IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 November 2024, accepted 24 November 2024, date of publication 27 November 2024, date of current version 9 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3507577

RESEARCH ARTICLE

Deep Reinforcement Learning-Based Multi-Access in Massive Machine-Type Communication

NASIM RAVI, (Student Member, IEEE), NUNO LOURENÇO, MARILIA CURADO, AND EDMUNDO MONTEIRO, (Senior Member, IEEE)
University of Coimbra, Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, 3030-290 Coimbra, Portugal

Corresponding author: Nasim Ravi (nasimravi@dei.uc.pt)

This work is funded in part by The Science and Technology Development Fund, Macau SAR. (File no. 0044/2022/A1); in part by the National Funds through the Foundation for Science and Technology (FCT), I.P., within the Scope of the Project CISUC under Grant UID/CEC/00326/2020; and in part by European Social Fund through the Regional Operational Program Centro 2020.

ABSTRACT The diverse applications of Machine-Type Communication (MTC) lead to exponential growth in Machine to Machine traffic. In connection with MTC deployment, a large number of devices are expected to access the wireless network simultaneously, resulting in network congestion. The conventional Random Access mechanism lacks the capability to handle the large number of access attempts expected from massive MTC (mMTC). Additionally, mMTC transceivers often operate by generating short data packets in a sporadic and sometimes unpredictable manner. To address the growing need for efficient communication in massive machine-type communication scenarios we propose an innovative solution, called Deep Reinforcement Learning-Based Multi-Access (DRLMA). Our model considers the Base Station (BS) as an agent navigating the landscape of machine-type communication devices. This agent dynamically switches between grant-based and grant-free access to leverage their strengths. We address the multi-access problem, formulating it as a Partially Observable Markov Decision Process (POMDP), to better understand and tackle challenges associated with dynamic access policies. Leveraging Deep Reinforcement Learning techniques, our approach optimizes sporadic traffic patterns, crafting an adaptable access policy to maximize both network throughput and energy efficiency under the battery constraint. Simulation results show that proposed DRLMA scheme outperforms traditional access schemes and existing access protocols in sporadic traffic in terms of the energy efficiency, throughput and network life time.

INDEX TERMS Massive machine-type communication, multiple access, deep reinforcement learning, grant-based, grant-free, sporadic traffic.

I. INTRODUCTION

Fifth generation (5G) and Beyond 5G (B5G) wireless networks have been developed to support a wide range of highly demanding services and applications by pushing network capabilities for improving performance. To meet all the needs and requirements of mobile networks in the future, the projected use of mobile networks is divided into three use-cases: Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC), and massive Machine-Type Communication (mMTC), each serving different application purposes. mMTC applications cover various sectors like transportation, utilities, health, environment, and security. In these scenarios, data packets often contain small amounts of information, such as control commands or sensor readings. Despite the relatively small data size, mMTC networks can comprise significantly more nodes compared to regular consumer mobile cellular networks [1].

3GPP initiated feasibility studies for Machine-Type Communication (MTC), defining it as a form of communication involving one or more entities that do not necessarily require human interaction. Most Machine-Type Devices (MTDs) are low-complexity devices and should operate for years entirely dependent on low cost batteries. Furthermore, these devices are usually placed in locations that suffer from

The associate editor coordinating the review of this manuscript and approving it for publication was Walid Al-Hussaibi.

significant penetration loss in the building (up to 50 dB), such as deep inside buildings or basements. In other words, they generally have unfavorable link budgets. Therefore, the main challenges in mMTC applications are to ensure the connection of a large number of MTDs while maintaining network performance and to minimize energy consumption by MTDs [2].

The Medium Access Control (MAC) layer is vital for managing channel access among nodes in a network. For mMTC, MAC protocols need to be efficient, scalable, low-power, and low-latency. However, conventional Random Access (RA) mechanisms, like those in Long Term Evolution (LTE) networks, struggle with a high number of connected devices, leading to network congestion and degraded Quality-of-Service (QoS). In cellular networks, the random RA procedure relies on grant-based and grant-free transmission schemes. Grant-based transmission involves devices requesting and receiving explicit permission (grant) from the base station before transmitting data, ensuring controlled and coordinated access with QoS guarantees but introducing overhead and delay. In contrast, grant-free transmission allows devices to autonomously initiate transmission without prior grant requests, reducing signaling overhead and latency. These grant-free schemes are suitable for mMTC scenarios with short packets, enabling direct transmission from active devices to the base station [3], often including metadata like preambles for device detection and channel estimation [4], but may face challenges such as increased collision probability and reduced coordination.

A. MOTIVATION AND CONTRIBUTIONS

As the number of MTC devices grows to hundreds of millions [5], the access channels in existing cellular networks are expected to face severe congestion and increased signaling overhead. This issue is particularly problematic with grant-based access methods, which require devices to request and receive permission before transmitting data, leading to significant delays and inefficiencies. On the other hand, in grant free, devices send data without requesting channel access and the lack of coordination results in tremendous amount of collisions due to the limited resources for massive number of MTCDs. Additionally, the heterogeneous nature of the network further complicates the design of an effective solution. Previous approaches have traditionally optimized either grant-based or grant-free schemes separately, for example, existing studies have worked on optimizing resource allocation in grant-based access or solving challenges like activity and signal detection in grant-free access, without addressing both methods simultaneously.

In our previous work [6], we introduced a Dynamic Switching Access (DSA) protocol that focused solely on switching from grant-free to grant-based access based on the continuously increasing number of active devices in a periodic traffic environment with constant arrival rates. The limitation of that work was that it only allowed switching in one direction, from grant-free to grant-based, due to the nature of the traffic model, which did not account for the sporadic traffic patterns seen in real-world MTC environments. In contrast, the current work significantly extends that approach by introducing a bi-directional switching mechanism that dynamically adapts between both access methods (grant-free and grant-based) based on real-time network conditions. This allows for more flexible and efficient management of network resources in environments where traffic is unpredictable. Additionally, unlike the previous work, which relied on a static threshold-based method, this paper formulates the access problem as a Partially Observable Markov Decision Process (POMDP), enabling more sophisticated decision-making in dynamic and uncertain environments. By applying Deep Reinforcement Learning (DRL), we optimize the access policy to handle sporadic traffic effectively, allowing the system to adapt flexibly between access protocols depending on network conditions. The major contributions of this paper can be summarized as follows:

- We propose a novel approach that leverages the strengths of distinct access protocols through an intelligent switching mechanism. This involves the base station guiding devices to adapt their access type based on real-time network conditions. In massive MTC environments with sporadic data transmissions, this approach allows devices to dynamically switch access type, enhancing network performance;
- We model the energy efficiency and throughput separately for both access types to demonstrate their dependency on different access protocols;
- We address the multi-access problem by formulating it as a POMDP to better understand and resolve challenges associated with dynamic access policies;
- Leveraging DRL techniques, our approach optimizes sporadic traffic, crafting an adaptable access policy to maximize both network throughput and energy efficiency. Additionally, we consider an essential parameter of device batteries by treating the BS as the agent, thereby avoiding unnecessary drain on device batteries during training;
- We evaluate our proposed method against benchmarks and other access protocols in the literature. Numerical results show that our DRL-based algorithm outperforms traditional access schemes, achieving significant average gains of throughput, energy efficiency and battery life time;

B. PAPER ORGANIZATION

The rest of the paper is organized as follows. Section II provides an overview of related literature. Section III describes the system model and problem formulation, including energy efficiency and throughput modeling. Section IV details the proposed DRL framework, including its training process and algorithm. Section V introduces the simulation environment used to evaluate our approach and demonstrates efficiency of proposed Deep Reinforcement Learning-Based

Multi-Access (DRLMA) compared to existing methods and finally, Section VI concludes the work.

II. RELATED WORK

3GPP identified the RA process as crucial for improving MTC support and proposed candidate solutions to enhance Physical Random Access Channel (PRACH) performance in overload scenarios [7]. Guo et al. [8] utilized Access Class Barring (ACB) for reducing preamble collisions, with the scheme's performance varying based on parameter configurations. However, recent studies combine learning methods with ACB to enhance network performance; [9] integrates Long-Short Term Memory (LSTM)-based traffic prediction and RS with ACB, improving throughput compared to grant-free benchmarks. Tello-Oquendo et al. [10] proposed a dynamic algorithm using reinforcement learning to adapt ACB's barring rate parameter, effectively reducing congestion and collisions in the Random Access Channel (RACH). In [11] an optimal ACB control and resource allocation scheme proposed to maximize system capacity while ensuring efficient resource utilization. Additionally, Salam et al. [12] and [13] proposed prioritized contention-based MTCD access with dynamic resource allocation by the aggregator to enhance access management in mMTC networks.

In recent years, there has been a decline in the focus on enhancing the grant-based, suggesting a reduced emphasis on further improvements. However, a new grant-based scheme introduced in [14] enhances the massive random access Gaussian channel by enabling coordinated user interactions and improving overall performance. This scheme utilizes short broadcast feedback from the base station, considering energy consumption and transmission delay, and benefits from a simplified design and optimization process with its closed-form expression for error probability approximation. Additionally, Liu et al. [15] propose a grant-based random access transmission scheme based on a sparse Tanner graph, employing source channel estimation and M-ary modulation to reduce transmission delay and achieve high reliability with the proposed message-passing decoder. For energy-efficient solution, the hybrid access protocol proposed in [16] combines contention-based and scheduled-based access, addressing energy efficiency, bandwidth, and delay optimization. Furthermore, Kim et al. [17] introduced spatial group-based preamble allocation to enhance preamble detection probability. Configured Grant (CG) with explicit Acknowledgment (ACK) is introduced in [18] to enhance uplink URLLC transmission scheme to improve reliability.

Several grant-free based approaches have been introduced in the literature. Collisions are a major obstacle such that grant-free scheme cannot stand alone to support mMTC. While the number of preamble sequences is very large, the probability that two or more users choose the same test sequence is non-zero. In this case, a collision is said to have occurred because these devices cannot be detected by the base station. To address collisions, work [19] introduced a distributed layer allowance-free Non-Orthogonal Multiple Access (NOMA) framework, dividing the cell into layers with different power levels to minimize interference. However, collisions within each layer remain severe. Additionally, in [20] a priority-enabled grant-free access is proposed to dynamically adapt the number of slots within a subframe based on traffic load estimation. Furthermore, the NOMA-based multichannel ALOHA scheme introduced in [21] demonstrates the potential of NOMA for non-coordinated transmissions, where devices select predetermined power levels to reduce collisions. However, their focus is primarily on improving throughput by varying power levels and the number of subchannels. Additionally, methods to determine the number of preambles for collision reduction have been proposed [22], [23], [24], with focusing on achieving high success probability while minimizing single-user failure probability [22]. For the problem of active user in grant-free, study [25] introduces an innovative deep learning architecture designed to tackle the Active User Detection (AUD) problem in GF-NOMA under frequency-selective fading channels without needing Channel State Information (CSI) or user sparsity information. Joint device activity detection, channel estimation and data recovery considered in [26], [27], [28], and [29]. Regarding optimization, Machine Learning (ML), particularly Reinforcement Learning (RL), is frequently employed as optimization tool across various studies. However, RL encounters scalability issues in large networks due to extensive exploration time, while DRL, encompassing techniques like Deep Q-Network (DQN) [30], [31], [32] and Deep Deterministic Policy Gradients (DDPG), aims to address the RL limitations. Zhang et al. [33] proposed a deep reinforcement learning approach to mitigate collisions in grant-free NOMA systems by clustering access resources and user equipment into separate subsystems.

Several studies have focused on Semi-Grant-Free (SGF) transmission to balance resources between grant-free and grant-based methods. Study [34] explores the outage performance of a Rate-Splitting Multiple Access (RSMA)-aided SGF transmission system, optimizing power allocation and decoding order to maximize grant-free user's rates while preserving the grant-based user's outage performance, supported by analytical and simulation results. For outage exploring, [35] proposed a fair CS-SGF method to address admission fairness. Theoretical analyses reveal limitations in achieving full diversity orders, prompting the introduction of a distributed power control strategy. Additionally, a joint channel assignment and power allocation approach for semi-grant-free NOMA systems is introduced in a separate study [36] maximizing network throughput while meeting individual device requirements. Furthermore, Double Deep Q Networks (DDQN) techniques are applied in [37] to optimize transmit power in SGF-NOMA IoT networks, achieving significant throughput gains, while [38] proposes a RB-oriented power pool design for SGF-NOMA, addressing residual errors using Multi-Agent DRL (MA-DRL) algorithm.

Table 1 provides a summary of the related works discussed in this section.

Existing research has predominantly focused on enhancing either pure grant-based or pure grant-free schemes to address their respective challenges, often neglecting the potential benefits of leveraging both approaches. Furthermore, most studies in grant-based access emphasize methods such as Aggregated Channel Bandwidth (ACB), device aggregation, or clustering, while those in grant-free access primarily focus on resource allocation, channel estimation, and active data detection, without adequately considering the access protocols themselves. The issue of sporadic traffic has also been largely overlooked, revealing a gap in improving random access protocols. Even in semi-grant-free scenarios, where resources are shared, the access protocol typically remains constant across the network. Additionally, many recent learning-based solutions treat devices as agents, leading to increased energy consumption without addressing this concern. This highlights the need for our approach, which aims to integrate access protocols and develop DRLMA.

III. SYSTEM MODELING

This section is divided into three parts: Traffic Modeling, access mechanism and problem Formulation. We will explore the underlying traffic patterns influencing the system, detail the access mechanisms followed by a comprehensive formulation of the problems related to energy efficiency and throughput, highlighting the dependencies of access protocols. For better comprehension, we report in Table 2 all the parameters that are used in the following.

A. TRAFFIC MODELING

We consider uplink transmission for a cellular network consisting of a single BS and n MTC devices. We assume Rayleigh fading where channel gains follow a distribution characterized by exponential randomness. We use a power-law path-loss model, and devices use full path-loss inversion power control to address the "near-far" problem, adjusting signal power based on downlink path-loss estimation to maintain a consistent received signal power at the BS, meeting a threshold ρ. We assume that in each time slot, packets that failed to transmit will retry in subsequent time slots, even if new packets arrive during those times. Each device maintains a queue for packet transmission, managed by the arrival of new packets and the presence of undelivered ones. We use a First Come First Serve (FCFS) scheduling scheme, where newly arrived packets are placed at the end of the queue. For simplicity, we assume each device has a sufficiently large buffer size, allowing for an infinite number of RACH attempts, ensuring no packet is dropped until successfully received by the BS.

For packet generation, we consider a sporadic traffic model where the arrival rate, denoted as λ, is a random variable uniformly distributed between 0 and $\lambda_{\max}$ ($\lambda \sim$ Uniform$(0, \lambda_{\max})$). This approach reflects the unpredictability of real-world scenarios, allowing the arrival rate to vary dynamically within a defined range. In each time slot, a device is deemed active if it has non-empty buffers, which is defined as $N_t^{\text{active}} = N_t^{\text{new}} + N_t^{\text{cum}} > 0$. Here, N_t^{new} represents the number of newly arrived packets, modeled as a Poisson random variable with an intensity of λ. Specifically, N_t^{new} follows a Poisson distribution defined by the probability mass function (PMF):

$$P(N_t^{\text{new}} = n) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n = 0, 1, 2, \ldots \tag{1}$$

The term N_t^{cum} accounts for previously failed packets that were not successfully transmitted in prior time slots. Thus, the total number of packets in the buffer can be influenced by both the sporadic nature of new arrivals and the accumulation of previously failed attempts.

To express the combined effect of new and cumulative packets mathematically, we define the PMF of the number of active devices in each time slot as follows:

$$P(N_t^{\text{active}} = k) = \sum_{n=0}^{\infty} P(N_t^{\text{new}} = n) \cdot P(N_t^{\text{cum}} = k - n) \tag{2}$$

where $P(N_t^{\text{new}} = n)$ follows the equation (1), and $P(N_t^{\text{cum}} = m)$ represents the distribution of accumulated packets from previous time slots.

B. ACCESS MECHANISMS

The BS has a preamble pool of ξ non-dedicated preambles known to the devices, each chosen with equal probability. We focus on Small Data Transmission RA (SDT RA) proposed by 3GPP [39] for both grant-based and grant-free access, as depicted in Figure 1. In the grant-based approach, devices transmit data with $Msg3$ without transitioning to the Radio Resource Control (RRC) Connected state. Devices initiate the process by selecting a preamble and transmitting it as $Msg1$ on the RA subframe. If the preamble is successfully detected, the BS grants a larger Physical Uplink Shared Channel (PUSCH) resource for both $Msg3$ and data transmission. Assuming the packet size is smaller than the maximum Transport Block Size (TBS), the data can be sent along with $Msg3$ in one slot. In the last step, the base station sends an acknowledge in $Msg4$. In the grant-free method, devices send data in $MsgA$, comprising both preamble and PUSCH data, transmitted separately over time with independent channels. If there is no collision during preamble or PUSCH

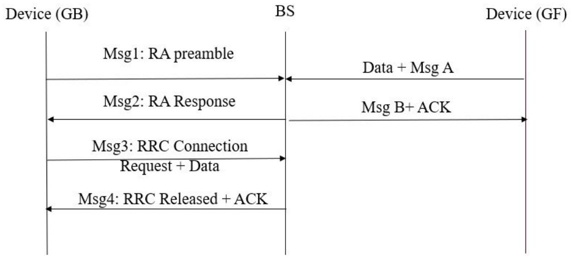


FIGURE 1. Grant-based and grant-free SDT RA schemes.

TABLE 1. Summary of related works.

Work	Methodology	Objectives	Use Cases	Access Type
[8]	ACB, Data Aggregation	Maximize number of successful MTDs and successful preamble utilization	mMTC	Grant-based
[9]	ACB, LSTM NN	Maximize accuracy for traffic prediction and alleviate congestion		
[10]	ACB, Q-learning	Maximize the access control of simultaneous H2H and M2M and reduce congestion		
[11]	ACB, Resource Allocation	Harmonize the access and transmission, optimize resource allocation		
[12]	Data Aggregation, Schedule Resource	Minimize outage probability		
[14]	Ordentlich Polyanskiy's Grantless	Minimize energy consumption and the collision overhead		
[15]	Physical-layer Transceiver Scheme	Minimize delay and high reliability		
[16]	Time-frames Separation	Maximize Energy efficiency, Spectrum efficiency and minimize delay		
[17]	Spatial-group-based	Minimize collision probability and access delay	URLLC	
[18]	Configured Grant-based	Minimize miss detection		
[19]	Distributed Layered	Maximize throughput	mMTC	Grant-free
[20]	Priority-based and Markov Chain	Maximize slot utilization and enhance traffic estimation		
[21]	Applying NOMA to ALOHA	Maximize throughput		
[22]	Number of Preamble-based	Minimize collision and number of failed users		
[23]	Number of Preamble-based	Maximize user success, minimizing retransmissions, latency, and energy consumption		
[24]	Number of Preamble-based	Maximize Preamble detection		
[25]	DNN	Minimize error rate of active device detection		
[26]	Data Length Diversity	Maximize activity level and channel estimation, active device and data detection		
[27]	BiGAMP*			
[28]	Bidirectional LTSM			
[29]	DRL	Maximize Channel estimation and data recovery performance		
[30]	DRL	Maximize successful access probability		
[31]	DRL	Maximize number of successful UEs	URLLC	
[32]	DRL	Minimize resource wastage and latency		
[33]	DRL	Maximize throughput	mMTC	
[34]	Resource Sharing	Maximize achievable rate	N/A	Semi-grant-free
[35]	Cumulative Distribution Scheduling	Maximize data rate and enhance outage performance	mMTC	
[36]	Lagrange Multipliers Subgradient	Maximize throughput		
[37]	DDQN	Maximize throughput and optimize power levels		
[38]	MA-DRL			

*BiGAMP: Bilinear Generalized Approximate Message Passing

decoding, the base station acknowledges with a message. Early preamble collision detection [40] is considered in grant-based access, it allows the base station to identify collisions when multiple devices select the same preamble during the initial message ($Msg1$). This mechanism works by detecting overlapping preambles, allowing the base station to recognize that multiple devices are attempting to access the network simultaneously, thereby facilitating prompt collision resolution. In our system modeling, we incorporate NOMA for grant-free access, allowing multiple devices to share the same resource. To manage interference, the Base Station employs Successive Interference Cancellation

TABLE 2. Parameters list.

Parameter	Description
ρ	Received Power
$\lambda_{\max}$	Maximum arrival rate
λ	Arrival rate in a specific time slot
N_t^{active}	Total number of active devices in the tth time slot
N^t_{New}	Number of newly arrived packets in the tth time slot
N^t_{Cum}	Number of accumulated packets in tth time slot that failed in earlier time slots
ξ	Number of Preambles
h_0	Channel from a typical device to the BS
I_{intra}	Aggregated intra-cell interference
σ_n^2	Noise power
B_i	Battery level of device i
E^t	Energy consumption in time slot t
E^m_{eff}	Energy efficiency during m time slots
S^m	Throughput during m time slots
T_s	Time slot length
N_{RAR}	Number of slots that RAR window occupies
N_{CRT}	Number of slots that CRT occupies
T_K	PUSCH scheduling parameter defined in the standards
T_D	Number of slots the D-th device sleeps before its data sending slot
P_s	Power consumption in the sleep state
P_r	Power consumption in receiving state
P_t	Power consumption in transmitting state
γ_{th}	SINR threshold for SIC decoding
a_1	Action of sending based on GF
a_2	Action of sending based on GB
S	Set of states used for DRL
A	Set of actions used for DRL
p	Transition probability from a state $s \in S$ to a state $s' \in S$
r	Reward received after taking an action
θ	Weights matrix of a multiple layers DNN
γ	Discount factor
η	Learning rate

(SIC). SIC decodes the strongest signal, subtracting it, and iteratively decoding the subsequent strongest signals. This iterative process enables the recovery of weaker signals based on the Signal to Interference and Noise Ratio (SINR), facilitating the retrieval of multiple packets. The SINR for PUSCH transmission in the tth time slot is formulated as [41]:

$$SINR^t = \rho(|h_0|)^2/(I^t_{intra} + \sigma_n^2) \tag{3}$$

where ρ is the full path-loss inversion power control threshold, h_0 is the channel from a typical device to the BS, σ_n^2 is the noise power and I_{intra} is the aggregated intra-cell interference, which is the interference generated by other users within the same cell. The SIC decoding process follows the descending order of received power. Given that the power control is constant, we have $|h_1|^2 > |h_2|^2 > \ldots > |h_n|^2$. Thus, the interference $\mathcal{I}_{\text{intra}}$ can be expressed as:

$$\mathcal{I}_{\text{intra}} = \sum_{j=1}^{n} 1_{\left\{N^t_{\text{New}_j}+N^t_{\text{Cum}_j}>0\right\}} \rho \left|h_j\right|^2 . \tag{4}$$

The above equation implies that the strongest signal is decoded first, and its impact on the overall interference is eliminated before decoding the next signal. Consequently, the remaining interference for the subsequent signals is reduced. In our system modeling, we utilize the downlink channel typically employed by the base station for granting access and sending acknowledgments. This channel is used to broadcast notifications to devices, ensuring they receive information about any changes in their random access method.

C. PROBLEM FORMULATION

In the context of mMTC, accommodating a large number of devices accessing the network with sporadic traffic presents a significant challenge. All the devices execute the same random access procedure during a given time slot by selecting one of the available RA methods. In this study, considering the battery-powered nature of MTC devices, we operate under the constraint that none of the devices have their batteries depleted, as

$$\forall i: \quad B_i \geq \epsilon \tag{5}$$

where B_i represents the battery level of device i and ϵ represents a very small positive value. Metrics such as throughput and energy efficiency are crucial for enhancing the performance of short-packet transmission, especially for battery-driven MTC devices in massive environments, providing valuable guidance for optimization endeavors. Our goal is to concurrently enhance these two metrics. This approach directs our optimization strategies towards achieving a balance between maximizing data throughput and enhancing energy efficiency within the MTC environment. An outline of the energy efficiency and throughput for each access type is further described.

1) ENERGY EFFICIENCY

It is calculated as the ratio of the number of packets successfully transmitted within the network to the total energy consumed during m time slot. It is expressed as

$$E^m_{eff} = \frac{\sum_{t=1}^{m} \mathcal{P}^t_{succ}}{\sum_{t=1}^{m} E^t} \tag{6}$$

However, the number of successful transmissions and energy consumption varies across different access protocols. We will discuss the calculation method for each protocol.

a: GRANT-BASED

- P_{succ}: In our scenario we assume that early collision detection has been considered. It means that probability of successful transmission is the probability of selecting a unique preamble [42] of active devices in equation (2). Suppose that all n active devices randomly choose preambles in $\{1, 2, \ldots, \xi\}$, therefore, the probability of the number of successful transmission in each time slot is

$$P_{succ} = P(N_t^{\text{active}} = k) n \left(1 - \frac{1}{\xi}\right)^{n-1} \tag{7}$$

- E: For ease of description, we illustrate the energy consumption of each packet transmission in Figure 2. As depicted, there are two devices that are trying to send date to the base station. T_s is the slot time length which is time duration for sending preamble in $Msg1$ N_{RAR} is the number of slots that Random Access Response (RAR) window occupies. N_{CRT} is the number of slots that Contention Resolution Timer (CRT) occupies. T_K is the PUSCH scheduling parameter defined in the standards [43] and T_D is the time that Dth device sleeps until its data sending slot. P_s,P_r are the power consumption when the device is in the sleep and receiving states separately, which are constants for all devices. The transmit power, P_t is considered a constant because, despite each device's radiated power depending on its distance to the BS due to full-path power control, the power consumed by the amplifiers and Radio Frequency (RF) hardware is significantly higher and nearly independent of the radiated power [44]. Therefore, the energy consumption of successful transmission can be written as:

$$E = P_t T_s + P_r N_{RAR} + (i-1) T_K P_s + N_{CRT} P_r \tag{8}$$

where i is the number of devices that successfully transmit a preamble and are granted allocation for $Msg3$. In equation $(i-1)$, the first device transmits data without queuing, while the remaining devices must wait for their turn.

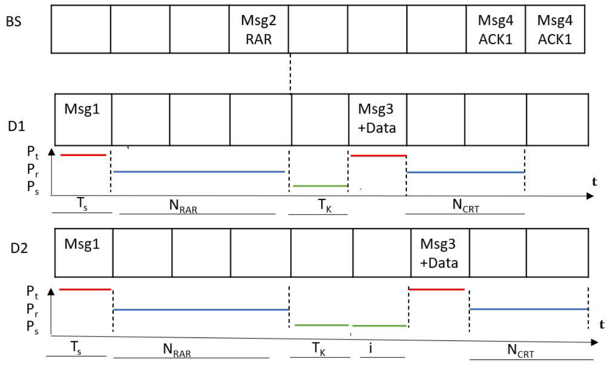


FIGURE 2. Timing relationship of Grant-based RA procedure.

b: GRANT-FREE

- P_{succ}: In grant-free access protocols, a packet's success depends on both the probability of success during the preamble phase and the SINR during the data phase. As a result, the number of successful transmission in each time slot is

$$\mathcal{P}_{succ} = P(N_t^{\text{active}} = k) n \left(1 - \frac{1}{\xi}\right)^{n-1} \times (\mathbb{1}_{(\rho(|h_0|)^2/(I^t_{intra}+\sigma_n^2))>\gamma_{th}}) \tag{9}$$

- E: In grant free, once devices have transmitted the preamble, they send data in another slot. Then they wait for Ack from base station at RAR windows. If their transmission is successful they receive Ack from base station. As shown in Figure 3, devices send preamble and data in each state of failure and success. Therefore energy consumption is constant as

$$E = (2 \times P_t T_s) + P_r N_{RAR} \tag{10}$$

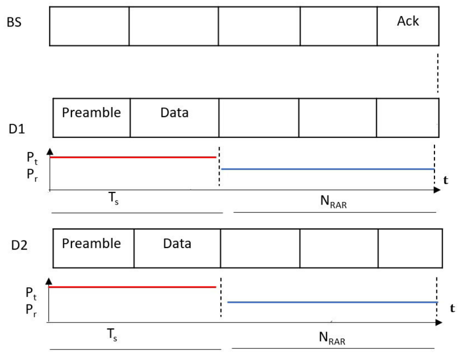


FIGURE 3. Timing relationship of Grant-free RA procedure.

2) THROUGHPUT

Throughput, denoted by S, can be derived by the number of successfully transmitted packets during a period in the mth time slot, as shown below.

$$S^m = \frac{\sum_{t=1}^{m} \mathcal{P}^t_{succ}}{\sum_{t=1}^{m} T^t} \tag{11}$$

where T can be determined as follows by referring to figures 2,3 for different access protocols.

a: GRANT-BASED

$$T = T_s + N_{RAR} + (i-1) T_K + N_{CRT} \tag{12}$$

b: GRANT-FREE

$$T = T_{data} = (2 \times T_s) + N_{RAR} \tag{13}$$

Referring to equations (10) and (13), while grant-free protocol exhibits lower energy consumption and time delay, based on the (9), its success probability decreases with increasing interference from other devices due to SINR effects. Optimizing both energy efficiency and throughput entails choosing the suitable random access method, especially in an environment marked by sporadic packet arrivals and fluctuating rates within each time slot. These irregularities in packet arrival rates present a significant challenge, complicating the ability to identify and track consistent trends within the network's operational patterns. In our system, the BS operates in an environment where it has incomplete information about the current network state, such as the number of active devices in a given time slot or the real-time channel conditions. This lack of full observability requires the BS to make decisions based on partial observations of the system, making the problem well-suited to be modeled as a POMDP. The state of the system in each time slot is derived from the previous time slot. Based on this historical information, the BS estimates the likely number of active devices in the upcoming slot and selects the appropriate access protocol. This process is essential for maximizing long-term throughput and energy efficiency. We emphasize long-term because the decision made in the current time slot can affect future time slots, as failed devices retry the same data in the next time slot, leading to buffer extension. As demonstrated earlier, energy efficiency and throughput are influenced by access protocols. As a result, our primary goal is to maximize both energy efficiency and throughput by identifying the most suitable actions for all devices. This objective can be formulated as:

$$\begin{aligned} &\max_{\{a_1,a_2\}} E^m_{eff}, S^m\,(a_1, a_2) \\ &\text{s.t. constraint(5).} \end{aligned} \tag{14}$$

where a_1 and a_2 are the action of sending based on GF and GB, respectively.

IV. DEEP REINFORCEMENT LEARNING-BASED MULTI-ACCESS

A POMDP model is represented by a tuple (S, A, p, r, O), where S denotes the set of states, A represents the set of actions, p is the transition probability from a state $s \in S$ to a state $s' \in S$ after executing action $a \in A$, r is the reward received after taking action a, and O includes both the observation set and the observation probability set. At each time step, the system is in a particular state s. After the agent selects an action a, it receives an observation o with a probability $O(o|s, a, s')$. Subsequently, the system transitions to a new state s' and provides a reward r to the agent.

To address the POMDP problem, we employ DRL. This method, highlighted for its capabilities in solving intricate sequential decision-making challenges, has showcased exemplary performance, even when dealing with partially observable environments. When the state and action spaces are large, RL algorithms can become computationally expensive and memory-intensive, making it challenging to converge to an optimal solution. To address this issue, DQN was introduced, combining Q-learning with Deep Neural Networks (DNNs) to efficiently train an accurate state-action value function for problems with high-dimensional state spaces. In this section, we propose employing a DQN-based algorithm to address the problem described in (eq 14). The rationale for adopting DQN within our framework stems from multiple considerations. One key reason is the demonstrated success of DNNs in addressing partially observable problems through function approximation. Given the continuous and dynamic nature of our state space, characterized by a wide range of numerical values due to the varying number of active devices in each time slot, the state space becomes too complex for a Q-table to handle effectively. DNNs, however, offer the flexibility and capacity to discern intricate patterns from the observed states, facilitating robust learning in such complex environments [31]. As our environment evolves dynamically, DQN's adaptive nature ensures its applicability across varied problem scenarios, making it a compelling choice for our research objectives.

A. RL FRAMEWORK

In our RL framework, to accurately calculate the energy consumption of devices in a real environment, an RL-agent is deployed at the BS to interact with the environment and progressively choose appropriate actions. However, this increases the power consumption and operational costs of the BS, which is assumed to be plugged in and has a constant power supply. An RL agent is defined by the tuple (S, A, p, r), where S is the set of states, A is the set of actions, p is a transition probability from state $s \in S$ to state $s' \in S$ after taking action $a \in A$ and r is the reward after taking action a. In each step, the system is at state s. After the agent takes an action a, transitions to a new state s', receiving a reward r. Based on our original problem, we define the action, state, and reward as follows:

- Action Space: The action space consists of an array of two elements, $[A1, A2]$, representing grant-based and grant-free actions, respectively. However, only one type of access protocol is used in each time slot. Devices must follow the random access protocol broadcasted by the base station and cannot perform both actions at the same time. They continue using their assigned access method until they receive a new random access alert from the base station, indicating a change in protocol.
- State Space: The state space is defined as an array encompassing three elements $[S_{devices}, S_{failure}, S_{energy}]$. $S_{devices}$ is the count of devices that both executed an action and had packets queued in their buffer during the preceding time slot. $S_{failure}$ represents the count of devices among the active ones in the last time slot whose packet transmission attempts ended in failure. S_{energy} encapsulates the total energy consumed during the previous time slot. These specific states serve as indicators, furnishing the agent with information to inform its decision-making process for subsequent time slots.
- Reward Function: A weighted summation function, $r(s, a)$, is employed to compute the reward for each configuration, defined as

$$r(s, a) = w_1.E^m_{eff} + (1 - w_1).S^m \tag{15}$$

where, E^m_{eff} represents the energy efficiency as (6), and S^m stands for throughput as (11). Furthermore, w_1 serve as weighting factors for each variable, allowing flexibility based on the specific requirements of the application. In our approach, we consider $w_1 = 0.5$ to treat them equally, providing a balanced optimization perspective.

The POMDP framework allows the agent to map partial observations (states) to optimal actions, such as selecting the most appropriate access protocol for the current time slot. The selection of an access protocol (grant-based or grant-free) influences the network's future state by reducing the number of queued packets and, thus, the number of active devices in subsequent time slots. This approach maximizes long-term throughput and energy efficiency, aligning with the goals of our optimization problem.

B. RL TRAINING

The DNN architecture of proposed DQN (Figure 4) incorporates a single neural network comprising of input layer, a LSTM layer, three Fully Connected (FC), and an output layer. The input sequence layer comprises state elements in the current time slot (t).

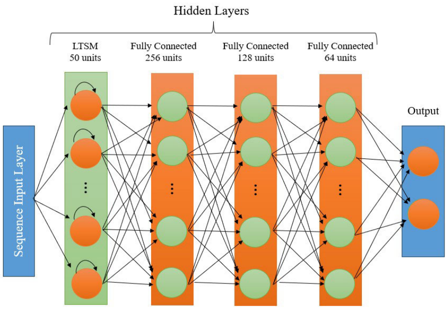


FIGURE 4. Architecture of the neural network used for the implementation of the DRLMA algorithm.

In the sporadic environment, the base station receives partial observations, leading to uncoordinated scenarios where decisions are made without information exchange. Traditional layers in neural networks struggle in these scenarios due to limited knowledge of the contention status from current network inputs. The LSTM layer, with its ability to retain and process information from previous time steps, is particularly well-suited for handling sequences of states in environments where the state is partially observable. This capability is crucial for making informed decisions based on historical context, which is often missing in traditional neural network layers. By integrating the LSTM layer with FC layers, the network not only captures these temporal dependencies but also refines and transforms the extracted features into actionable insights. This combination allows the network to effectively learn about the contention status in the system and generate accurate resource access probabilities, thereby reducing the likelihood of collisions. Here, the Q-function input encompasses both the current state and historical state, expressed as $s(t) = \{s(t) \cup H(t)\}$.

The choice of three FC layers balances the model's ability to capture complex patterns with computational efficiency. This structure allows for progressive refinement and abstraction of features from the LSTM, enabling the network to capture hierarchical relationships essential for evaluating action trade-offs. With the output dimensionality from the LSTM being manageable, three FC layers effectively reduce and transform this dimensionality into Q-values for binary actions (grant-based and grant-free) determining the optimal action.

During training, the DNN is updated using experiences collected by the RL agent, represented as tuples (S, A, r, S'), where S is the current state, A is the action taken, r is the reward received, and S' is the next state. These experiences are stored in an experience replay buffer and sampled to train the DNN. The DNN approximates the Q-value function $Q(S, A; \theta)$, where θ represents the weights matrix of a multiple layers DNN. Hence keeping a large storage space for state-actions pair (Q-values), DRL agent only memorize θ weights in its local memory that reduces the memory and computation complexity. The Q-value function is defined as:

$$Q(S, A; \theta) = \mathbb{E}\left[r + \gamma \max_{A'} Q(S', A'; \theta^{-}) \mid S, A \right]. \tag{16}$$

This equation computes the expected cumulative reward r when taking action A in state S, followed by selecting the optimal future action A' in the next state S', with γ as the discount factor. The term θ^{-} represents the parameters of the target network, which helps stabilize the training process [45].

The DNN is trained by minimizing the loss function, which quantifies the difference between the predicted Q-values $Q(S, A; \theta)$ and the target Q-values in eq (16). The loss function is given by:

$$L(\theta) = \mathbb{E}\left[\left(r + \gamma \max_{A'} Q(S', A'; \theta^{-}) - Q(S, A; \theta) \right)^2 \right]. \tag{17}$$

Loss function measures how far the DNN's current predictions are from the target Q-values. By iteratively minimizing this loss, the DNN updates its parameters θ, improving its ability to accurately estimate Q-values and make better action decisions. This process allows the DQN to progressively refine its policy through the agent's interactions with the environment, optimizing long-term rewards.

When the agent selects an action with a higher Q-value, it is anticipated that the environment will encounter fewer collisions, leading to the agent receiving increased rewards as a result.

C. RL ALGORITHM

Similar to [31], we consider online training and the process of DQN is summarized in Algorithm 1. In the initialization procedure, several training parameters are configured, including the discount factor γ, batch size $\mathbf{N_b}$, ϵ-greedy probability ϵ, copy frequency of network weights T_s, T_p for smooth and periodic respectively, and experience replay size. In the beginning of each episode, all buffers are set to 0 and the battery capacities reset to full capacity. In each tth slot, buffers are allocated with 0 or 1 packet. Value of 0 indicates that there is no packet in the buffer, while value of 1 indicates that there is a packet in the buffer ready for sending. Agent inputs current state $s(t)$ into the primary Q-network and obtains Q-values for all actions. The action a(t) is determined using ϵ-greedy policy and the agent collects states $s(t)$ and receives reward $r(t+1)$ from the environment. Following this, the environment transitions to a new state denoted as $s(t+1)$. The agent then generates a new training data point $(s(t), a(t), r(t+1), s(t+1))$ based on these observations and stores it in its memory. We consider the application of minibatch training, instead of a single sample, for training the primary Q-network which improves the convergent reliability. Therefore, the expectation is taken over the N_b minibatch randomly selected from the experience replay. In each slot, the primary Q-network is trained with the gradient descent method and using RMSProp optimizer [31]. Periodic smoothing was employed for the target network in this study. This technique involved gradual updates of the target network's weights towards the Q-network's weights, contributing to enhanced stability and convergence during the training process.

Algorithm 1 DLRMA Training Process

```
Initialize training parameters $\gamma, \epsilon, N_b, T_s, T_p$ and experience replay size;
Initialize primary Q-network and target Q-network with random weights;
for episode = 1 to MaxEpisodes do
    Reset environment, device battery and device buffer;
    while All (battery > 0) do
        Generate random packets for each device [0, 1];
        BS choose an action using $\epsilon$-greedy and broadcasts it to the devices;
        BS observes the number of failures and energy consumption based on active devices in $S_{t+1}$;
        Calculate reward based on Eq.15;
        Store transmission $(S_t, a_t, r_{t-1}, S(t+1)$ in the memory;
        Sample random $\mathbf{N_b}$ mini-batch of transmission $(S_t, a_t, r_{t-1}, S(t+1))$;
        Perform gradient descent step and update primary network using RMSProp optimizer;
        Update target network periodically $T_p$ with smoothing $T_s$ and copy primary network weights to target Q-network weights;
    end while
end for
```

D. COMPUTATIONAL AND TRAINING COMPLEXITY

The computational complexity of the proposed DRL algorithm can be expressed in two distinct aspects: the complexity associated with the model architecture and the complexity related to the training process. The computational complexity is primarily a function of the number of connections through the deep neural network, defined as $O(U)$, where U is given by $U = Ku_1 + \sum_{g=1}^{G-1} u_g u_{g+1} + Mu_G$ [46]. Here, K is the size of the input layer, equivalent to the length of the state, M is the size of the output corresponding to the length of the action set, and u_g is the number of neurons in the g-th layer. The computational complexity is a linear function directly proportional to the number of connections, highlighting the need for a well-structured network architecture to manage computational demands effectively. Furthermore, the training complexity is determined by the number of agents, episodes, and time steps involved in the learning process. For N agents during one mini-batch of E episodes and F time-steps until convergence results in computational complexity of order, the training complexity can be expressed as $O(NEFU)$. This implies that the training complexity grows linearly with the number of agents and the number of time-steps, indicating the importance of efficient training strategies. However, our DRL model mitigates these concerns by using three hidden layers with decreasing neurons, balancing expressiveness and computational efficiency. With just two actions as output and three states as input, the complexity is further reduced. Additionally, employing a single agent for network-wide decisions limits agent interactions, facilitating efficient learning and ensuring scalability can be managed effectively.

The computational efforts required for decision-making and inference in production remain manageable even as the number of connected devices scales up. This stability is due to our model's use of an aggregated state representation–$[S_{devices}, S_{failure}, S_{energy}]$–which condenses key information across all devices. Instead of tracking each device individually, the base station processes only the total number of active devices, overall energy consumption, and total failed attempts within each time slot. This approach means that regardless of how many devices are present, the system operates on these three collective metrics, ensuring that computational complexity does not increase with more devices. This design choice allows the model to efficiently support massive IoT environments without escalating inference demands.

V. EXPERIMENTAL STUDY

The proposed approach was evaluated using a MATLAB simulation environment with comparisons with random GF, traditional GB, the DSA method [6] and the fallback method mentioned in [39]. In DSA method, a threshold is determined based on the arrival rates for switching from grant-free to grant-based. The fallback method involves devices switching to grant-based access by receiving a grant from the base station if they send a successful preamble but experience data transmission failure. The impact of device number, maximum arrival rate on the performance of DRLMA, as well as on system throughput, energy efficiency and battery lifetime is investigated.

A. SYSTEM HYPER-PARAMETERS AND SIMULATION SETUP

The simulations consider the effect of the frequency band, assuming it only impacts the duration of the slot. Channel inversion power control is employed by each device, and a buffer is simulated on each device to track the arrival and accumulation of new packets over time.

TABLE 3. Simulation parameters.

Parameter	Value
ρ	-90 dBm
σ_n^2	-100.4 dBm
γ_{th}	-10 dB
T_s	0.5 ms
N_{RAR}	40
N_{CRT}	48
T_k	0.5 ms
P_s	15 uW
P_r	80 mW
P_t	500 mW
Full battery capacity	250J [47]

The parameter values used are listed in Table 3 but in the training phase, to make the training feasible and reduce complexity, we assume an initial training battery level of 5J. As learning parameters we trained agent in 10000 episodes and the learning rate, discount factor and batch size are $\eta = 0.001$, $\gamma = 0.95$, $N_b = 32$, respectively. The ϵ-greedy probability ϵ decreases from 1 to 0.001.

To account for the diverse nature of the mMTC environment, we took into consideration varying arrival rates that are not consistent across different time slots. Furthermore, to ensure the robustness of our findings, we conducted five independent runs, each initialized with different random seeds. These repetitions contribute to the reliability and generalizability of our results. It is noteworthy that the standard deviation associated with these simulations is very small, signifying the stability and consistency of our outcomes.

B. CONVERGENCE OF THE DLRMA TRAINING

Figure 5 illustrates the training process of the agent, depicting the system convergence of the proposed DRL learning framework through a plot of the average received reward (dark blue) based on (15). In the initial training phase, reward increases from 1 to almost 2.5 and the agent learns lots of experiences in action selection, and begins to exploit the potential better action selection probability that can improve the reward. After that, the agent learns from good memory samples resulting from the new action selection probability.

Figure 5 presents the convergence behavior of our DRL model during the training process. The rapid convergence observed, with the system reaching satisfactory performance in as few as 200 episodes, illustrates the effectiveness of our complexity-reducing strategies. This visualization highlights the low computational burden of our proposed scheme, making it suitable for an agent with limited computing resources.

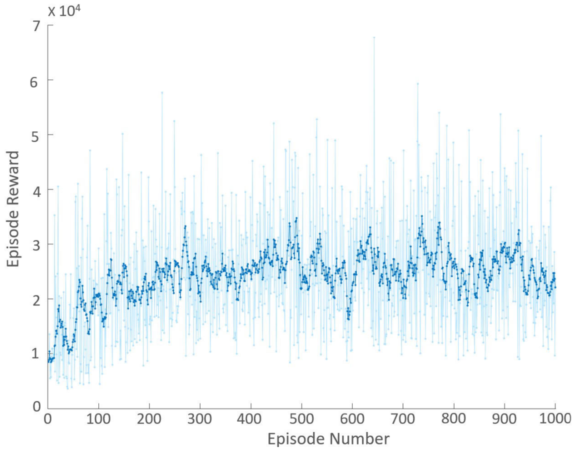


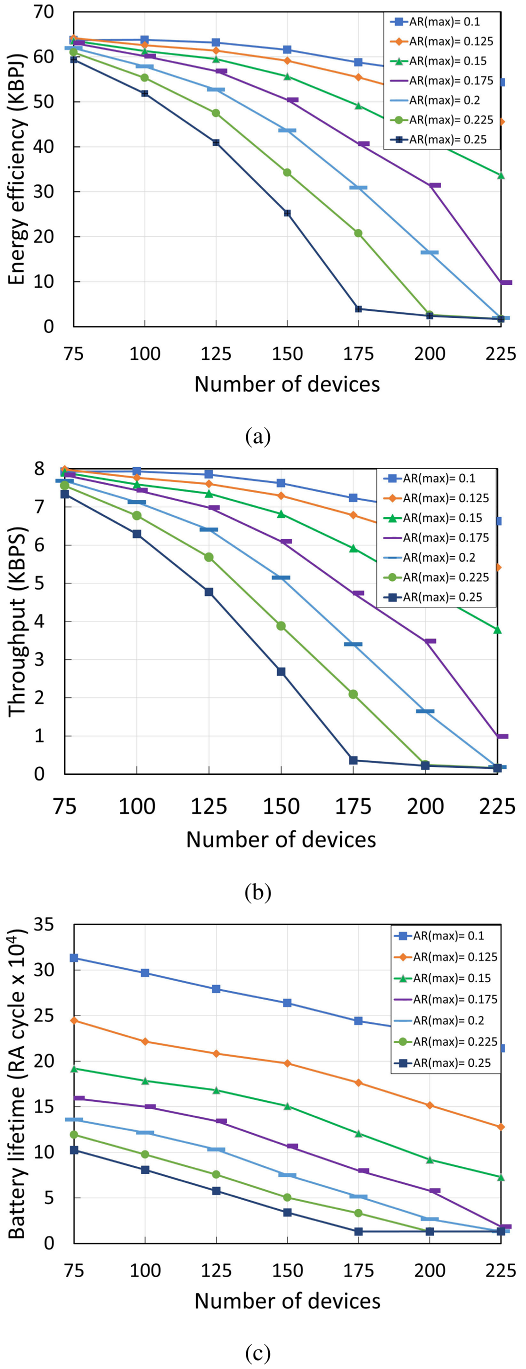
FIGURE 5. Training process of the DRLMA agent. Light blue indicates the reward for each episode, the dark blue shows the average reward over the last 10 episodes.

C. PERFORMANCE OF DRLMA

We conduct a separate analysis focusing on the variation in arrival rate of packets with determining the maximum boundary specific to individual numbers of devices. This method aims to provide insights into device-specific behaviors, allowing for a more granular understanding of performance factors and variations in the system.

Figure 6 illustrates the performance of the proposed DRLMA across different arrival rate boundaries based on the number of devices, focusing on energy efficiency, throughput, and battery lifetime. When the number of arrival rate boundaries increases, the number of active data also rises. Based on Figures 6a and 6b, when there are 75 devices, the energy efficiency and throughput are nearly the same. However, as the number of active devices increases with higher arrival rate boundaries, the difference in performance becomes more pronounced. This is due to the rise in collisions, which reduces the number of successful transmissions and increases both energy and time consumption. Consequently, based on equations (6) and (11), the performance of energy efficiency and throughput declines. For battery lifetime, the scenario differs slightly; here, energy consumption directly impacts the number of RA cycles. As energy consumption rises with the increase in active data due to higher arrival rates, there is an initial difference between the various arrival rates. After reaching a specific device number, such as 175, the number of active devices becomes significantly higher, leading to a substantial increase in collisions. As a result, the performance of throughput and energy efficiency approaches zero.

This behavior is typical for all access protocols as the number of active data increases. In the following sections, we compare our proposed DRLMA with other access protocols based on the number of devices and arrival rate boundaries separately.

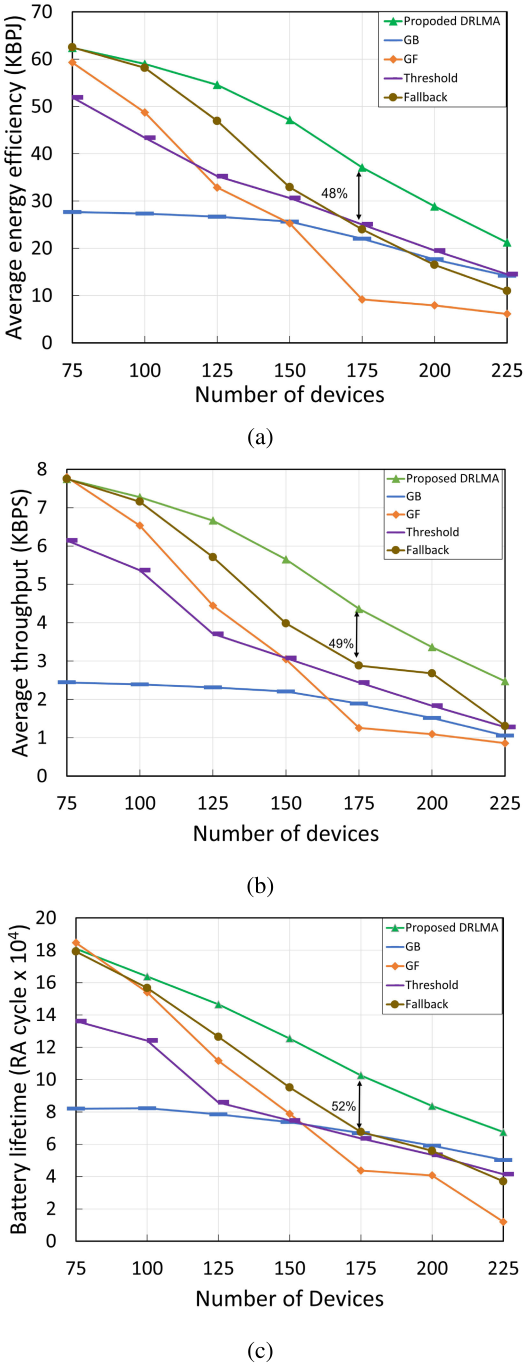


(a)

(b)

(c)

FIGURE 6. Performance of DRLMA based on number of devices for different maximum arrival rates.



(a)

(b)

(c)

FIGURE 7. Comparison of different access protocols based of number of devices for (a) energy efficiency, (b) throughput and (c) battery lifetime.

D. IMPACT OF THE NUMBER OF DEVICES

In this section we compare our proposed DRLMA with pure grant-free and grant-based as well as other access protocols in the literature. Figure 7 illustrate the average energy efficiency, average throughput and battery lifetime concerning the number of devices. we consider a sporadic traffic which there are not a constant arrival rate in each RA cycle. Different packet arrival rates are mentioned in Figure 6, and we calculate the average of these rates for each number of devices.

Based on the figures, with fewer devices, the grant-free, fallback, and DRLMA methods outperform the grant-based and threshold approaches. This is because, with fewer devices, collisions are less frequent, and grant-free methods can manage successful transmissions with low energy consumption as shown in equations (10) and (13). However, as the number of devices increases, the performance of grant-free methods declines due to heightened collision rates in an uncoordinated environment. The threshold method shows improved performance with an increasing number of devices by switching from grant-free to grant-based access. However, due to sporadic traffic, it cannot switch back to grant-free based on the traffic state, so its performance remains below that of DRLMA.

Fallback also performs better than some other protocols but declines as the number of devices increases because in the network state of high active data, it initially attempts grant-free access and then switches to grant-based, causing a waste of time and energy. As seen from figure 7a, DRLMA shows a 48% gain in energy efficiency, from figure 7b a 49% gain in throughput, and from figure 7c a 52% gain in battery lifetime.

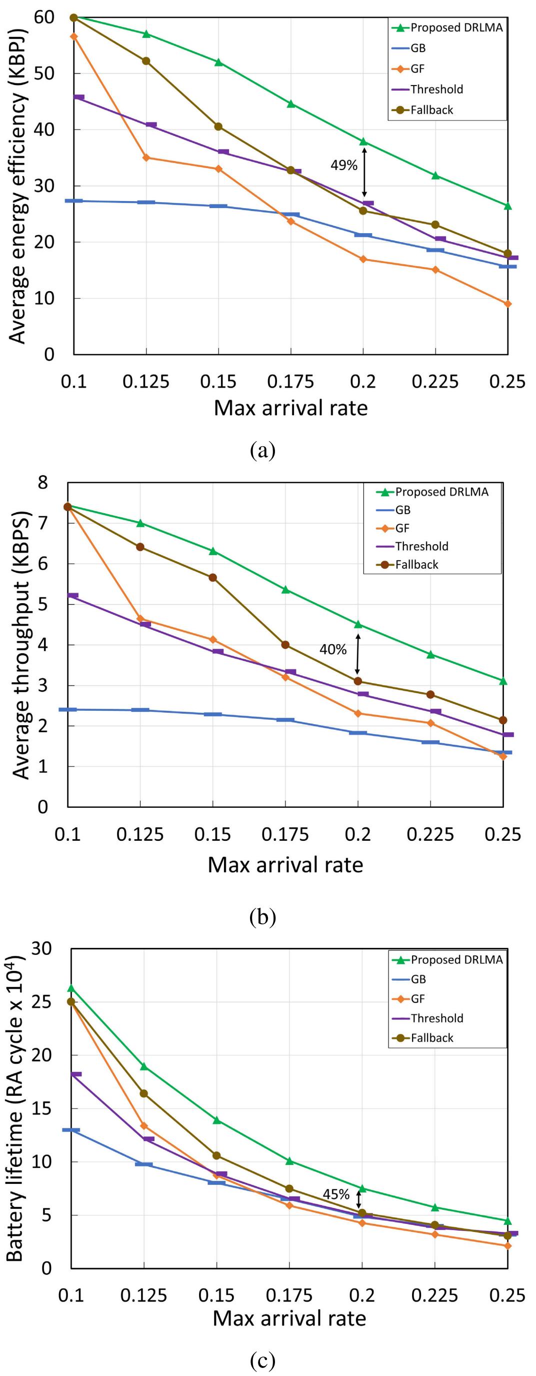
Hence, DRLMA demonstrates prowess in swiftly learning and seamlessly adapting between access protocols, responding dynamically to the ever-changing network status. Our method surpasses other access protocols, affirming its selection of a suitable access type. This not only prevents network congestion but also optimizes resource efficiency, leading to noticeable improvements in energy efficiency, throughput, and battery lifetime.

E. IMPACT OF ARRIVAL RATE BOUNDARIES

We analyzed our results based on the arrival rate diversities for a specific number of devices to focus solely on the variety in the number of active devices.

Figure 8 illustrates the impact of different arrival rates on energy efficiency, throughput, and battery lifetime by comparing different access protocols. To account for variations in the number of devices, we considered all device numbers shown in Figure 7. We calculated values for each device number based on different arrival rate boundaries separately and then averaged the results. It is worth mentioning that the arrival rates in each time slot are not constant and vary to meet sporadic traffic, with boundaries set to a maximum arrival rate corresponding to active devices. As arrival rates increase, the number of active devices also rises, leading to increased congestion. This causes a sharp decline in the performance of grant-free and fallback methods. Another contributing factor to this decline is that, with each failed transmission, the buffer extends, increasing access requests. The threshold method also experiences a performance decrease, though not as severely as grant-free and fallback methods, due to its switch to grant-based access after reaching a threshold. Grant-based access maintains almost constant performance as it manages access by allocating resources for a specific number of active devices.

Based on Figures 8a and 8b, DRLMA shows a gain of 49% in energy efficiency and 40% in throughput. These gains are significant as they indicate that DRLMA can effectively manage the transmission power and resource allocation, leading to more efficient use of energy and higher data transmission rates, which are crucial in scenarios with a high number of devices and limited power resources. As shown in Figure 8c, in terms of battery lifetime, grant-free initially performs better due to lower energy consumption. However, as the arrival rate boundary increases, our proposed DRLMA outperforms the other access protocols with a gain of 45%. This improvement is particularly important in massive MTC environments, where extending the battery life of devices is critical to maintaining long-term, reliable operation without frequent battery replacements.



(a)

(b)

(c)

FIGURE 8. Comparison of different access protocols based on the arrival rate boundaries for (a) energy efficiency, (b) throughput and (c) battery lifetime.

VI. CONCLUSION

Multiple access schemes are pivotal in facilitating extensive connectivity within mMTC and future network advancements, particularly given the sporadic traffic and battery-driven devices characteristic of mMTC. To mitigate collisions in massive and heterogeneous environments and enhance both throughput and energy efficiency, this paper introduces an intelligent RL-based access switching mechanism employing a DQN within a single neural network architecture. Initially, we model energy efficiency and throughput based on different access protocols to demonstrate these objectives as functions of the access protocols. Adopting a POMDP approach due to the irregular and uncoordinated nature of the vast environment, we address this through the proposed DRLMA algorithm. Comparative evaluations against pure grant-based, grant-free, threshold-based switching protocol and fallback reveal that our RL agent, which selects the appropriate access type based on varying arrival rates and device numbers, outperforms others. Results show that with a specific number of devices, throughput increases by up to 49%, allowing the network to handle more data efficiently. Energy efficiency improves by 48%, reducing power consumption and extending battery life for IoT devices. Additionally, a 52% increase in battery lifetime means devices can operate longer without frequent recharging, enhancing the sustainability and reliability of large-scale massive MTC deployments.

DRLMA framework is a novel model and in networks with high heterogeneity–where devices vary in power levels, high interference or traffic surges, our model's performance may be affected. We recognize these as potential limitations and opportunities for further development. Moving forward, we plan to integrate power level allocation and interference management techniques into the DRLMA framework. This enhancement aims to optimize access protocols and resource management, ultimately improving the performance and scalability of mMTC networks.

REFERENCES

[1] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2020.

[2] O. Liberg, M. Sundberg, E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of Things: Technologies, Standards, and Performance*. New York, NY, USA: Academic, 2017.

[3] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.

[4] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[5] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road toward 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1117–1174, 2nd Quart., 2022.

[6] N. Ravi, M. Curado, and E. Monteiro, "Dynamic switching access in massive machine type communication," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jul. 2024, pp. 1230–1235.

[7] *Study RAN Improvements for Machine-Type Communications*, document V11.0.0, Sep. 2011.

[8] J. Guo, S. Durrani, X. Zhou, and H. Yanikomeroglu, "Machine-type communication with random access and data aggregation: A stochastic geometry approach," in *Proc. Global Commun. Conf.*, Dec. 2017, pp. 1–7.

[9] H. L. D. Santos, J. H. I. de Souza, J. C. M. Filho, and T. Abrão, "LSTM-ACB-based random access for mixed traffic IoT networks," in *Proc. IEEE 8th World Forum Internet Things (WF-IoT)*, 2022, pp. 1–6.

[10] L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, "Reinforcement learning-based ACB in LTE-A networks for handling massive M2M and H2H communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[11] J. Li, Q. Du, L. Sun, and P. Ren, "Queue-aware joint ACB control and resource allocation for mMTC networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[12] T. Salam, W. U. Rehman, and X. Tao, "Cooperative data aggregation and dynamic resource allocation for massive machine type communication," *IEEE Access*, vol. 6, pp. 4145–4158, 2018.

[13] T. Salam, W. ur Rehman, R. Khan, I. Khan, and X. Tao, "Dynamic resource allocation and mobile aggregator selection in mission critical MTC networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 64–69.

[14] G. K. Facenda and D. Silva, "Efficient scheduling for the massive random access Gaussian channel," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7598–7609, Nov. 2020.

[15] J. Liu and X. Wang, "A grant-based random access scheme with low latency for mMTC in IoT networks," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 18211–18224, May 2023.

[16] M. M. Karchegani and B. S. Ghahfarokhi, "P-persistent massive random access mechanism for machine type communication," *Telecommun. Syst.*, vol. 78, no. 2, pp. 169–185, Oct. 2021.

[17] T. Kim, H. S. Jang, and D. K. Sung, "An enhanced random access scheme with spatial group based reusable preamble allocation in cellular M2M networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1714–1717, Oct. 2015.

[18] C. Liang, S. Xia, X. Han, and P. Hao, "Configured grant based URLLC enhancement for uplink transmissions," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 1053–1058.

[19] H. Jiang, Q. Cui, Y. Gu, X. Qin, X. Zhang, and X. Tao, "Distributed layered grant-free non-orthogonal multiple access for massive MTC," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–7.

[20] T. N. Weerasinghe, V. Casares-Giner, I. A. M. Balapuwaduge, and F. Y. Li, "Priority enabled grant-free access with dynamic slot allocation for heterogeneous mMTC traffic in 5G NR networks," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3192–3206, May 2021.

[21] J. Choi, "NOMA-based random access with multichannel Aloha," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.

[22] S. Jha, H. G. Srinath, and N. M. Balasubramanya, "On determining the number of preambles in grant-free mMTC uplink to reduce collisions," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2023, pp. 1–6.

[23] H. Grama Srinath, M. Rana, and N. M. Balasubramanya, "Grant-free access for mMTC: A performance analysis based on number of preambles, repetitions, and retransmissions," *IEEE Internet Things J.*, vol. 9, no. 16, pp. 15169–15183, Aug. 2022.

[24] Y. Wang, W. Xu, M. Juntti, J. Lin, and M. Pan, "Composite preambles based on differential phase rotations for grant-free random access systems," *IEEE Internet Things J.*, vol. 10, no. 19, pp. 17035–17046, May 2023.

[25] Z.-S. Lien and C.-H. Lee, "Deep neural network based active user detection for grant-free multiple access," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 13007–13022, Sep. 2024.

[26] H. Xiao, W. Chen, J. Fang, B. Ai, and I. J. Wassell, "A grant-free method for massive machine-type communication with backward activity level estimation," *IEEE Trans. Signal Process.*, vol. 68, pp. 6665–6680, 2020.

[27] S. Zhang, Y. Cui, and W. Chen, "Joint device activity detection, channel estimation and signal detection for massive grant-free access via BiGAMP," *IEEE Trans. Signal Process.*, vol. 71, pp. 1200–1215, 2023.

[28] S. Khan, S. Durrani, M. B. Shahab, S. J. Johnson, and S. Camtepe, "Joint user and data detection in grant-free NOMA with attention-based BiLSTM network," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1499–1515, 2023.

[29] Y. Bai, W. Chen, B. Ai, Z. Zhong, and I. J. Wassell, "Prior information aided deep learning method for grant-free NOMA in mMTC," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 112–126, Jan. 2022.

[30] D. Wu, Z. Zhang, Y. Huang, and X. Qin, "Priority-aware access strategy for GF-NOMA system in IIoT: The device-specific allocation approach," *IEEE Internet Things J.*, vol. 11, no. 2, pp. 2152–2165, Jul. 2023.

[31] Y. Liu, Y. Deng, H. Zhou, M. Elkashlan, and A. Nallanathan, "Deep reinforcement learning-based grant-free NOMA optimization for mURLLC," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1475–1490, Mar. 2023.

[32] M. Elsayem, H. Abou-zeid, A. Afana, and S. Givigi, "Reinforcement learning-based dynamic resource allocation for grant-free access," in *Proc. IEEE Global Commun. Conf.*, Dec. 2022, pp. 1091–1096.

[33] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6369–6379, Jul. 2020.

[34] F. Xiao, X. Li, L. Yang, H. Liu, and T. A. Tsiftsis, "Outage performance analysis of RSMA-aided semi-grant-free transmission systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 253–268, 2023.

[35] H. Lu, X. Xie, Z. Shi, H. Lei, H. Yang, and J. Cai, "Advanced NOMA assisted semi-grant-free transmission schemes for randomly distributed users," *IEEE Trans. Wireless Commun.*, vol. 22, no. 7, pp. 4638–4653, Dec. 2022.

[36] D. Pliatsios, A. A. Boulogeorgos, T. Lagkas, V. Argyriou, I. D. Moscholios, and P. Sarigiannidis, "Semi-grant-free non-orthogonal multiple access for tactile Internet of Things," in *Proc. IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2021, pp. 1389–1394.

[37] M. Fayaz, W. Yi, Y. Liu, S. Thayaparan, and A. Nallanathan, "Toward autonomous power control in semi-grant-free NOMA systems: A power pool-based approach," *IEEE Trans. Commun.*, vol. 72, no. 6, pp. 3273–3289, Jun. 2024.

[38] M. Fayaz, W. Yi, Y. Liu, and A. Nallanathan, "Competitive MA-DRL for transmit power pool design in semi-grant-free NOMA systems," 2021, *arXiv:2106.11190.*

[39] H. Zhou, Y. Deng, L. Feltrin, and A. Höglund, "Analyzing novel grant-based and grant-free access schemes for small data transmission," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2805–2819, Apr. 2022.

[40] H. S. Jang, S. M. Kim, H.-S. Park, and D. K. Sung, "An early preamble collision detection scheme based on tagged preambles for cellular M2M random access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 5974–5984, Jul. 2017.

[41] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan, "Random access analysis for massive IoT networks under a new spatio-temporal model: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5788–5803, Nov. 2018.

[42] C.-W. Pyo, K. Takizawa, M. Moriyama, M. Oodo, H. Tezuka, K. Ishizu, and F. Kojima, "A throughput study of grant-free multiple access for massive wireless communications," in *Proc. 20th Int. Symp. Wireless Pers. Multimedia Commun. (WPMC)*, Dec. 2017, pp. 529–534.

[43] *Physical Layer Procedures for Control*, document TS 38.213 V16.2.0, 2020.

[44] *Study on User Equipment (UE) Power Saving in NR*, document TR 38.840 V16.0.0, 3GPP, 2019.

[45] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

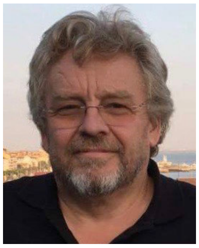
[46] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[47] A. Azari, "Energy-efficient scheduling and grouping for machine-type communications over cellular networks," *Ad Hoc Netw.*, vol. 43, pp. 16–29, Jun. 2016.

NASIM RAVI (Student Member, IEEE) received the B.Sc. degree in information technology engineering from Azerbaijan Shahid Madani University, Iran, in 2014, and the M.Sc. degree in computer engineering from Istanbul Technical University (ITU), Türkiye, in 2019. She is currently pursuing the Ph.D. degree with the University of Coimbra (UC). She is a Researcher with the Research Center for Informatics and Systems of the University of Coimbra (CISUC), Portugal. Her current research interests include machine type communication, access protocols, resource allocation, and reinforcement learning.

NUNO LOURENÇO received the Ph.D. degree in information science and technology from the Department of Informatics Engineering, University of Coimbra, in 2016. He is currently the Coordinator of the Evolutionary and Complex Systems (ECOS) Group. He has been a member of the Centre for Informatics and Systems of the University of Coimbra (CISUC), since 2009. Formerly, he was appointed as a Senior Research Officer with the University of Essex, U.K. He is currently an Assistant Professor with the Department of Informatics Engineering, University of Coimbra. He is the Co-Creator of Structured Grammatical Evolution, Probabilistic Grammatical Evolution, and DENSER, a novel approach to automatically design deep artificial neural networks using evolutionary computation. He has authored or co-authored more than 60 articles in journals and top conferences from the evolutionary computation and artificial intelligence areas and he has been involved as a Researcher in 13 projects (national and international). His main research interests include bio-inspired algorithms, optimization, and machine learning. He is a member of the Program Committee of GECCO, PPSN, and EuroGP; a member of the Steering Committee of EuroGP; and an Executive Board Member of SPECIES. He served as the Chair in the main conferences for the Evolutionary Computation field, namely EuroGP 2020 and 2021 as the Program Chair, and PPSN 2018 and EuroGP 2019 as the Publication Chair.

MARILIA CURADO received the Ph.D. degree in computer engineering from the University of Coimbra, Coimbra, Portugal, in 2005. She is a Full Professor with the Department of Informatics Engineering, University of Coimbra. She is currently the Director of the Laboratory for Informatics and Systems, Pedro Nunes Institute, Coimbra. She has participated in a large number of national and international projects. Her research interests include resilience and quality of service in 5G networks, the Internet of Things, and communications in the cloud. She is a member of the Editorial Board of *Computer Networks* (Elsevier), *Computer Communications* (Elsevier), and *Internet Technology Letters* (Wiley) and has been involved in the scientific organization and coordination of several international conferences.



EDMUNDO MONTEIRO (Senior Member, IEEE) received the degree in electrical engineering (informatics specialty) from the University of Coimbra (UC), Portugal, in 1984, and the Ph.D. degree in electrical engineering (computer communications), and the Habilitation degree in informatics engineering, in 1996 and 2007, respectively. He is a Full Professor with UC. He is currently the Head of the Informatics Engineering Department, UC. He is a Senior Researcher with the Centre for Informatics and Systems of the University of Coimbra (CISUC). He has more than 35 years of research and industry experience in the fields of computer communications, wireless and mobile communications, quality of service, network and service management, cybersecurity, critical infrastructure protection, cloud networking, the Internet of Things, and sensor networks. He participated in many Portuguese, European, and international research projects, and initiatives. His publications include six books (authored and edited), several book chapters, and more than 200 articles in international refereed journals and conferences. He is the co-author of nine international patents. He is a member of the Editorial Board of *Wireless Networks* (Springer) and *ITU Journal on Future and Evolving Technologies* journals. He was involved in the organization of many national and international conferences and workshops. He is a member of Ordem dos Engenheiros (the Portuguese Engineering Association), and a Senior Member of the IEEE Communication Society, the IEEE Computer Society, and ACM SIGCOMM. He is also the Portuguese representative in IFIP TC6 (Communication Systems).