Deep Reinforcement Learning Based Load Balancing Scheme in Dense Cellular Network Using RoF Technology

Mahfida Amjad Dipa, Syamsuri Yakoob, Fadlee Rasid, Faisul Ahmad, and Azwan Mahmud

***Abstract*—In a dense cellular network, the small cell size and limited frequency make it hard to control the traffic, and hence, there is a necessity for the transmission points to know how much traffic they can handle. To fix this problem in the network, this study suggests a Load Balancing (LB) scheme based on Reinforcement Learning (RL) named DRL-LB adopting a Deep Deterministic Policy Gradient (DDPG) RL approach for a dense cellular network utilizing the RoF technologies. The DRL-LB technique is based on self-exploration in the continuous action space to speed up the execution process. The SNR of the dense network has been taken into account to increase the network spectral efficiency concerning the number of users. The number of users per base station satisfying the minimum SNR value acts as the LB constraints in the scheme. The result analysis shows that it can achieve the required 10 dB of SNR value with 1.6 bits/s/Hz spectral efficiency. It attains a higher spectral efficiency and rewards around 78% compared to the non-LB approach in the scheme. Furthermore, the simulation process also depicts that DRL-LB is 73% more efficient in running time.**

***Index Terms*—reinforcement learning, deep deterministic policy gradient, DDPG, load balancing, radio over fiber, RoF, dense network.**

I. Introduction

With the advancement of future wireless networks, new technologies are being developed and suggested to meet 5G design goals. To meet the growing demands on flexibility, reliability, and transmission capacity of wireless networks, microwave photonics has emerged as an effective tool and in such states, networks like Radio over Fiber (RoF) [1] which is a kind of cellular network integration, is the combination of an optical fiber infrastructure and Radio Frequency (RF) technology can provide a scalable solution towards millimeter-Wave (mmW) signal for enhancing Mobile Broadband (eMBB) for 5G network transmission.

When a cell experiences significant network traffic, cellular networks may become inconsistent. In addition to being unable to meet user demands, an overloaded cell may cause issues like throughput and latency. As the number of users increases, the system's performance can be impacted by the relatively limited network coverage. Hence, the traditional rule-based Load-Balancing (LB) methods are not suitable for dense networks. Considering the issue of small cellular coverage with a huge traffic load, developing an efficient LB scheme for a dense cellular network is crucial. However, some recent LB techniques are using rule-based methods in different networks like Cloud Radio Access Network (C-RAN) [1] and LiFi [2]. Regarding the learning techniques, such as Reinforcement Learning (RL), have shown effectiveness for dealing with communication LB [3]. RL attempts to learn control policies through interaction with the environment.

Based on a distributed multi-agent deep Q-network, an LB mechanism has been depicted in [3] focusing on user association for dense networks. It uses a matching game-based policy for LB, where each Base Station (BS) maintains a preference list to make association decisions utilizing the sum data rates. Another LB scheme has been presented in [4] for heterogeneous LiFi/WiFi networks. To maximize average network throughput and user satisfaction in terms of user data rate, it shows three distinct RL reward functions are shown using the Trust Region Policy Optimization (TRPO) learning algorithm. Another RL technique has been reported in [5] using the Deep Deterministic Policy Gradient (DDPG) method for increasing Spectral Efficiency (SE) in the massive Multiple Input Multiple Output (MIMO) communication system, considering the Signal to-Interference-plus-Noise Ratio (SINR). Another study is reported in [6], conducted on an LB approach in Self Organized Network (SON), which employs a ranked buffer strategy where BSs compete to achieve LB while considering quality of service metrics through a multi-agent DDPG scheme that emphasizes throughput, resource block utilization, and network latency. It uses a competitive framework where each agent aims to maximize its received reward under the assumption of worst-case scenarios, while simultaneously, other agents strive to minimize the rewards of their competitors. Each agent, represented by a BS, tries to

Manuscript received April 23, 2025; revised May 12, 2025. Date of publication July 15, 2025. Date of current version July 15, 2025.

This research is funded by the Organization for Women in Science for the Developing World (OWSD), fund reservation no. 3240318596 and the Fundamental Research Grant (FRGS), Ministry of Higher Education, Malaysia, No. FRGS/1/2023/TK07/UPM/02/11.

M. A. Dipa, S. Yakoob, F. Rasid, and F. Ahmad are with the Wireless and Photonics Networks Center, Faculty of Engineering, Universiti Putra Malaysia, Malaysia (e-mails: amjad.mahfida@student.upm.edu.my, {syamsuri, fadlee, faisul}@upm.edu.my).

A. Mahmud is with the Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia (e-mail: azwan.mahmud@mmu.edu.my).

Digital Object Identifier (DOI): 10.24138/jcomss-2025-0056

enhance its throughput while observing the constraints related to latency and resource block utilization.

But to the best of the author's knowledge, LB for RoF infrastructure based on DDPG has not been reported yet. Hence, the contributions of this article can be listed as follows:

i. this study introduces the Deep RL based LB scheme named as DRL-LB scheme designed to address the LB challenge within dense cellular networks through a self-exploration approach. The objective of this scheme is to dynamically evaluate the User Entities (UEs) per BS in a self-exploratory manner, adapting the appropriate Signal to noise Ratio (SNR) by employing a DDPG algorithm and this self-exploration mechanism enhances the SE of the system.
ii. DRL-LB is an off-policy RL based algorithm for LB, which utilizes deep neural networks to approximate the LB policy, significantly enhancing the model's performance capabilities over RoF infrastructure. The DRL-LB can be trained within a learning framework to autonomously derive the optimal LB policy without requiring any prior knowledge of the underlying dense environments.
iii. by utilizing the DDPG agent in the DRL-LB scheme it can find the most effective direction of action to maximize the estimated cumulative long-term reward based on the model-free RL method in the continuous action space and hence it can increase the SE based on the SNR value of the dense network. Additionally, by utilizing the DDPG method it accelerates the execution phase.

The rest of this paper is organized as follows: Section II is about dense cellular network design using RoF technology, Section III presents a proposed deep reinforcement learning-based load balancing scheme, result analysis and discussion have been described in Section IV, and finally, Section V concludes this paper.

II. Dense Cellular Network Design using RoF Technologies

The Optical Heterodyne (OH) [7] is a RoF technology that acts as a photonics-assisted RF signal synthesis scheme, providing an alternative to traditional electronic approaches in terms of cost, complexity, and bandwidth. Among these techniques. The OH scheme of direct frequency downmixing of two optical carriers on a Photodiode (PD) is one of the most straightforward and effective ways to achieve high-capacity wireless communication systems flexibly. For creating the two cells OH RoF technique has been utilized, considering two RF signals, 4 GHz and 30 GHz, representing macro and microcell, respectively, in the dense network by 3rd Generation Partnership Project (3GPP) specifications, according to [8]. The reason for choosing the two RF signals, 4 GHz and 30 GHz, is to design the two layers in the dense network. In 5G New Radio (5G NR), the 3GPP standard employs both low-frequency bands such as 4 GHz and high-frequency bands near 30 GHz. The 4 GHz band, which is included in Frequency Range 1 (FR1), is regarded for its ability to provide wider coverage and support higher mobility applications. Conversely, the 30 GHz band, categorized under Frequency Range 2 (FR2), is frequently linked to mmW technology, delivering elevated data rates. The 3GPP is responsible for the development and standardization of mobile network technologies, which includes

TABLE I
Design Parameters of the Dense Cellular Network using RoF Technology

Symbol	*Name*	*Values*
f	RF Signals	Macrocell (4GHz), Microcell (30 GHz)
D	Cell Coverage	Macrocell (500m), Microcell (200m)
λ	Wavelength	1550 nm
L_p	Laser power	-10 dBm
N_L	Laser relative intensity noise	-140 dB/Hz
d	SMF Fiber Length	10 km
V_π	Drive voltage of MZM	4 V
σ	Fiber chromatic dispersion	16 ps/nm/km
R	Responsivity of PD	0.8 A/W
m	Modulation	64 QAM
W	UE Channel Bandwidth	20 MHz
SCS	Subcarrier spacing	15 KHz

5G New Radio (5G NR) that operates on both FR1 and FR2. As a reference the default specifications by 3GPP of dense network scenario has been presented in appendix section in Fig. A1. The design parameters of the dense network simulation of this article are presented in Table I. And the block diagram of the network design has been presented in Fig.1. A Continuous Wave (CW) laser with a wavelength, λ of 1550 nm is used to generate an optical light wave producing an output power, L_p of -10 dBm with an intensity noise, N_L of -140 dB/Hz. The optical signal passes through a Polarization Controller (PC) to maximize the optical signal coupling into the Mach Zehnder Modulator (MZM) driven by a drive voltage, V_π of 4V while at the same time ensuring as small as possible polarization-dependent loss. Next, the RF signal, f, is transmitted into the MZM, which is driven by a RF signals 4 GHz RF signal with direct modulation and a 15 GHz RF signal with external modulation. Then the RF signal is launched into a fiber length, d of 10 km Single Mode Fiber (SMF) with a chromatic dispersion, σ of 16 ps/nm/km. The optical signal is detected by a high-speed PD with a 3 dB bandwidth and a responsivity, R, of 0.8 A/W. The PD output is fed into a spectrum analyzer to be monitored and measured. The Delay Interferometer (DI) has been used to separate two optical signals at the receiving end. At the receiving end, the PD converts the optical field to current, followed by a Transimpedance Amplifier (TIA) for amplification, and then demodulated by Amplitude (AM) demodulators to get back the electrical signals. The generated photocurrent can be analyzed by the RF spectrum analyzer and a BER analyzer. The channel bandwidth, W, and the subcarrier spacing, SCS, for the dense network simulation are 20 MHz and 15 KHz, respectively, for 64 QAM modulation, m technique.

The 30 GHz RF signal is generated using the Dual Sideband (DSB) mechanism due to its simplicity and system efficiency for high-speed wireless data transmission [9], [10], [11]. The DSB module can be realized by a MZM and the output of the DSB technique can be given as [12]

$$DSB\ (t) \propto E1\ exp\ (j2\pi ft)\ exp\ [j2\pi ft + 2\pi hm(t)] + exp\ [-j2\pi ft - 2\pi hm(t)] + \alpha A \quad (1)$$

where f is the RF signal, E_1 is the intensity of the electrical field, $2\pi h$ is the modulation index, $m(t)$ is the message signal,

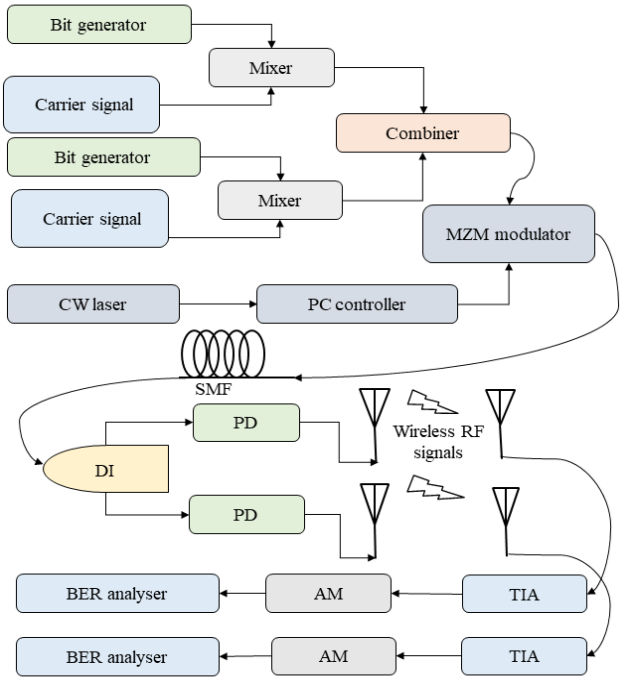

Fig. 1. The block diagram of the dense cellular network design using RoF technology

A is the amplitude of the optical carrier, and α denotes the carrier suppression factor. For the DSB optical signal dominated by the first-order sidebands, α is close to zero. The modulated signal is further sent to the PD for heterodyne mixing, as follows

$$PD(t) = \mathrm{R}|E_1 DSB(t) + E_2 exp(j2\pi ft)|2 = 2\mathrm{R}E_1E_2\ cos\ [2\pi ft + 2\pi hm(t)] + cos\ [2\pi ft - 2\pi hm(t)] + \alpha A\ cos[2\pi ft] \quad (2)$$

where R is related to the responsivity of PD and E_2 denotes the amplitude of the optical carrier. Since the low-frequency terms ($\cos(2\pi ft)$) can be filtered by the RF components like power amplifiers or antennas operating at the high-frequency RF band and can be ignored in Eq. (2) [13]. Thus, only the RF signals centered around $2f$ are generated through DSB. For calculating the Signal to Noise Ratio (SNR), this paper adopted the following equation from [34]

$$\gamma = \mathrm{s_k} - N \quad (3)$$

where γ is the SNR in dB, s_k is the received power and N is the noise power in dBm. From the SNR values to calculate Error Vector Magnitude (EVM) which is the metric used to quantify the performance of a radio transmitter or a receiver, the following equations have been considered

$$\mathrm{EVM} = \sqrt{\frac{1/T}{P_0}\sum_{t=1}^{T}\left[\left|n_{I,t}\right|^2 + \left|n_{Q,t}\right|^2\right]} \quad (4)$$

where $n_{I,t}$ and $n_{Q,t}$ are in-phase and quadrature components, respectively. P_0 is the power of the normalized ideal constellation or the transmitted constellation. According to Eq. (4) can be written as

$$EVM \approx \sqrt{1/SNR} \quad (5)$$

The key performance indicators of the dense network are average Spectral Efficiency (SE) and user throughput denoted as Γ. If $R_i\ (t)$ denotes the number of correctly received bits by the total number (N) of UEs, and M is the number of transmission-reception points. The channel bandwidth is denoted by W, and t is the time over which the data bits are received. The SE has been estimated by running system-level simulations over a number of N drops N_{drops}. Each drop gives a value of $\sum_{i=1}^{N} R_i(t)$ denoted as $R_1(t), \ldots R_{(N_{drops})}\ (t)$ and the estimated SE resulting is given by [15]

$$SE = \frac{\sum_{j=1}^{N_{drops}} R^{(j)}(t)}{N_{drops}.t.W.M} = \frac{\sum_{i=1}^{N_{drops}}\sum_{j=1}^{N_{drops}} R_i^{(j)}(t)}{N_{drops}.t.W.M} \quad (6)$$

where SE is the estimated average spectral efficiency, and it will approach the actual average with an increasing number of N_{drops} of UEs in the network, and $R_i^{(j)}(t)$ is the simulated total number of correctly received bits for UE_i in drop j. For calculating throughput, UE_i in drop j correctly decode $R_i^{(j)}(t)$ accumulated bits in [0, t]. During this total time UE_i receives an accumulated service time of $t_i \leq t$, where the service time is the time duration between the first packet arrival and when the last packet of the burst is correctly decoded. In the case, of a full buffer, $t_i = t$. Hence, the rate of normalization by service time t_i and channel bandwidth (W) of UE_i in drop j, $r_i^{(j)}$, is

$$r_i^{(j)} = \frac{R_i^{(j)}(t)}{t_i.W} \quad (7)$$

Running all N_{drops} (UE_n) simulations leads to $N_{drops} \times N$ values of $r_i^{(j)}$ of which is the lowest 5th percentile point of the Cumulative Distribution Function (CDF) is used to estimate the 5th percentile user SE. The network performance metrics have been evaluated considering the throughput derived from the received signal power and the noise power of the RF signals and can be calculated using the following equations adopted from [16]. The throughput has been measured using the following equation

$$\Gamma = W * log\ (1 + \gamma) \quad (8)$$

where Γ represents throughput in Mbps and γ represents SNR in dB derived from Eq. (3). And W is the channel bandwidth which is 20 MHz for 15 KHz subcarrier spacing and 10 ms time frame using 64 QAM modulation. The SE can be derived using Eq. (3) as the below equation

$$SE = log\ (1 + \gamma) \quad (9)$$

where SE measured in bits/s/Hz and γ represents SNR in dB derived from Eq. (3).

III. Proposed Deep Reinforcement Learning Based Load Balancing Scheme

Machine learning techniques like RL have shown effectiveness in dealing with communication load balancing [17]. RL attempts to learn control policies by interacting with the environment. However, RL-based techniques have inherent challenges. RL requires frequent interactions with the environment to learn a satisfactory policy and a reward function to achieve the desired performance [18]. This article proposes the DRL-LB scheme for the dense cellular network using RoF technologies, using a DDPG agent to find the most effective direction of action that maximizes the estimated cumulative long-term reward based on the model-free RL method in the continuous action space [35]. The goal of the RL is to train an agent to complete a task within an unknown environment. The agent receives observations and a reward from the environment and sends actions to the environment. The reward is a measure of how successful an action is concerning completing the task goal. In each training time step the current observation from the RoF environment selects an action based on the number of UE and SNR and returns the corresponding action that maximizes the long-term reward. Then it observes the reward and the next observation from the environment. It stores the experience in the experience buffer to update the policy through the RL algorithm. To compute the cumulative reward, the agent first computes the next action by passing the next observation from the experience buffer to the target action to find the cumulative reward by minimizing the loss across all the experiences.

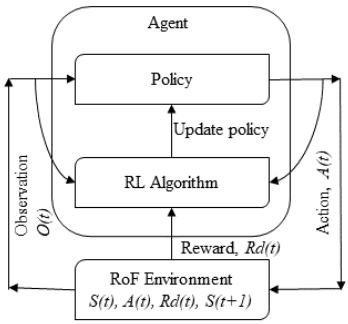

Fig. 2. Schematic diagram of DRL-LB scheme

Fig. 2 presents the schematic diagram of the proposed DRL-LB scheme. In the scheme, the agent and the environment interact at each of a sequence of discrete time steps. At a given time step *t*, the environment is in a state *S*(*t*), which results in the observation *O*(*t*). Based on *O*(*t*) and its internal policy function, the agent calculates an action *A*(*t*). According to its internal dynamics the RoF environment updates its state to *S*(*t*+1), which results in the next observation *O*(*t*+1) based on both the state *S*(*t*) and the action *A*(*t*). Based on *S*(*t*), *A*(*t*), and *S*(*t*+1), the environment also calculates a reward *Rd*(*t*+1). The reward is an immediate measure of how good the action *A*(*t*) is. At the next time step *t*+1, the agent receives the observation *O*(*t*+1) and the reward *Rd*(*t*+1). Based on the memory buffer of observations and rewards received, the learning algorithm updates the agent's policy parameters in an attempt to improve the policy function. The parameter update may occur at each step or after a sequence of steps. Finally, the agent calculates the next action *A*(*t*+1) based on *O*(*t*+1) and on its policy function, and the process is repeated. The LB policy and the rewards of the DRL-LB scheme are presented in the following two subsections, and the algorithm and flow of the algorithm have been depicted in Algorithm I and Fig. 3, respectively.

A. LB Policy

The agent for the DRL-LB scheme is the UE. This paper takes into account the SNR values defined in Eq. (3). The agent is being trained to determine whether the number of UEs is in an acceptable range or not, considering the SNR value. As this article considers the number of UE per BS in a cell in the network, it is necessary to provide information regarding the load of a particular BS to the RL agent. The load constraint of the LB technique is the number of UE per BS, and by restricting the maximum number of UEs at each BS, it can balance the loads among all BSs in a dense wireless network [22]. Furthermore, the minimum SNR value for connection establishment is 10 dB [4]. Therefore, it is more efficient to transmit the SNR information to the LB controller. The UEs are under the BS coverage of the dense network and can be connected if they get acceptable SNR values. From the 3GPP standards [15] the number of users served by each BS is 10 and is denoted as $\mathcal{U}$. Thus, the number of linked users at each BS cannot exceed $\mathcal{U}$. The number of BSs in the cell is denoted as $\mathcal{B}$ and the SNR at the $\mathcal{U}_{th}$ attended by the BS can be written as ξ. Hence, to maximize the SE using Eq. (9), the LB constraint in the RoF dense network formulated as

$$C = \sum_{i=1}^{\mathfrak{b}} \mathfrak{b} \sum_{j=1}^{\mathfrak{b}} \mathfrak{b}[\mathfrak{b}|\mathfrak{b} < \mathfrak{b}_{th}] \sum_{j=1}^{\mathfrak{b}} \xi[\xi|\xi > \xi_n] \tag{10}$$

where *C* denotes the LB constraint for the LB policy; $\mathcal{B}$, $\mathcal{U}$ and ξ denote number of BSs, the number of UEs, and the SNR values of the $\mathcal{U}_{th}$ respectively. And

$$\mathfrak{b} \in (1,2,3,\dots,10), \text{ such as } [1 < \mathfrak{b} < 10] \tag{11}$$

$$\xi \in (10,11,12,\dots,40), \text{where } [1 < \xi < 40] \tag{12}$$

B. Rewards

If the LB policy is true during the simulation, then it is assumed that the proposed network is balanced. The reward function, *Rd* is calculated based on the SNR values. Hence, the modified R can be written

$$Rd_j = \sum_{i=1}^{\mathfrak{b}} \left[Rd_j \middle| Rd_j \epsilon (Rd) \right] \tag{13}$$

where, $\mathcal{U}$ is the number of UEs connected to BS_j, and Rd_j corresponds to the *Rd* measured at time *t* by UE_i. The *Rd* is intended that if the request of a UE is rejected by the BS, it would get a lower reward value as it is violating the LB constraints through the LB policy by designing a loss or error function, *e* as

$$e = [e \mid e \in \{(e > 1) \cup (e < -1)\}] \quad (14)$$

Hence, the corresponding *Rd* from Eq. 14 can be written as

$$Rd = \sum_{i=1}^{b} (Rd_j + e) \quad (15)$$

If the value of *e* is greater than 1, it would penalize the average UE performance metrics, whereas if *e* is less than 1, the reward will be high, indicating the network is balanced. The algorithm of the DRL-LB has been presented in Algorithm I.

Algorithm I DRL-LB Scheme

Step 1: Initialize the agent UE
Step 2: Reset the environment according to Eq. (9)
Step 3: Get initial observation from the environment according to Eq. (12)
Step 4: Compute initial action based on current policy according to Eq. (10)
Step 5: From the current observation, the agent selects an action according to Eq. (9) and Eq. (12)
Step 6: Based on the action, the reward is calculated according to Eq. (15)
Step 7: Update the current action & observation with the next action & observation
Step 8: If it does not reach the stopping criteria, then go to step 2; otherwise, go to step 9
Step 9: End of training

From Algorithm I, at first, the LB scheme is initiated by the agent UE. Then it reset the RoF environment according to the Eq. (9). After getting the initial observation from the environment based on Eq. (12), it computes the initial action according to the current LB policy according to Eq. (10). Then from the current observation the agent selects the action according to Eq. (9) and Eq. (12). Based on the action the reward is calculated according to Eq. (15). Then, it updates the current action and observation with the next action and observation and this process continues until it reaches the stopping condition. The process flow of the DRL-LB scheme has been depicted in Fig. 3.

TABLE II
DESIGN PARAMETERS OF DRL-LB SCHEME

Symbol	*Name*	*Values*
β	Learning rate	10^{-3}
Opt	Optimizer	Adam
$\|B\|$	Mini batch size	64
$\|E\|$	Number of episodes	500
$\|T_{steps}\|$	Time steps per episode	1000
γ	Discount factor	0.99

C. Datasets

The training dataset is generated by software simulation through Matlab vR2023b. The design parameters of the DRL-LB scheme are presented in Table II. The training is done by considering the number of UEs, received signal power and the noise power of the UE in the network. The dataset consists of 500 episodes for each UE and 1000 timesteps for each episode.

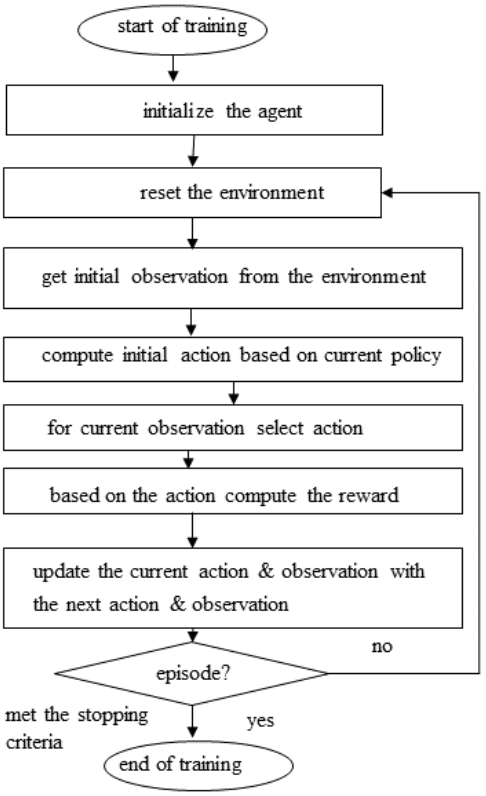

Fig. 3. Process flow of DRL-LB scheme

Each agent's RL structure is built using three hidden layers of the Rectified Linear Unit (ReLU) activation function (64, 32, 32) with a discount factor, γ, with a value of 0.9, a learning rate, β, of 10^{-3}. Using the optimizer, *Opt* as Adam optimizer and a mini-batch size, $|B|$ of 64, the weights of the DRL-LB method are updated at each time step. The datasets are used to train system models to be more subtle and recognize the input data. Through this data collection, the model becomes stable and can realize the data for testing. The testing process of the DRL-LB scheme can recognize the data effectively to produce optimal SE values [5]. The deep neural network of the DDPG agent has been trained based on the dataset of independent realizations of the UEs' received signal power and noise power, obtained by the outputs from the DRL-LB scheme in RoF environments by applying Eq. (15).

D. Training and Validation Procedure

The DRL-LB scheme considers an episodic training procedure, where in each episode, the number of UEs and the SNR values of UEs are randomly selected following a set of probability distributions and constraints based on Eq. (11) and Eq. (12). Each new episode allows the scheme to experience a potentially unexplored subset of the observation space based on UE and SNR. The LB policy takes place based on Eq. (8) as a new episode begins and remains fixed for the duration of that episode. An episode consists of a fixed number of time steps, where at each time step, the agents decide on which UE and SNR actions the policy should take [21]. The DRL-LB scheme has undertaken training over 500 episodes, each consisting of 1000 steps. The RL architecture for each agent incorporates

three hidden layers utilizing the Rectified Linear Unit (ReLU) activation function, configured as (64, 32, 32). The discount factor is set at 0.9, while the learning rate is established at 10^{-3}. The adam optimizer is employed for weight updates at each time step, with a mini-batch size of 64. In the dataset, 80% of the samples are allocated for training purposes, 10% for validation, and the remaining 10% constitutes the independent test dataset. The simulations are performed on a workstation with Intel(R) Core (TM) i5-10210U CPU @ 1.60GHz, 2.10 GHz processor, 8.00 GB of RAM, using a 64-bit operating system. To verify the performance of the proposed DRL-LB scheme, this work considers its own simulation and training results in terms of using LB constraints and without LB constraints. However, this work makes a comparison based on the performance and design parameters of some recent works based on DDPG, as it is difficult to directly compare them due to different network scenarios and constraints.

IV. Result Analysis and Discussion

This section presents the network performance analysis regarding SNR, and EVM to received optical power and launched optical input power, to show the effectiveness of using RoF technology in the dense network design. And it is also essential as it is related to the proposed LB scheme, where SNR values are considered for training the DRL-LB scheme to increase the SE in this article. In Fig. 4 (a), it shows that at 0 dBm launched optical input power, the received optical input power is -40 to -42 dBm for 10 km fiber length for both 4 GHz and 30 GHz signals, which is within the acceptable range for the proper transmission in the network [22]. Fig. 4 (b) shows the variation of calculated SNR values using Eq. (3) with the launched optical input power. As the input power increases, the SNR also increases for both cells. From 5 dBm and above, the launched input power is getting greater than a 10 dB SNR value, which is the minimum level to establish a connection [4]. Fig. 4(c) shows the EVM vs the received optical power for 64 QAM both RF signals with a 10 km fiber link and the EVM values are calculated using Eq. (5). For 64-QAM, the average required optical power at the receiver to get an EVM of less than 8% is −14.8 dBm. As the received optical power increases, the EVM values decrease, and it is evident that the results have favorable compliance [8] Consequently, in Fig. 4 (d), it is also shown that as the SNR increases, it decreases the EVM values, as it is obvious that both the EVM and SNR are in upright alignment [23].

Figures 5 - 8 examine how the number of UE performs based on SNR, maintaining the SE in the system while satisfying the LB policy of the DRL-LB scheme. The learning was carried out in episodes, where each episode contains multiple learning time steps. During each episode, the DDPG agent is learning and updating every time step, and each UE is carried out once per episode. The system performance depicted in Fig. 5 in terms of SE concerning SNR that are examined by using Eq. (3) and Eq. (9) for two cells. The graph shows that the SE of the cells gradually increases with the increasing value of SNR. At the 10 dB SNR the SE for two cells is within 1 to 1.2 bits/s/Hz. In contrast, the DRL-LB scheme outperforms at 38 dB SNR with a SE of 1.7 bits/s/Hz, which is in the acceptable range according to the 3GPP specifications [8]. This behavior suggests that the SE increases with the increase of SNR, providing good system performance.

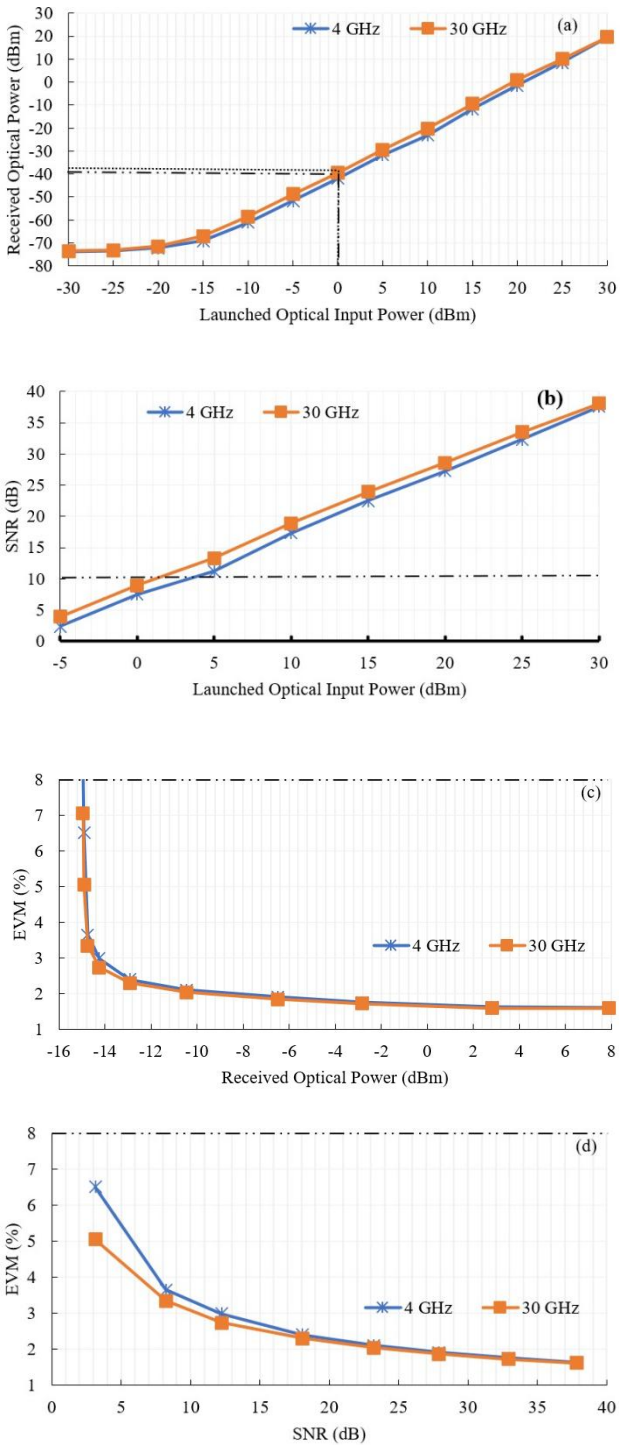

Fig. 4. Network performance analysis: (a) Received optical power vs launched input power, (b) SNR vs launched optical power, (c) EVM vs received optical power, (d) EVM vs SNR

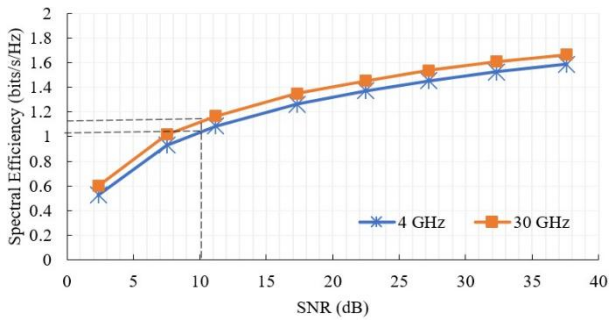

Fig. 5. Performance analysis: Spectral Efficiency vs SNR

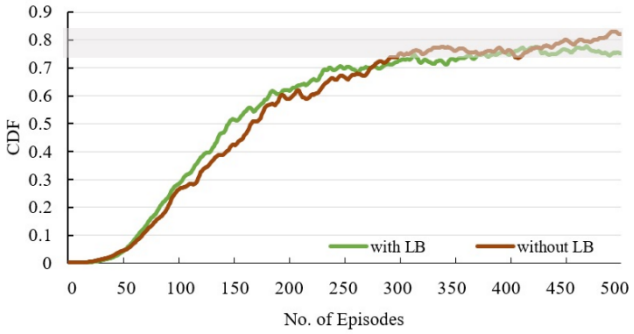

Fig. 6. Convergence behavior: the CDF values of the SE obtained by the DRL-LB scheme per UE

The convergence behavior of the DRL-LB scheme has been depicted in Fig. 6. It shows the correlation of the scheme without LB constraints. It is evident from the graph that the difference between the two curves in terms of CDF values is only 0.03, which indicates that the proposed scheme outperforms in achieving the SE [3]. Additionally, it can be stated that the scheme for both curves with LB and without LB justifies that the testing and validation of the scheme is 97% accurate [5]. As seen in Fig. 6, the DRL-LB scheme converges to a local optimum for both curves as each agent trains in each episode. Furthermore, the proposed scheme shows a similar and stable convergence behavior in the dense network, even though the values change from one episode to the next. It also shows that the proposed scheme, considering the SNR values of the UE per BS, enhances the SE. This result verifies the effectiveness of the proposed DRL-LB scheme with the LB policy utilising the DDPG RL agent.

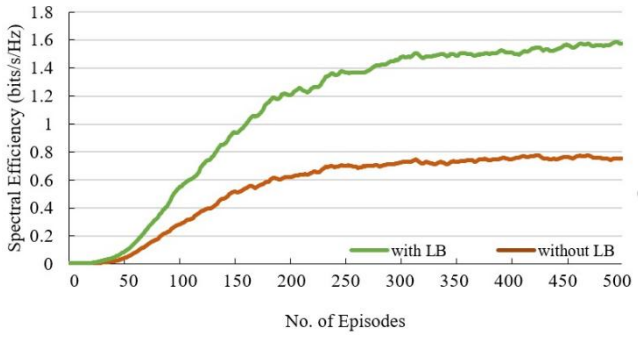

Fig. 7. Spectral efficiency obtained by DRL-LB scheme per UE

The system performance depicted in Fig. 7 in terms of SE for each user is obtained by Eq. (9) for a 20 MHz channel bandwidth for a 10 ms time frame. The graph shows that the SE gradually increases with each episode of the training. At the 500th Episode, the SE obtained using LB is 1.6 bits/s/Hz, which is in the acceptable range by 3GPP specifications defined in [8], whereas it is 0.78 bits/s/Hz for not using LB in the scheme. Hence, it is also noticeable that the DRL-LB is performing around 78% higher in gaining SE by applying the LB constraints in the scheme compared to the non-LB approach. Lastly, Fig. 8 shows the performance analysis of the DRL-LB scheme in terms of rewards obtained by Eq. (15). Fig. 8 shows a similar performance like Fig. 7 where the average rewards of the LB approach are 78% higher reward compared of the non-LB approach in the DRL-LB scheme.

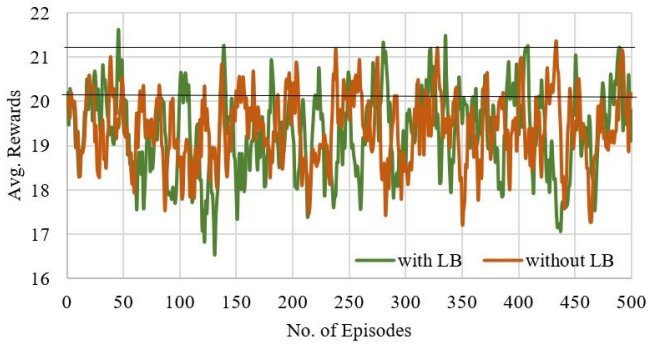

Fig. 8. Performance analysis of the DRL-LB scheme in terms of average. Rewards

To assess the efficacy of the proposed DRL-LB scheme, this study examines its simulation and training outcomes both with and without LB constraints. Furthermore, a comparative analysis is conducted based on the performance and design parameters of several recent studies utilizing DDPG, as it is difficult to directly compare them due to different network scenarios and constraints. The DRL-LB scheme has been compared with three recent DDPG approaches in Tables III for application type, design parameters, performance metrics, outcome and result of the methods. From Table III it states that two methods are applicable for MIMO communication, another one is for Self-Organizing Networks (SON), whereas the proposed scheme is for dense networks using RoF technologies for LB. Among these, in [5] it states that compared to not employing the DDPG approach, it may obtain a higher SE of 85% and an average SINR of 19.54 dB, whereas the conventional method only yields an average SINR of 15.32 dB. And in [24] it presented a downlink power control method using DDPG in a MIMO communication system was presented to increase the implementation speed of the system. And the scheme presented in [6] is for SON which tries to increase the throughput and decrease the training run time. On the contrary, the DRL-LB scheme has been proposed for LB in the dense cellular network adopting the DDPG RL method. And the result depicts that, it can outperform at 38 dB SNR with a SE of 1.7 bits/s/Hz. And hence, the DRL-LB can achieve 78% higher SE and average rewards compared to the non-LB approach in the scheme.

TABLE III
COMPARATIVE ANALYSIS

Method	*Application*	*Design Parameters*	*Performance Metrics*	*Result*
DDPG [5]	MIMO	SINR	SE	85% higher SE
DDPG [24]	MIMO	Power control	Speed	Faster execution process
DDPG [6]	SON	No. of users	Run time	70.49% improvement in run time
DRL-LB (proposed) using DDPG	Dense cellular network using RoF	SNR	SE, run time	78% higher SE than non-LB approach, 73% improvement in run time

The run time analysis of the proposed DRL-LB has been compared with [6] in Table IV. The DDPG work in [6] shows that it is running the training for 500 episodes and 1000 time steps each, around 30 minutes and claiming that it is gaining 70.49% improvement. Hence, this article is making a comparison with [6] where the proposed DRL-LB scheme takes only 22 minutes for the runtime, and thus it can achieve 73% improvement in run time for the same number of episodes and time steps in the scheme. Additionally, it takes 26 minutes to train the model without LB constraints, whereas it takes only 22 minutes for LB constraints, which is 4 minutes less than without LB constraints. Hence, in other words, it can be said that DRL-LB is 60% more efficient in running time than without using the LB constraints.

TABLE IV
RUN TIME ANALYSIS

Method	*No. of Episodes*	*No. of Time Steps*	*Total Run Time*	*Avg. Improvement*
DDPG [6]	500	1000	30 minutes	70.49%
DRL-LB (proposed)	500	1000	22 minutes	73%

V. CONCLUSION

This paper proposes an RL-based LB technique named as DRL-LB scheme for a dense cellular network that considers multiband 4 GHz and 30 GHz RF signals representing macro and microcells utilizing RoF technology. To enhance the performance efficiency of the DRL-LB scheme in the reinforcement training and execution process, this paper adopts model-free off-policy deep learning using a DDPG agent. The agent searches for an optimal policy that maximizes the expected cumulative long-term reward to satisfy the LB constraints. In the network performance study, the SNR and EVM performance metrics are considered to demonstrate the effectiveness of using RoF technology in dense network design. From the result analysis, it appears that the dense network performs well, maintaining acceptable values of 10 dB SNR and less than 8% EVM. Similarly, the numerical values indicate that the DRL-LB scheme delivers comparable performance in terms of SNR and SE in the network. It demonstrates that the suggested network can sustain an SNR value of 10 dB for effective signal transmission and maintain a user SE of 1.6 bits/s/Hz, which is within the acceptable range according to 3GPP specifications. Furthermore, the simulation results of the DRL-LB scheme shows that, when each agent trains with the DDPG agent, the network remains stable and converges similarly to a local optimum, while the difference in the CDF value for the SE is only 0.03 between the curves for LB and without LB, justifying that the testing and validation of the scheme is 97% accurate. Additionally, the DRL-LB can achieve 78% higher SE and rewards compared to the non-LB approach in the scheme.

This work is limited to software simulation only and does not include any experimental analysis. The SNR value plays a vital role in this LB scheme, achieving a higher SE value compared to the non-LB approach. One of the key benefits of the DRL-LB method is that it does not require knowledge of the policy used to create existing data logs, as DDPG is an off-policy method. This enables the model to be trained using random actions that remain within operational limits, with episode termination occurring when those limits are exceeded, and it operates independently of any specific policy that could enhance long-term behaviour. Thus, it would be fascinating to explore additional types of referring data related to specific real-world scenarios in future studies. Consequently, a future direction for further research could involve testing real-world data from various fields, such as cloud radio access networks or 5G/6G fibre-wireless networks.

REFERENCES

[1] K. Suresh *et al.*, "Enhanced Metaheuristic Algorithm-Based Load Balancing in a 5G Cloud Radio Access Network," *Electronics (Switzerland)*, vol. 11, no. 21, 2022, doi: 10.3390/electronics11213611.

[2] Y. Al-Karawi, H. Al-Raweshidy, and R. Nilavalan, "Optimizing the Energy Efficiency Using Quantum Based Load Balancing in Open Radio Access Networks," *IEEE Access*, vol. 12, pp. 37903–37918, 2024, doi: 10.1109/ACCESS.2024.3375530.

[3] B. Lim and M. Vu, "Distributed Multi-Agent Deep Q-Learning for Load Balancing User Association in Dense Networks," *IEEE Wireless Communications Letters*, vol. 12, no. 7, pp. 1120–1124, Jul. 2023, doi: 10.1109/LWC.2023.3250492.

[4] R. Ahmad, M. D. Soltani, M. Safari, and A. Srivastava, "Reinforcement Learning-Based Near-Optimal Load Balancing for Heterogeneous LiFi WiFi Network," *IEEE Syst J*, vol. 16, no. 2, pp. 3084–3095, Jun. 2022, doi: 10.1109/JSYST.2021.3088302.

[5] N. Nasaruddin, A. Risky, Y. Yunida, and R. Adriman, "Deep Deterministic Policy Gradient-Based Spectral Efficiency for Massive MIMO Communication System," *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 14, no. 1, pp. 13–22, 2025, doi: 10.18178/ijeetc.14.1.13-22.

[6] P. E. I. Rivera and M. Erol-Kantarci, "Competitive Multi-Agent Load Balancing with Adaptive Policies in Wireless Networks," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.07050

[7] K. Zeb *et al.*, "Broadband Optical Heterodyne Millimeter-Wave-over-Fiber Wireless Links Based on a Quantum Dash Dual-Wavelength DFB Laser," *Journal of Lightwave Technology*, vol. 40, no. 12, pp. 3698–3708, Jun. 2022, doi: 10.1109/JLT.2022.3154652.

[8] TSGS, "TS 122 261 - V17.11.0 - 5G; Service requirements for the 5G system (3GPP TS 22.261 version 17.11.0 Release 17)," 2022. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[9] J. Bohata, D. N. Nguyen, P. T. Dat, Z. Ghassemlooy, B. Ortega, and S. Zvanovec, "A Scalable 26 GHz RoF Relay System Using Optoelectronic Extender," *IEEE Photonics Technology Letters*, vol. 35, no. 6, pp. 293–296, 2023, doi: 10.1109/LPT.2023.3241438.

[10] S. Yaakob, R. M. Mahmood, Z. Zan, C. B. M. Rashidi, A. Mahmud, and S. B. A. Anas, "Modulation index and phase imbalance of dual-sideband optical carrier suppression (DSB-OCS) in optical Millimeter-wave system," *Photonics*, vol. 8, no. 5, 2021, doi: 10.3390/photonics8050153.
[11] J. Bohata, L. Vallejo, B. Ortega, and S. Zvanovec, "Optical CS-DSB Schemes for 5G mmW Fronthaul Seamless Transmission," *IEEE Photonics J*, vol. 14, no. 2, Apr. 2022, doi: 10.1109/JPHOT.2022.3161087.
[12] Z. Jia, J. Yu, and G. K. Chang, "A full-duplex radio-over-fiber system based on optical carrier suppression and reuse," *IEEE Photonics Technology Letters*, vol. 18, no. 16, pp. 1726–1728, Aug. 2006, doi: 10.1109/LPT.2006.879946.
[13] P. Li *et al.*, "Constant-Envelope OFDM for Power-Efficient and Nonlinearity-Tolerant Heterodyne MMW-RoF System with Envelope Detection," *Journal of Lightwave Technology*, vol. 40, no. 20, pp. 6882–6890, 2022, doi: 10.1109/JLT.2022.3199439.
[14] R. A. Shafik, M. S. Rahman, and A. H. M. R. Islam, "On the extended relationships among EVM, BER and SNR as performance metrics," *Proceedings of 4th International Conference on Electrical and Computer Engineering, ICECE 2006*, no. December, pp. 408–411, 2006, doi: 10.1109/ICECE.2006.355657.
[15] ITU-R, "Guidelines for evaluation of radio interface technologies for IMT-2020 M Series Mobile, radiodetermination, amateur and related satellite services," 2017. [Online]. Available: http://www.itu.int/ITU-R/go/patents/en
[16] R. Ahmad, M. D. Soltani, M. Safari, and A. Srivastava, "Reinforcement Learning-Based Near-Optimal Load Balancing for Heterogeneous LiFi WiFi Network," *IEEE Syst J*, vol. 16, no. 2, pp. 3084–3095, Jun. 2022, doi: 10.1109/JSYST.2021.3088302.
[17] D. Wu *et al.*, "Reinforcement learning for communication load balancing: approaches and challenges," *Front Comput Sci*, vol. 5, p. 1156064, May 2023, doi: 10.3389/FCOMP.2023.1156064/BIBTEX.
[18] D. Wu *et al.*, "Reinforcement learning for communication load balancing: approaches and challenges," 2023, *Frontiers Media S.A.* doi: 10.3389/fcomp.2023.1156064.
[19] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," Sep. 2015, [Online]. Available: http://arxiv.org/abs/1509.02971
[20] A. Alizadeh and M. Vu, "Load balancing user association in millimeter wave MIMO networks," *IEEE Trans Wirel Commun*, vol. 18, no. 6, pp. 2932–2945, Jun. 2019, doi: 10.1109/TWC.2019.2906196.
[21] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, "Resource Management in Wireless Networks via Multi-Agent Deep Reinforcement Learning," *IEEE Trans Wirel Commun*, vol. 20, no. 6, pp. 3507–3523, Jun. 2021, doi: 10.1109/TWC.2021.3051163.
[22] H. Yu *et al.*, "Deducing of Optical and Electronic Domains Based Distortions in Radio over Fiber Network," *Applied Sciences 2022, Vol. 12, Page 753*, vol. 12, no. 2, p. 753, Jan. 2022, doi: 10.3390/APP12020753.
[23] H.-H. Lu *et al.*, "A combined fibre/free-space-optical communication system for long-haul wireline/wireless transmission at millimetre-wave/sub-THz frequencies," *Communications Engineering*, vol. 2, no. 1, May 2023, doi: 10.1038/s44172-023-00068-1.
[24] L. Luo, J. Zhang, S. Chen, X. Zhang, B. Ai, and D. W. K. Ng, "Downlink Power Control for Cell-Free Massive MIMO With Deep Reinforcement Learning," *IEEE Trans Veh Technol*, vol. 71, no. 6, pp. 6772–6777, Jun. 2022, doi: 10.1109/TVT.2022.3162585.

APPENDIX

TABLE A.1
THE LIST OF ABBREVIATIONS USED IN THIS MANUSCRIPT

Name	*Abbreviations*
Radio over Fiber	RoF
Radio Frequency	RF
millimeter-Wave	mmW
enhancing Mobile Broadband	eMBB
Load-Balancing	LB
Cloud Radio Access Network	C-RAN
Reinforcement Learning	RL
Trust Region Policy Optimization	TRPO
Deep Deterministic Policy Gradient	DDPG
Spectral Efficiency	SE
Signal to-Interference-plus-Noise Ratio	SINR
Base Station	BS
User Entities	UE
Signal to noise Ratio	SNR
Deep RL based LB scheme	DRL-LB scheme (Proposed)
Optical Heterodyne	OH
Photodiode	PD
Continuous Wave	CW
Polarization Controller	PC
Mach Zehnder Modulator	MZM
Polarization Controller	PC
Single Mode Fiber	SMF
Delay Interferometer	DI
Dual Sideband	DSB
Error Vector Magnitude	EVM
Spectral Efficiency	SE
Cumulative Distribution Function	CDF
Rectified Linear Unit	ReLU
Multiple Input Multiple Output	MIMO
Self Organized Network	SON
3rd Generation Partnership Project	3GPP

	Dense Urban eMBB		
Parameters	Spectral Efficiency and mobility evaluations		User experienced data rate evaluation
	Configuration A	Configuration B	Configuration C
Baseline evaluation configuration parameters			
Carrier frequency for evaluation	1 layer (Macro) with 4 GHz	1 layer (Macro) with 30 GHz	1 or 2 layers (Macro + Micro) 4 GHz and 30 GHz available in macro and micro layers

Fig. A1. Frequency bands for dense cellular network by 3GPP

Mahfida Amjad Dipa received her B.Sc. degree in Computer Science and Engineering (CSE) from Manarat International University, Dhaka, Bangladesh, in 2007 and her Master degree in Information Technology (IT) from the Institute of Information Technology (IIT), University of Dhaka (DU), in 2009. Currently, she is pursuing her Ph.D. degree in Wireless Communications and Networks Engineering at Universiti Putra Malaysia (UPM), Malaysia. She started her teaching profession at the university level in 2012. At present, she is a faculty member in the department of CSE at Stamford University Bangladesh, Dhaka, Bangladesh (currently on study leave). Her research area is wireless communication, mobile ad hoc networks, and software engineering.

Syamsuri Yaakob received the B. Eng and M. Eng. Sc. degrees in electronic engineering from King's College London and Multimedia University (MMU), Cyberjaya, in 1998 and 2010, respectively. He received his Ph.D. in Millimeter-wave Radio over Fiber System from the Universiti Teknologi Malaysia (UTM) in 2014. He is currently an Associate Professor at the Dept. of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM). Before this appointment, he was a satellite engineer at Telekom Malaysia (1998) and a senior researcher at TM Research and Development Sdn. Bhd. (TMR&D) (2004). His current research interests are millimeter-wave radio over fiber system, passive optical networks, optical communication systems and access network technologies.

Mohd Fadlee A. Rasid is with the Faculty of Engineering Universiti Putra Malaysia (UPM). He received a B.Sc. in electrical engineering from Purdue University, USA and a Ph.D. in electronic and electrical engineering (mobile communications) from Loughborough University, U.K. He directs research activities within the Wireless Sensors group and his work on wireless medical sensors had gained importance in health care applications involving mobile telemedicine and has had worldwide publicity, including BBC news. He was a research consultant for a British Council UKIERI project on wireless medical sensors project. He was also part of the French Government STIC Asia Project on ICT-ADI: Toward a human-friendly assistive environment for Aging, Disability & Independence. He had led a few projects on sensor networks from Ministry of Communications and Multimedia Malaysia as well as Economic Planning Unit (EPU) under the Prime Minister's Department, particularly for agriculture and environmental applications. He was involved with a project under Qatar National Research Fund by Qatar Foundations on Ubiquitous Healthcare and recently with Korean Development Institute (KDI) on Smart City related.

Faisul Arif Ahmad is currently a senior lecturer in Department of Computer and Communication Systems Engineering, Faculty of Engineering, fUniversiti Putra Malaysia (UPM). He was graduated as B. Eng. in information technology in 2001 from Muroran Istitute Technology, Hokkaido, Japan and joined a research center at Panasonic AVC (formerly known as Matsushita Audio Video) (Malaysia) in 2001. He joined Universiti Putra Malaysia in November 2003 as a Tutor and finished his Master in Engineering, majoring Electrical in Universiti Teknologi Malaysia. He was granted his PhD from Universiti Putra Malaysia in 2016. His research is focused on mobile robotic system and particularly in the area of intelligent system. He is registered as a graduate engineer of the Board of Engineers Malaysia (BEM) and also member of Institute of Electrical and Electronics Engineers (IEEE) Systems, Man and Cybernetic (SMC), and Robotics and Automation Society (RAS).

Azwan Mahmud (Member, IEEE) received the B.Sc. degree (Hons.) in Electrical and Electronic engineering from University College London, London, U.K., in 1998, and the M.B.A. degree in strategic management from the University of Technology Malaysia, in 2008. He pursued his Ph.D. degree in wireless communications with The University of Manchester, U.K in 2014. Currently, he is working as a professional at Multimedia University (MMU), Cyberjaya Malaysia. His current research interests cover 5G and 6G, wireless systems, IoT, heterogeneous systems, to seek new knowledge and to explore and advance general scientific understanding. Additionally, his research also tries to solve practical problems and improve the quality-of-life utilizing technology and software.