20], it was used for spectrum sharing.

In the proposed solution, the DQN agent is located at the RAC (see Fig. 1) that makes decisions for the RUEs associated with the BSs in the scenario. At time t, the DQN agent of BS b selects an action $\boldsymbol{a(t)}$ that contains the RUE activation configuration $\mathbf{C_b}(t)$ to be applied to the set of RUEs in the next time window of duration ΔT. The selection of a given action is dependent on the state observed at time t denoted as $\boldsymbol{s(t)}$ together with the available policy π at that time. The state is obtained by processing the data from the network monitoring module and the candidate RUE database at the data processing

module located in the RAC. This processing is needed to adapt the information to the format required by the DQN agent.

The outcome of applying a certain action that defines a RUE activation configuration is assessed by means of a reward signal $r(t+1)$. This reward is delivered to the DQN agent at the end of the time window ΔT. It essentially measures how effective or ineffective the selected action was. The reward signal obtained over time after selecting the different actions is utilized to progressively enhance the DQN-agent decision-making policy. The main components of the proposed DQN-based solution along with the policy learning process are described below.

A. State, action and reward definition

The state $\boldsymbol{s}(t)$ is represented as a vector associated with a particular BS b, and it has different components listed in the following:

- $S_{eff}(t)=\{S_{b,1}(t), S_{b,2}(t),\ldots,S_{b,R}(t)\}$ represents the spectral efficiency of the RUEs in BS b, computed according to (1).
- $C_b(t)=\{a_{b,1}(t),a_{b,2}(t),\ldots,a_{b,R}(t)\}$ denotes the configuration of all RUEs at time t.
- $N_b(t)=\{N_{b,1}(t), N_{b,2}(t),\ldots, N_{b,R}(t)\}$ corresponds to the average number of UEs in the outage that have been served by each RUE measured at time t.

The total number of components in the state is $3\cdot R$.

A given action $\boldsymbol{a}(t) \in \mathcal{A}$ can be seen as a vector $C_b(t)=\{a_{b,r}(t)\}$ that contains the RUE activation configuration applied every time window accounting for all the considered RUEs. The so-called action space $\mathcal{A}$ contains all eligible RUE activation configurations. Since an RUE has only 2 possible modes (activated and deactivated), the total number of possible actions in the action space is 2^R.

The reward signal $r(t+1)$ that assesses the efficiency of the action $\boldsymbol{a}(t)$ is selected for state $\boldsymbol{s}(t)$ in relation to the optimization criterion (3). Thus, the reward can be expressed as the obtained value of *AGE*:

$$r(t+1)= \frac{1}{R}\left[\sum_{r=1}^{R} E_{b,r}(a_{b,r}(t))\right] \quad (4)$$

B. Policy learning process

A stage of major relevance when applying the DQN technique is the training of the DQN agent. By means of the training, the agent actively learns a decision-making policy π that is used for selecting which action to apply under a particular state. The training process considered in this work is based on the algorithm presented in [16] but customizes the DQN agent according to the previously defined state, action and reward. The overall process is summarized in the following. It is worth mentioning that the same algorithmic procedure has also been successfully applied by the authors for addressing the capacity sharing and the multiconnectivity problems in [17] and [19], respectively.

The fundamental objective of RL-based algorithms is to determine the optimal policy π^* that maximizes the so-called discounted cumulative future reward defined as $\sum_{j=0}^{\infty}\tau^k r(t+j+1)$, where τ represents the discount factor that takes values between 0 and 1. In the case of the DQN algorithm, the optimal policy results from determining the optimal action-value function denoted as $Q^*(\boldsymbol{s},\boldsymbol{a})$. This function represents the maximum expected discounted cumulative reward that can be obtained by applying an action $\boldsymbol{a}$ for a particular state $\boldsymbol{s}$ starting at a certain time t and following the policy π. This can be expressed recursively by means of the Bellman equation as:

$$Q^*(\boldsymbol{s},\boldsymbol{a}) = E[r(t+1)+ \tau\cdot max_{a'} Q^*(s(t+1),a') \mid s(t)=s,a(t)=a,\pi] \quad (5)$$

Based on this definition, the optimum policy π^* is the one that selects the action that maximizes the action-value function, that is:

$$\pi^* = arg\ max_a\ Q^*(\boldsymbol{s},\boldsymbol{a}) \quad (6)$$

The DQN algorithm makes use of a DNN to approximate the optimum action-value function $Q^*(\boldsymbol{s},\boldsymbol{a})$. In particular, the DNN takes as input each one of the components of state $\boldsymbol{s}$ and provides an output $Q(\boldsymbol{s},\boldsymbol{a},\theta)$ that represents the approximation of the optimum action-value function for each one of the eligible actions. The term θ denotes the weights of the different interconnections between neurons in the DNN. In this respect, the structure of the DNN includes an input layer with a number of neurons equal to the number of components in state $3\cdot R$, an output layer with a number of neurons equal to the number of possible actions 2^R and one or more hidden layers. The number of hidden layers and the number of neurons in each layer are the hyperparameters of the DQN that are specified as part of the configuration.

The optimal action-value function can then be learned by iteratively updating the function $Q(\boldsymbol{s},\boldsymbol{a},\theta)$ during the training stage by varying the values of the weights θ in accordance with the experienced rewards. To update the weights, the DQN agent includes the following components:

- **Evaluation DNN $Q(\boldsymbol{s},\boldsymbol{a},\theta)$**: It is the DNN that approximates the optimum value function $Q^*(\boldsymbol{s},\boldsymbol{a})$. Based on this DNN, the policy π for deciding the different RUE activation actions to apply is given by:

$$\pi = arg\ max_a\ Q(\boldsymbol{s},\boldsymbol{a},\theta) \quad (7)$$

- **Target DNN $Q(\boldsymbol{s},\boldsymbol{a},\theta^-)$:** This is another neural network with the same structure in terms of the number of layers and neurons as the evaluation DNN but with weights θ^-. It is used to calculate the time difference (TD) target expressed as $r(t+1)+\tau \max_{\boldsymbol{a}'} Q(\boldsymbol{s}(t+1),\boldsymbol{a}',\theta^-)$, which allows updating the weights of the evaluation DNN while conducting the training process.
- **Experience dataset $\boldsymbol{D}$**: This dataset gathers the experiences obtained by the DQN agent during the training process. A given experience is expressed by means of a tuple $<\boldsymbol{s}(t), \boldsymbol{a}(t), r(t+1), \boldsymbol{s}(t+1)>$ composed of the state and action performed at time t along with the obtained reward and the new state at time $t+1$. The total length of the dataset is denoted as l.

When the training process starts, the weights of the target and the evaluation DNN are initialized randomly. Then, during the training process, these weights are progressively updated. Overall, the training stage involves two main parts, namely, data collection and the process of updating the weights.

The data collection is the process of filling experience dataset D with the gathered experiences. For this purpose, at each training step, the agent observes the state and chooses an action $\boldsymbol{a}(t)$ following an ε-greedy policy that selects the action based on the current policy (7) with probability 1-ε and a random action with probability ε. This random action selection is needed for incorporating in the training process the capability to explore new actions that are different from the ones that the current policy would select. After applying the selected action, the obtained reward is measured and placed in the experience tuple that is saved in the dataset.

Every time that the experience dataset reaches its storage capacity l, older experiences are removed and substituted by recent ones. Moreover, at the beginning of the training, the agent selects actions randomly (i.e., ε is set to 1) to gather a wide variety of experiences. This is maintained during a number of InitialCollectSteps training steps.

The update of the weights of the evaluation DNN is done at every training step by considering the experiences accumulated in the experience dataset. An updating process consists of making a random selection of a mini-batch $U(D)$ of past experiences J belonging to the dataset. These experiences are expressed as e_j, j=1,...,J, where e_j is an experience tuple denoted as $<\boldsymbol{s}_j,\boldsymbol{a}_j,r_j,\boldsymbol{s}_j*>$. Then, the update is performed by means of a mini-batch gradient descent procedure. To this end, the average mean squared error (MSE) for all the experiences in $U(D)$ is computed first as:

$$L(\theta)= E_{e_j \in U(D)} [(r_j+\tau \max_{a'} Q(s_j*,a',\theta^-) - Q(s_j,a_j,\theta))^2] \quad (8)$$

Then, the mini-batch gradient descent of $L(\theta)$ is computed by the derivative of $L(\theta)$ with respect to θ as follows:

$$\nabla L(\theta)=E_{e_j \in U(D)} [(r_j+\tau \max_{a'} Q(s_j*,a',\theta^-) - Q(s_j,a_j,\theta)) \cdot \nabla_\theta Q(s_j,a_j,\theta))] \quad (9)$$

The final step consists of updating the weights of the evaluation DNN $Q(\boldsymbol{s},\boldsymbol{a},\theta)$ as follows:

$$\theta \leftarrow \theta + \alpha \cdot \nabla L(\theta) \quad (10)$$

where α represents the learning rate.

Following each update of θ, the obtained $Q(\boldsymbol{s},\boldsymbol{a},\theta)$ will be used to select new actions. In relation to the weights θ^- of the target DNN, they are updated as $\theta^-=\theta$ after every P updates of the evaluation DNN.

The pseudocode that summarizes the abovementioned procedure is presented in Algorithm 1, which is based on [19] but has been adapted to outline the DQN-agent training procedure for the relay activation problem of BS b. The duration of the training procedure is given by the parameter MaxNumberOfTrainingSteps.

Algorithm 1. DQN training for BS *b*

```
Initialize DNN counter p=0.
For t=0… MaxNumberOfTrainingSteps
    Collect state s(t) (see section III.A)
    Generate a random number ε' between 0 and 1.
    If ε'<ε  (where ε=1 if t<=InitialCollectSteps)
        Select a random RUE activation configuration a(t).
    Else
        Get a RUE activation configuration a(t) based on π.
    End if
    Compute reward r(t+1) and s(t+1) as a function of action a(t).
    If D is full (l samples are stored)
        Delete the oldest experience.
    Store experience < s(t), a(t), r(t+1), s(t+1)> in D.
    Randomly sample a minibatch of experiences U(D) from D of length J.
    Compute the loss function L(θ).
    Compute the mini-batch gradient descent ∇L(θ).
    Update weights θ of evaluation DNN using (10).
    If p==P
        Update the weights of target DNN θ⁻=θ and set p=0.
    Else
        p=p+1
    End if
End for
```

IV. Performance Evaluation

This section presents the performance assessment of the proposed RUE activation strategy by performing different system-level simulations. Section IV. A describes the scenario used for the evaluation together with the algorithm parameters. Then, Section IV. B describes the considered key performance indicators (KPIs) for assessing the performance of the proposed model. Based on this, Section IV. C presents the evolution of the training process to obtain the RUE activation policy, and Section IV. D presents the obtained performance results by comparing the proposed approach against different benchmarking strategies.

A. Considered Scenario

The studied scenario is a 200 m × 200 m square area consisting of one 5G NR BS and four UEs able to act as relays. This is illustrated in Fig. 2. The key parameters of both BS and RUEs are shown in Table II. The traffic generation of the different UEs is based on a Poisson session arrival process with an average session generation rate of 0.6 sessions/s and an exponentially distributed session duration with an average of 120 s. A UE remains static for the entire duration of its session.

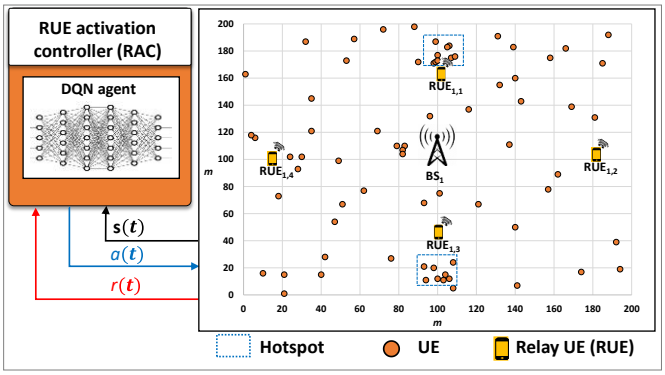

Fig. 2. Graphic representation of the scenario used for training and evaluation.

TABLE II
BS CONFIGURATION PARAMETERS

Parameter	Value	
Type of RAT	5G NR	Relay UE (RUE)
Position [x, y] m	[100,100]	[164,93] [100,180] [30,130] [100,20]
Frequency	26 GHz	3.5 GHz
Channel bandwidth	100 MHz	100 MHz
Transmitted power	21 dBm	21 dBm
Antenna gain	26 dB	3 dB
Height	10 m	1.5 m
UE antenna gain	10 dB	
UE noise figure	9 dB	
UE height	1.5 m	
Path loss model	Model of Sec 7.4 of [22]	UE-to-UE propagation model of [23]

TABLE III
DQN ALGORITHM CONFIGURATION PARAMETERS

Parameter	Value
Initial collect steps	500
MaxNumberOfTrainingSteps	200000
Experience Replay buffer maximum length (l)	$100 \cdot 10^3$
Mini-batch size (J)	64
Time window (ΔT)	30 sec
DNN updating period (P)	500 Training Steps
Discount factor (τ)	0.9
Learning rate (α)	0.0003
ε value (ε-Greedy)	0.1
DNN architecture	Input layer: 12 nodes Two hidden layers: 100 and 50 nodes Output layer: 16 nodes

The traffic spatial distribution assumes that 40% of the UEs are randomly located inside two square hotspots of 15 m × 15 m and 20 m × 20 m, located in the upper and lower center of the scenario area, respectively (see Fig. 2). The remaining UEs are randomly distributed in the whole scenario. The model assumes a connectivity model in which a given UE attempts to connect via an RUE just in the case of being in outage with its corresponding BS (i.e., $S_D < S_{min}$). The model has been developed in Python by using the *TF-agents* library [21], which provides tools for the development of DRL models, including DQN. The DQN parameters are detailed in Table III.

B. Key Performance Indicators

This section describes the KPIs considered for assessing the performance of the proposed solution:

- **Average reward R_w:** This measures the average of the reward values (i.e., the *AGE* values) obtained for all the actions taken by the DQN agent during the evaluation of the learned policy, which is:

$$R_w = \frac{1}{N}\sum_{n=1}^{N} r(t_n) \tag{11}$$

where N denotes the total number of actions selected during the evaluation and $r(t_n)$ is the reward obtained as a result of the action made at time t_n, n=1,...,N.

- **RUE time in active mode M_T:** This measures the total cumulative time that all RUEs have been active during all evaluations, that is:

$$M_t = \sum_{r=1}^{R}\sum_{n=1}^{N} \Delta T \cdot a_{b,r}(t_n) \tag{12}$$

- **Efficiency of RUEs in active mode F :** This KPI measures how much of the time that an RUE is in active mode is actually efficient, meaning that it has served at least x=1 UE on average. If we define $K(t_n)$ as the total number of activated RUEs given an action $a(t_n)$, the KPI is given by:

$$F = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{K(t_n)} \sum_{k=1}^{K(t_n)} Ae_k \tag{13}$$

where Ae_k takes the value 1 if the average number of UEs served by the k-th activated RUE during period (t_n-ΔT, t_n) has been at least x, and it takes the value 0 otherwise.

- **Average Spectral efficiency S:** This KPI measures the average spectral efficiency obtained during the evaluation process. For this purpose, during each time period (t_n-ΔT, t_n), we measure the spectral efficiency obtained by all the UEs with an active session, taking one sample per UE every second, resulting in a total of $N_{samp}(t_n)$ samples denoted as $s_{eff}(i,t_n)$. Then, the average spectral efficiency is given by:

$$S = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{N_{samp}(t_n)} \sum_{i=1}^{N_{samp}(t_n)} s_{eff}(i,t_n) \tag{14}$$

- **Outage probability O:** This KPI computes the outage probability of a policy during the evaluation process. For this purpose, during each time period (t_n-ΔT, t_n), we take one sample every second for each UE with an active session, where the *i-th* sample is $o(i,t_n)$=1 if the UE experiences a spectral efficiency lower than S_{min}=1 b/s/Hz and $o(i,t_n)$=0 otherwise. Then, denoting the total number of samples for each period as $N_{samp}(t_n)$, the outage probability is given by:

$$O = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{N_{samp}(t_n)} \sum_{i=1}^{N_{samp}(t_n)} o(i,t_n) \tag{15}$$

C. Assessment of the training stage

The training process of the DQN algorithm is assumed to be performed offline in order to learn the policy that will be applied later on in the real or physical network in the so-called inference stage. This approach is aligned with recent initiatives such as O-RAN [24], which consider that ML models are trained on a training host before deploying them in the physical network. On the one hand, the offline training allows that the DQN agent is exposed to a wide variety of network conditions during the training and, on the other hand, it avoids that wrong decisions made during training negatively impact the performance of the physical network. To assess the training stage, this paper considers a training scenario in which UEs act according to the operation and parameters presented in Section IV.A. As explained in Section III, in each training step of ΔT seconds, the DQN agent selects a given action $\boldsymbol{a}(t)$ that results in the activation or deactivation of the available RUEs. After applying an action, the scenario continues its normal operation, and the reward is measured to progressively update the policy. The training was executed until reaching the maximum number of training steps MaxNumberOfTrainingSteps=$2 \cdot 10^5$.

In the following, some results are presented to study how the policy is progressively improved during the learning process. For this purpose, every 500 training steps, we obtained the current policy of the DQN agent, and we executed an evaluation of this policy by applying it for one hour in a given evaluation scenario that corresponds to a certain realization of the random traffic generation and spatial traffic distribution processes (in this way, the evaluations of all the policies every 500 training steps were performed under the same conditions, so they were comparable).

As a result of this one-hour evaluation, we collected the average reward. Fig. 3 plots the evolution of this average reward obtained with the learned policies every 500 training steps. The results show that during the first $80 \cdot 10^3$ training steps, the behavior of the learned policy is quite unstable. In fact, at approximately $75 \cdot 10^3$ steps, there is a significant decrease in the performance, reflecting that the training with this number of steps is still insufficient and the policy has not learnt to select optimal actions yet. However, after this period, the average reward starts to increase and tends to stabilize after approximately $175 \cdot 10^3$ training steps. It is worth mentioning that, being an offline training, the performance degradation observed at $75 \cdot 10^3$ does not have any impact on the physical network. Instead, it is the policy obtained at the end of the training process the one that will determine the performance in the physical network, as studied in the next sub-section.

D. Performance evaluation of the learnt DQN-based policy

This section assesses the performance obtained by the RUE activation in the inference stage using the policy learned by the DQN agent after completing the training process. For benchmarking purposes, the following reference strategies of RUE activation were considered:

Fig. 3. Evolution of the average reward as a function of the training steps.

1. **Random RUE activation:** this strategy consists of randomly selecting an action $\boldsymbol{a}(t) \in \mathcal{A}$ in every time window ΔT. The actions are selected with equal probability. Therefore, an RUE has a probability of being activated of 50% in every time window.
2. **All RUEs activated:** this strategy maintains all the RUEs activated during the whole time of evaluation. Therefore, this strategy will provide the best possible performance in terms of average spectral efficiency and outage probability, but at the cost of having all RUEs always activated even if they may not be necessary during certain periods of time.
3. **All RUEs deactivated:** this strategy keeps deactivating the RUEs during the entire evaluation time, so it can be considered the classical RAN in which no relay UE capabilities are exploited.
4. **Genetic algorithm:** A genetic algorithm is an optimization technique inspired by the principles of natural selection and genetics [25]. It is used to find solutions to complex problems by mimicking the process of evolution. This strategy is executed in every time window ΔT to decide the combination of relays to be activated/deactivated at that time. The algorithm starts from an initial population of potential solutions that consists of N_p individuals (i.e., actions $\boldsymbol{a}(t) \in \mathcal{A}$ in the problem considered here). These individuals are in turn constituted by a number of genes G, where each gene corresponds to the value of $a_{b,r}(t)$ for the r-th RUEs in BS b. Each individual is evaluated in terms of their fitness, which measures its suitability to solve the given problem. Specifically, the value of fitness considered here is the *AGE* in (3). Then, through a process of selection, crossover and mutation (modeled with probability of mutation P_{mut}), new generations of individuals are created (details can be found in [25]). This cycle continues for several generations, gradually improving the quality of the solutions. The algorithm terminates when reaching a maximum number of generations N_{gen}. Then, the final solution is the individual with the highest fitness found in all the generations. The considered parameters for the genetic algorithm are N_p=8, G=4, P_{mut}=0.1 and N_{gen}=100.

5. **Algorithm 'G' from [10]:** this algorithm intends to activate the RUEs that minimize the outage while maximizing the spectral efficiency. To achieve the latter, each UE in outage is associated with the candidate RUE that provides the highest value of spectral efficiency and fulfills $S_R > S_{min}$. After analyzing this for all the UE devices, the strategy activates all the relays that could serve at least one UE device. This strategy assumes perfect instantaneous knowledge of the spectral efficiency conditions for all the active UEs in the links with their serving BSs and in the links with all the candidate RUEs.
6. **Exhaustive search strategy:** this strategy applies at each time step an exhaustive search process among all possible configurations $\mathbf{C_b}(t)$ to find the one that ensures the maximum aggregate global efficiency *AGE*, thus maximizing the target of the optimization problem formulated in Section II. This exhaustive search strategy also assumes perfect instantaneous knowledge of the spectral efficiency conditions of all the active UEs in the links with their serving BSs and in the links with all the candidate RUEs.

The comparison between RUE activation strategies was performed by conducting 100 different policy-evaluation procedures of one-hour duration for each strategy. Each evaluation procedure consisted of applying the assessed strategy under a given realization of traffic generation from the UE in the scenario. Then, by varying the traffic generation realization in each evaluation procedure, the performance of the different strategies was assessed in a wide range of situations in the considered scenario. It is also worth noting that, for a given evaluation procedure, all the strategies were applied with the same traffic realization, so that they were evaluated under exactly the same conditions.

Fig. 4 shows the average reward or, equivalently, the average value of the *AGE* obtained as a result of all the evaluations for the different strategies. The *Exhaustive search strategy* corresponds to the theoretical upper bound of the achievable performance since it assumes perfect instantaneous knowledge of the spectral efficiency conditions in all the different links. *Algorithm 'G'* performs very similarly to the *Exhaustive search strategy* because it also considers the same theoretical assumption of perfect knowledge. The difference can be explained as follows. The *Exhaustive search strategy* is optimized to maximize the *AGE*, while *Algorithm 'G'* targets outage minimization and spectral efficiency maximization.

The results reflect that the *DQN-based strategy* achieves an efficiency equivalent to 84.7% of the *Exhaustive search strategy* and 86% of *Algorithm 'G'*. When compared to the other strategies different from the *Exhaustive search* and the *Algorithm 'G'*, it is observed from Fig. 4 that the *DQN-based strategy* outperforms all of them. Although the result with respect to the *All RUEs deactivated* strategy was expected, when compared to the *random* strategy, the *DQN-based strategy* is 119% more efficient in terms of the reward. Moreover, the comparison with the *All RUEs activated* strategy shows that maintaining all the relays activated during the whole evaluation time does not result in an efficient strategy. In fact, our proposed algorithm outperforms the *All RUEs activated* strategy by 75% since this strategy keeps some RUEs activated even when this is not needed according to the network dynamics.

The comparison between the *Genetic algorithm* and the *DQN-based strategy* in Fig. 4 shows that DQN achieves a bit better performance with an average improvement of 2.6%. This reflects that both algorithms properly solve the optimization problem. However, from an implementation perspective, an importance difference between both approaches is that the *Genetic algorithm* has to conduct an optimization every time that a decision has to be made and this involves an evolutionary search process to find a solution. In contrast, the *DQN-based strategy* benefits from the experience acquired during the training, so that it makes decisions much faster. In fact, using a computer with a Core i5-6400 2.7 GHz processor and 8 GB of RAM, and considering all evaluation procedures, it has been obtained that an execution of the DQN agent in the inference stage lasts on average 0.017 ms, while the *Genetic algorithm* requires on average 104.3 ms.

To deeply analyze the performance of the different strategies in a wide range of situations and to explore how the improvements achieved by the *DQN-based strategy* vary in different policy evaluation procedures, Fig. 5 plots a boxplot of the percentage of reward increase achieved by the DQN-based strategy with respect to all the other strategies considering the 100 policy-evaluation procedures. The boxplot reflects the distribution of this reward improvement and allows us to establish the ranges of best and worst performance. Specifically, for each strategy, the top and bottom lines shown in the plot reflect the maximum and minimum values, while the box represents the range between the 25th and 75th percentiles of the distribution. With respect to the upper bound strategies *Exhaustive search strategy* and *Algorithm 'G'*, the boxplot shows that, although they obtain a higher reward than the *DQN-based strategy*, the differences are small. In the best case, the reward reduction of the *DQN-based strategy* with respect to the *Exhaustive search strategy* is only 4.5%, and it is 2.7% with respect to *Algorithm 'G'*. In the worst case, the reward reduction is 30% with respect to the *Exhaustive search strategy* and 28% with respect to the *'G' algorithm*. The median reward reduction is 14.9% for the *Exhaustive search strategy* and 14.8% for *Algorithm 'G'*.

When compared with the rest of the strategies, the median of the DQN improvement is 1.89% against *Genetic algorithm*, 75.6% against the *All RUEs activated* strategy, 119% against the *Random RUEs activation* strategy and 191% against the *All RUEs deactivated* strategy. In the best situations, the improvements are 20.1%, 105%, 178%, and 373%, respectively, while in the worst cases, they are -16.5%, 42.6%, 61.6%, and 77.7%. These numbers confirm that, thanks to the training process, the DQN-agent is able to learn when the activation of the RUEs will be beneficial, which in turn will translate into a benefit for both network users and MNOs.

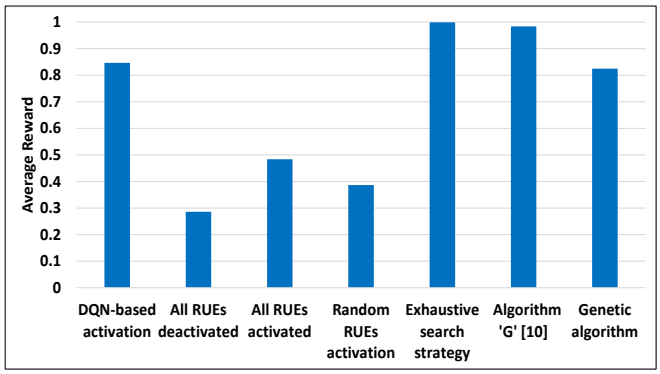

Fig. 4. Average reward for the different strategies.

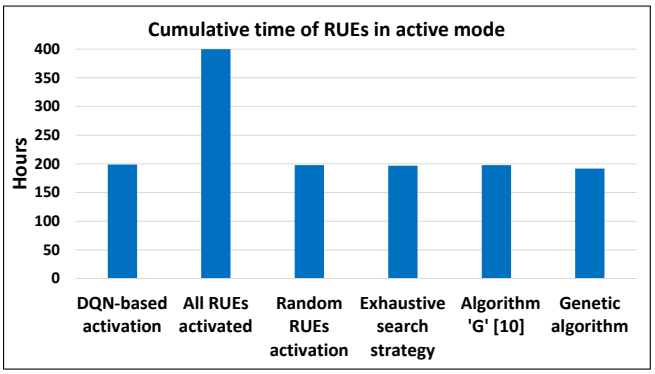

Fig. 6. Time of RUEs in active mode for the different strategies.

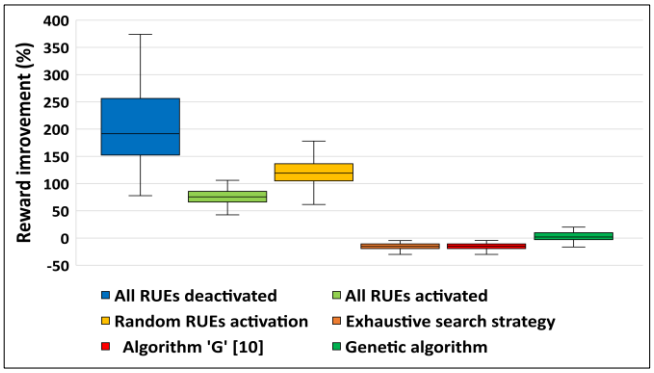

Fig. 5. Boxplot of the reward improvement of the DQN-based strategy against the benchmarking strategies for the 100 policy-evaluation procedures.

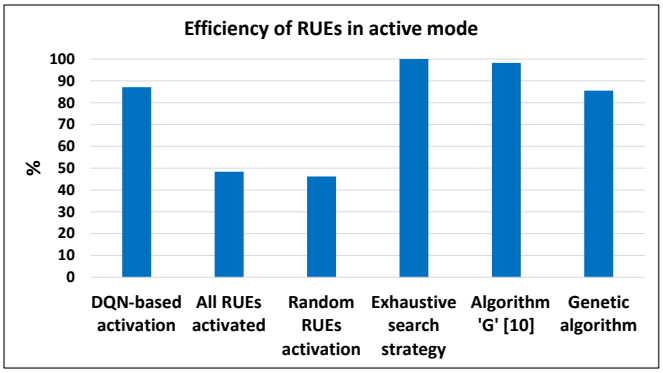

Fig. 7. Efficiency of RUEs in active mode for the different strategies.

The RUE time in active mode is shown in Fig. 6, where we compare the total time for different strategies. For the *All-RUEs activated* strategy, the RUEs are activated for a total of 400 hours (i.e., 4 RUEs activated during 100 hours each), while for the *Exhaustive search strategy* and *Algorithm 'G'*, the RUEs are active for a total of 196.7 hours and 197.7 hours, respectively. These numbers are very close to the *DQN-based strategy*, which has the RUEs active for a total time of 198.75 hours, i.e., a difference of less than 1% with the upper bound strategies. This value is also similar to the one obtained with the random strategy because this strategy tends to activate on average approximately the same number of RUEs, but its random behavior leads to poorer performance in terms of *AGE*. A similar situation occurs when comparing with the *Genetic algorithm*, since it maintains RUEs active during 191.7 hours but at the cost of a slightly worse performance in terms of *AGE* as seen in the previous results of Figs. 4 and 5. It is clear that by performing a proper activation, the DQN strategy performs closely to the upper bound strategies and clearly reduces by more than 200 hours the time that RUEs remain in active mode with respect to the *All-RUEs activated* strategy. This will translate into significant energy savings.

Beyond this time reduction, it is also important to measure how much of the time that an RUE is activated is actually useful. For this purpose, Fig. 7 shows the efficiency of RUEs in active mode considering the different activation strategies.

As expected, the highest efficiency values correspond to the upper bound strategies, where *Algorithm 'G'* reaches a value of 98.2% while the *Exhaustive search strategy* reaches 100%. The *DQN-based strategy*, on the other hand, has an efficiency of 87.1% and the *Genetic algorithm* reaches an efficiency of 85.5%. In contrast, the efficiency is substantially lower with the *Random RUEs activation* and *All-RUEs activated* strategies, whose values are 46.2% and 48.3%, respectively. Thus, the *DQN-based strategy* is slightly superior to the *Genetic algorithm* and achieves almost twice the performance of the *Random RUEs activation* and *All-RUEs activated* strategies due to the capability of the DQN-based algorithm to consider network context information when deciding which RUEs to activate. Additionally, the comparison against the *All-RUEs activated* strategy reflects that keeping the RUEs activated all the time does not result in an efficient method. The reason is that, given the network dynamics, at certain times, it is not necessary or efficient to use relays; therefore, activation may result in a waste of resources.

To assess the network performance improvement brought by the activated RUEs, Fig. 8 plots the obtained outage probability for the different strategies evaluated in the paper. A relevant result extracted from Fig. 8 is the significant difference between using or not using relays. For the case when there are no activated relays in the network, the outage probability is 10%, while by using the *DQN-based activation* strategy, the outage is reduced down to approximately 1.1%. In other words, by using and activating relays correctly as the DQN algorithm

does, the outage probability is reduced by approximately 89%. This is a stronger reduction than the one obtained with random activation, which provides an outage of approximately 5.1% or a reduction of only 49%.

Note that in the case of *Algorithm 'G'*, *Exhaustive search strategy* and *All-RUEs activated* reduce the achieved outage probability to almost 0. For the case of the *All-RUEs activated* strategy, it was expected because all relays are always active, but as seen previously, this is at the expense of a much longer time in active mode. The *DQN-based activation* is superior to the *Genetic algorithm* at minimizing the outage probability. Indeed, the *DQN-based activation* is able to reach an outage probability similar to the *All-RUEs activated* strategy but with 50% less time spent by relays in active mode. It is also worth remarking that from the perspective of outage, the performance of *Algorithm 'G'* is better than that of the *Exhaustive search strategy* because the former intends to minimize the outage, while the *Exhaustive search strategy* targets the maximization of the *AGE*.

Fig. 9 plots the average spectral efficiency obtained by applying the different strategies. The *DQN-based strategy* clearly outperforms the *random* and *All RUEs-deactivated* strategies, with improvements of 5.3% and 12.3%, respectively and also showed some improvement against the *Genetic algorithm*. In turn, the *All RUEs Activated* strategy provides the highest spectral efficiency but is only 1.2% higher than with the *DQN-based strategy*, which achieves very similar performance to the *Exhaustive search strategy* and *Algorithm 'G'* strategies. Considering that the *DQN-based strategy* keeps the relays active for less than half of the time than the *All RUEs Activated* strategy did, it is concluded that this strategy achieves a better trade-off between spectral efficiency improvement and RUE activation time.

To assess the range of spectral efficiency values, Fig. 10 plots the cumulative distribution function (CDF) of the spectral efficiency for the different activation strategies. The CDF as a function of a given value of spectral efficiency, s_e, is given by the probability that a spectral efficiency sample $s_{eff}(i,t_n)$ is lower than s_e:

$$CDF(s_e)=P(s_{eff}(i,t_n) \leq s_e) \quad (16)$$

this computation is performed considering all the spectral efficiency samples taken during the 100 policy-evaluation procedures.

When comparing the DQN strategy against both upper bound references, the good performance of our proposed approach is exhibited since a very similar CDF is observed. In fact, if we measure the probability that a given UE experiences a spectral efficiency of at least 5 b/s/Hz, with the *Exhaustive search strategy* and *Algorithm 'G'*, this probability is 84.1% and 86%, respectively, while with the *DQN-based strategy*, it is 83%, which is a small difference. It is worth emphasizing the importance of proper activation of the relays that our DQN approach achieves. When performing an activation of the relays based on the *Genetic algorithm*, the probability of experiencing a spectral efficiency of at least 5 b/s/Hz is 79% and takes a value of 56% when performing a random activation. Overall, a significant difference is observed between using or not using RUEs. In fact, it can be observed that when there are no RUEs activated, the probability of having a spectral efficiency higher than 5 b/s/Hz is only 25%, which is much lower than that of the *DQN-based activation strategy*.

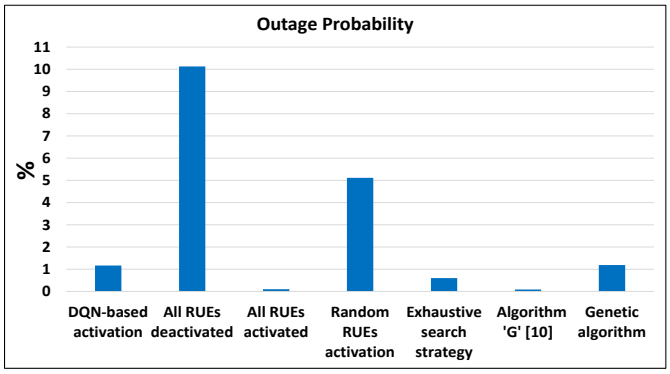

Fig. 8. Outage probability of UEs for different activation strategies.

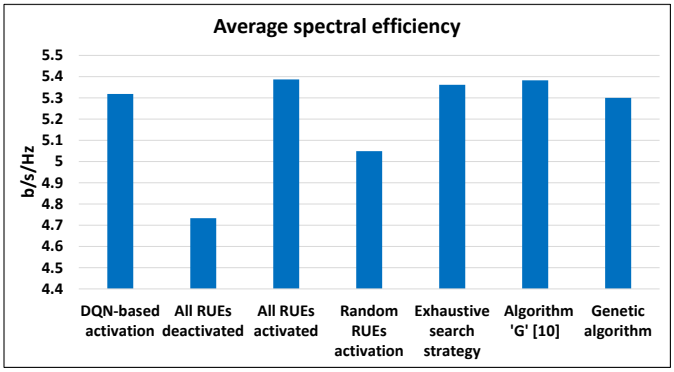

Fig. 9. Average spectral efficiency for different activation strategies.

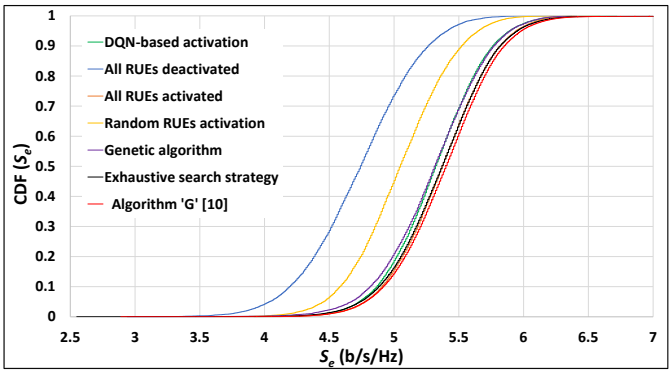

Fig. 10. CDF of the spectral efficiency obtained for different activation strategies.

V. Conclusion and Future Work

This paper has presented a novel approach for exploiting UE's capability to be activated as relays (RUEs) as a way of augmenting the radio access network (RAN) infrastructure and thus increasing the network coverage and improving the network and UE performance. Specifically, the paper has

proposed a new strategy for efficiently deciding when to activate or deactivate a certain relay UE. The activation function is based on the deep Q-network algorithm that makes use of the network context information and a training process to learn a policy for activating only those RUEs that are useful under the given the network conditions.

The behavior of the proposed DQN-based RUEs activation function has been assessed by means of system-level simulations and contrasted against five other reference strategies. The results have shown that (i) the DQN-based activation strategy is able to learn from the network context information and consequently to activate the relays properly. This is reflected in high performance in terms of aggregate global efficiency. (ii) The policy learned by the DQN agent substantially reduces the time that relays remain active and significantly increases the average global efficiency. (iii) The proposed approach leads to important outage probability reductions and spectral efficiency increases with respect to the strategies without RUEs and when RUEs are activated randomly. Specifically, the average spectral efficiency is improved in 12.3% and the outage probability is reduced in 89% with respect to the case without RUEs. Moreover, it provides a very similar performance to the strategy that keeps the RUEs activated all the time, although the relays are active less than 50% of the time. (iv) When compared against a classical optimization strategy such as a genetic algorithm the DQN approach exhibits a bit better performance and an important reduction in the execution time for making decisions. (v) The results show that the performance of the DQN strategy is quite close to that of two theoretical upper bound strategies that operate with perfect instantaneous knowledge of all the link conditions.

To summarize the overall performance of the proposed DQN strategy, Fig. 11 plots a radar chart displaying the normalized results for all the studied metrics comparing the DQN-based strategy, against the *All RUEs-deactivated strategy* and the *Exhaustive search strategy*. The performance of the DQN approach has only minimal deviations from the *Exhaustive search strategy*, and it achieves important improvements with respect to the *All RUEs-deactivated* strategy. This highlights two main findings: firstly, it emphasizes the network performance benefits gained through the efficient use of relays; secondly, considering that the exhaustive search approach acts as a theoretical upper bound benchmark, the similarity between the DQN-based approach and this upper bound confirms the effectiveness of our proposed approach for tackling the relay activation problem.

Following the promising results obtained, our future work intends to study the performance of the model based on actual network measurements. Moreover, the identification of the mechanisms required for the practical implementation of the proposed solution, including the implications on current management interfaces and message exchanges between involved nodes, is also envisioned as future work.

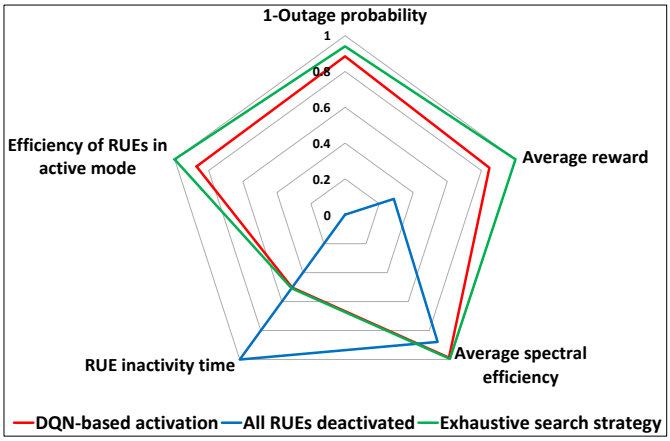

Fig. 11. Radar plot of the performance obtained by different activation strategies for the considered normalized metrics.

REFERENCES

[1] Ericsson Mobility Report, June, 2022. [Online]. Available: https://www.ericsson.com/49d3a0/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-june-2022.pdf.

[2] J. Pérez-Romero and O. Sallent, "Leveraging User Equipment for Radio Access Network Augmentation," in *Proc. IEEE Conference on Standards for Communications and Networking (CSCN)*, Thessaloniki, Greece, 2021, pp. 83-87.

[3] J. Sydir and R. Taori, "An evolved cellular system architecture incorporating relay stations," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 115-121, Jun. 2009.

[4] O. Teyeb, A. Muhammad, G. Mildh, E. Dahlman, F. Barac and B. Makki, "Integrated Access Backhauled Networks," in *Proc. IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Honolulu, HI, USA, 2019, pp. 1-5.

[5] M. Polese *et al*., "Integrated Access and Backhaul in 5G mmWave Networks: Potential and Challenges," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 62-68, Mar. 2020.

[6] Study on architecture enhancements for vehicle-mounted relays (Release 18), v1.1.0, 3GPP Report TR 23.700-05, Oct. 2022.

[7] *Enhanced Relays for Energy Efficiency and Extensive Coverage; Stage 1 (Release 17), V17.1.0*, 3GPP Report TR 22.866, Dec. 2019.

[8] *Service requirements for 5G system; Stage 1 (Release 18), V18.5.0*, 3GPP Standard TS 22.261, Dec. 2021.

[9] P. Mach, and Z. Becvar, "Device-to-Device Relaying: Optimization, Performance Perspectives, and Open Challenges Towards 6G Networks," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 3, pp. 1336-1393, third quarter 2022.

[10] J. Pérez-Romero, and O. Sallent, "On the Value of Context Awareness for Relay Activation in Beyond 5G Radio Access Networks," in *Proc. IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, Helsinki, Finland, 2022, pp. 1-6.

[11] J. Pérez-Romero, O. Sallent, and O. Ruiz, "On Relay User Equipment Activation in Beyond 5G Radio Access Networks," in *Proc. IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, London, United Kingdom, 2022, pp. 1-6.

[12] A. Abdelreheem, O. A. Omer, H. Esmaiel, and U. S. Mohamed, "Deep Learning-Based Relay Selection In D2D Millimeter Wave Communications," in *Proc. International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2019, pp. 1-5.

[13] H. Kim, T. Fujii, and K. Umebayashi, "Relay Nodes Selection Using Reinforcement Learning," in *Proc. International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Jeju Island, Korea (South), 2021, pp. 329-334.

[14] *NR; Physical layer procedures for data (Release 17), V17.0.0*, 3GPP Standard TS 38.214, Dec. 2021.

[15] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26-38, Nov. 2017.

[16] V. Mnih, et al. "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[17] I. Vilà, J. Pérez-Romero, O. Sallent, and A. Umbert, "A Multi-Agent Reinforcement Learning Approach for Capacity Sharing in Multi-Tenant Scenarios," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9450-9465, Sep. 2021.

[18] Y. Zhang, C. Kang, Y. Teng, S. Li, W. Zheng, and J. Fang, "Deep Reinforcement Learning Framework for Joint Resource Allocation in Heterogeneous Networks," in *Proc. IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1-6.

[19] J.J. Hernández-Carlón, J. Pérez-Romero, O. Sallent, I. Vilà, and F. Casadevall, "A Deep Q-Network-Based Algorithm for Multi-Connectivity Optimization in Heterogeneous Cellular-Networks," *Sensors*, vol. 22, no. 16, Aug. 2022.

[20] U. Challita and D. Sandberg, "Deep Reinforcement Learning for Dynamic Spectrum Sharing of LTE and NR," in *Proc. IEEE International Conference on Communications (ICC)*, Montreal, QC, Canada, 2021, pp. 1-6.

[21] Guadarrama, S.; Korattikara, A.; Ramirez, O.; Castro, P.; Holly, E.; Fishman, S.; Wang, K.; Gonina, E.; Wu, N.; Kokiopoulou, E.; et al. TF-Agents: A Library for Reinforcement Learning in TensorFlow. 2018. [Online]. Available: https://github.com/tensorflow/agents.

[22] *Study on Channel Model for Frequencies From 0.5 to 100 GHz (Release 16), V16.1.0*, 3GPP Report TR 38.901, Dec. 2019.

[23] "TDD UE-UE Interference Simulations", Siemens AG, R4-030189 document of the 3GPP TSG-RAN Working Group 4 meeting #26, Feb. 2003.

[24] *AI/ML Workflow Description and Requirements 1.03*, document O-RAN.WG2.AIML-v01.03, O-RAN Alliance, Oct., 2021.

[25] L. Davis, *Handbook of Genetic Algorithm.* New York, NY, USA: Van Nostrand, 1991.

Juan Jesus Hernández-Carlón received the B.E. degree in electronics and telecommunication engineering and the M.E. degree in telecommunication engineering from the National Polytechnic Institute (IPN), Mexico City, Mexico, in 2014 and 2017, respectively. In 2019, he joined the Mobile Communication Research Group (GRCM), Department of Signal Theory and Communications (TSC), UPC, where he is currently pursuing a Ph.D. supported with FPI grant by the Ministry of Science and Innovation of the Government of Spain. His research interests include radio access network (RAN) management, multiconnectivity in heterogeneous networks, and the application of deep reinforcement learning to radio resource management.

Jordi Pérez-Romero (Member, IEEE) received a degree in telecommunications engineering and a Ph.D. degree from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1997 and 2001, respectively. He is currently a Professor with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya. He is working in the field of wireless communication systems, with a particular focus on radio resource management, cognitive radio networks and network optimization. He has been involved in different European projects with different responsibilities, such as researcher, work package leader, and Project Responsible, has participated in different projects for private companies and has contributed to the 3GPP and ETSI standardization bodies. He has authored or coauthored more than 250 papers in international journals and conferences and three books and has contributed to seven book chapters. He is an Associate Editor for the IEEE Vehicular Technology Magazine.

Oriol Sallent is currently a Professor with the Universitat Politècnica de Catalunya, Barcelona, Spain. He has participated in a wide range of European and national projects, with diverse responsibilities as a principal investigator, co-ordinator, and work package leader. He regularly serves as a consultant for a number of private companies. He has been involved in the organization of many different scientific activities, such as conferences, workshops, and special issues in renowned international journals. He has contributed to standardization bodies such as 3GPP, IEEE, and ETSI. He is the coauthor of 13 books and has authored or coauthored more than 250 papers, mostly in high-impact IEEE journals and renowned international conferences. His research interests include 5G RAN (Radio Access Network) planning and management, artificial intelligence-based radio resource management, virtualisation of wireless networks, cognitive management in cognitive radio networks and dynamic spectrum access and management, among others.

Irene Vilà received a B.E. degree in Telecommunication Systems Engineering in 2015, an M.S. degree in Telecommunication Engineering in 2017, and a Ph.D. degree in Signal Theory and Communications in 2022, all from Universitat Politècnica de Catalunya (UPC). She is currently a postdoctoral researcher between the CONNECT center in Trinity College Dublin (TCD) and the Department of Signal Theory and Communications (TSC) at UPC. She has been involved in different national and European research projects founded by both public and private organizations, and she has published more than 15 papers in international journals and conferences. Her research interests focus on the field of mobile communications, particularly on radio resource management, network optimization, radio access network (RAN) slicing and the application of artificial intelligence and, particularly, machine learning to RAN management.

Ferran Casadevall (M'87) received the Engineer (1977) and Doctor Engineer (1983) degrees in Telecommunications Engineering from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. ` In 1978, he joined UPC, where he was an Associate Professor from 1983 to 1991. He is currently a Full Professor with the Department of Signal Theory and Communications, UPC. Over the last 30 years, he has been mainly concerned with the performance analysis and development of digital mobile radio systems. He has published approximately 250 technical papers in both international conferences and magazines, most of them corresponding to IEEE publications. His particular research interests include cellular and personal communication systems, cognitive radio issues, radio resource management techniques, and network optimization. He has participated in more than 40 research projects funded by both public and private organizations. In particular, he has actively participated in 20 research projects funded by the European Commission, being the Coordinator and Project Manager for three of them: Advanced Radio Resource management for Wireless Systems (ARROWS), Evolutionary Strategies for Radio Resource Management in Cellular Heterogeneous Networks (EVEREST), and Advanced Resource Management Solutions for Future All IP Heterogeneous Mobile Radio Environments (AROMA). Prof. Casadevall has been a Technical Program Committee Member for different international IEEE supported conferences and a Reviewer for several IEEE magazines. He has also participated in several research evaluation panels. From October 1992 to January 1996, he was in charge of the Information Technology Area, National Agency for Evaluation and Forecasting (Spanish National Research Council).