

Intelligent Energy and Carbon Management in 6G: A Hybrid DRL-PPO Optimization Framework

Harshita Patil , Kaushlendra Sharma , and Shishupal Kumar 

Abstract—6th generation (6G) networks are anticipated to deliver highly-reliable, low-latency communications (URLLC); therefore, the increased use of Terahertz (THz) transceivers and Holographic MIMO surfaces will result in significant sustainability challenges as a result of the excessive amount of energy these devices will require for their operations. As part of this research, we present a framework using deep reinforcement learning (DRL) techniques known as Proximal Policy Optimization (PPO) that can effectively address two goals (conflicting) - minimizing energy consumption through intelligent channel sensing techniques and minimizing carbon footprints by using grid-aware resource allocation techniques. We demonstrate how a Carbon-Aware PPO agent can decouple network performance from carbon emissions by collecting operational energy usage records from live 5G networks and dynamically scaling them based on 6G power restrictions. Overall, we found that this approach could reduce carbon emissions approximately 84.5% versus traditional maximum performance baselines without causing any quality of service (QoS) violations; thus, effectively circumventing all limitations of static sleep-scheduling heuristic approaches.

Link to graphical and video abstracts, and to code:
<https://latam.ieee9.org/index.php/transactions/article/view/10658>

Index Terms—6G Networks, Deep Reinforcement Learning, Green AI, Carbon-Aware Networking, Intelligent Sensing.

I. INTRODUCTION

IN the future of wireless networking—the direction we’ll continue to move in, as we work toward the deployment of 6G networks—growth in the density of base stations and increased computational workloads associated with THz beamforming are anticipated to significantly increase energy consumption by the network. In terms of expectations for continued growth, energy efficiency approaches that have been traditionally deemed “Efficiency-First” and that are primarily focused on providing throughput and coverage above all else do not appear to be sustainable, given the trajectory toward carbon neutrality globally. Rohit Kumar, Saurav Kumar Gupta, Hwang-Cheng Wang, C. Shyamala Kumari, and Sai Srinivas Vara Prasad Korlam [1], [14] have noted that moving toward using higher frequency bands (e.g. mmWave and THz) will require a complete rethinking of how we consume energy in the Radio Access Network (RAN).

The associate editor coordinating the review of this manuscript and approving it for publication was Carolina Del-Valle-Soto (*Corresponding author: Harshita Gangadhar Patil*).

Harshita Gangadhar Patil, K. Sharma, and S. Kumar are with the Department of Computer Science and Engineering, Indian Institute of Information Technology, Nagpur, India (e-mails: dtea23cse017@iiitn.ac.in, ksharma@iiitn.ac.in, and skumar@iiitn.ac.in).

6G Base Stations consume energy for transmitting and receiving signals as well as the necessary processing of the signals (such as complex operations for Massive MIMO or Intelligent Reflecting Surfaces (IRS)). However, in 5G, energy costs were generally proportional to traffic load, whereas in 6G, 6G infrastructure incurs significant “static costs” just to keep the THz hardware in an “active” state. Therefore, efficient sleep mode management will become even more important [11], [15].

As a result of rigid energy-first heuristics (such as static sleep), service quality is often reduced. According to the work of Xavier Fernando and George Lazaroiu [2], the wake-up delay period associated with deep sleep states will make it difficult for those applications that are latency critical (in IIoT) to operate properly. This therefore represents a significant opportunities to optimize carbon emissions with respect to the stringent reliability conditions necessary for mission critical 6th generation applications [12].

This research fills a significant gap by proposing a “Carbon-First” approach to energy consumption through the development of self-learning Artificial Intelligence. By exploring the taxonomy of self-learning networks created by Leila Ismail and Rajkumar Buyya [3], we present an agent capable of utilizing a Deep Reinforcement Learning (DRL) model to predict: traffic spikes; and grid carbon footprint (intensity). By incorporating this predictive information into the system, it is possible to dynamically balance small energy expenditures against large carbon savings; thus decoupling the energy used in the system from the associated carbon impacts [16], [17].

The primary contributions of this paper are:

- 1) A dynamic **6G Power Scaling Model** that transforms operational 5G traces into THz-constrained power profiles.
- 2) A **Carbon-Aware PPO Agent** that optimizes resource allocation based on real-time grid intensity and traffic volatility.
- 3) A **Hybrid Edge-Cloud Architecture** that orchestrates cooperative transmission to distribute loads to green energy zones.
- 4) An empirical evaluation demonstrating an 84.5% reduction in carbon emissions with zero QoS violations.

II. RELATED WORK

A. Green AI and Smart Cities

Integrating artificial intelligence (AI) into the infrastructure of smart cities is a vital support system for achieving sustainability within those smart cities. According to Tan,

Mehmood, and Corchado [4], “Green AI” is, thus, understood not simply for application to achieve a more sustainable environment through the use of AI technologies, but to also achieve a more sustainable form of AI through other means of making sustainable energy. The implications of 6G network capabilities, therefore, would suggest that energy-conserving AI algorithms developed will also be able to conserve energy as they control the 6G network. Our work is consistent with this idea by using a new policy-based reinforcement learning algorithm, PPO (Proximal Policy Optimization), which has been shown to have faster convergence rates than DQNs (Deep Q-Network) while also reducing the amount of computing power needed [18].

B. Comparison with Existing Methods

Table I highlights the differences between existing static approaches and our proposed dynamic framework. While traditional methods rely on fixed thresholds, our PPO-based approach adapts continuously [20].

TABLE I
COMPARISON OF OPTIMIZATION FRAMEWORKS

Feature	Static Sleep	Standard DQN	Proposed PPO
Decision Basis	Fixed Threshold	Q-Values	Policy Gradient
Adaptability	Low	Medium	High
Carbon Aware	No	No	Yes
Convergence	N/A	Slow	Fast
QoS Risk	High	Medium	Low

C. Digitalization in 6G Infrastructure

According to Walter Konhäuser [6], digitalization within buildings and smart cities is needed before or in preparation for the deployment of 6G networks, so that they can be interacted with by the built environment to optimize energy consumption. Additionally, a group of researchers Maurizio Murrioni, Matteo Anedda, Mauro Fadda, Pietro Ruii, Vlad Popescu, Corneliu Zaharia, and Daniele Giusto [5] conducted a survey of enabling technologies for 6G-enabled smart cities and have identified “Intelligent Sensing” as a major component. Our work extends theirs by utilizing sensing to not only estimate channels but also create carbon-awareness; we will do this by directly integrating grid data APIs into the closed control loop process [13].

D. Energy-Efficient IIoT

Xavier Fernando and George Lazaroiu [2] suggest clustering techniques to decrease energy consumption in Industrial IoT by using 6G sensor networks that are efficient for low-powered devices, and do not consider the heavy energy consumption incurred by equipment into the network itself. Our approach focuses on creating an energy-efficient backbone network using the infrastructure of Edge Nodes and Cloud Servers to ensure both sides of a connection (i.e., the user’s devices and the backbone of the network) are equally efficient [19].

III. SYSTEM MODEL

A. THz Channel Propagation and Power Scaling

To ensure ecological validity, we utilize real-world energy consumption traces from a functional 5G base station network. However, 6G operates in the THz spectrum (0.1–10 THz), which suffers from high molecular absorption loss. To compensate, 6G base stations require significantly higher power density.

We model the received power P_r using the specific THz path loss model:

$$P_r = P_t \cdot G_t \cdot G_r \cdot \left(\frac{c}{4\pi f d} \right)^2 \cdot e^{-k_{abs}(f)d} \quad (1)$$

Where $k_{abs}(f)$ is the frequency-dependent molecular absorption coefficient. To overcome this loss, the active power P_t must be scaled up. To accurately capture 6G energy dynamics, total base station power must be viewed as a composition of hardware components: $P_{total} = P_{RF} + P_{BB} + P_{cooling} + P_{backhaul}$. For THz systems, the Radio Frequency power (P_{RF}) requires massive scaling due to beamforming and path loss compensation.

Let $P_{5G}(t)$ be the measured power at time t . The predicted 6G power consumption $P_{6G}(t)$ is modeled as:

$$P_{6G}(t) = \begin{cases} \alpha \cdot P_{5G}(t) & \text{if Active (Boost)} \\ P_{5G}(t) & \text{if Eco Mode} \\ \beta \cdot P_{5G}(t) & \text{if Deep Sleep} \end{cases} \quad (2)$$

Where $\alpha = 3.5$ represents the scaling factor for THz active transmission and $\beta = 0.2$ represents the efficiency of AI-enabled “Zero-Energy” sleep modes. These factors are verified against the power consumption models presented by Rohit Kumar, Saurav Kumar Gupta, Hwang-Cheng Wang, C. Shyamala Kumari, and Sai Srinivas Vara Prasad Korlam [1].

B. Traffic Arrival Modeling (MMPP)

Unlike 4G/5G traffic which often follows predictable daily patterns, 6G traffic—driven by Holographic Telepresence and XR—is highly bursty. To model this accurately, we utilize a Markov Modulated Poisson Process (MMPP).

The packet arrival rate $\lambda(t)$ is controlled by a continuous-time Markov chain $J(t) \in \{1, \dots, M\}$. In state j , packets arrive according to a Poisson process with rate λ_j . We define two primary states:

- 1) **Background State** (S_0): Represents low-bandwidth IIoT signaling and keep-alive messages. $\lambda_0 \approx 10$ Mbps.
- 2) **Burst State** (S_1): Represents active XR sessions or holographic frame transmission. $\lambda_1 \approx 10$ Gbps.

The transition matrix Q determines the state switching probabilities. Furthermore, to physically evaluate Quality of Service (QoS) and realistically model packet drops under such high loads, we introduce a finite-capacity data buffer modeled as an $M/M/1/K$ queue. With a maximum buffer Q_{max} , the queue length $Q(t)$ evolves over a decision epoch Δt as:

$$Q(t+1) = \max(0, \min(Q_{max}, Q(t) + (\lambda(t) - \mu(a_t)) \cdot \Delta t)) \quad (3)$$

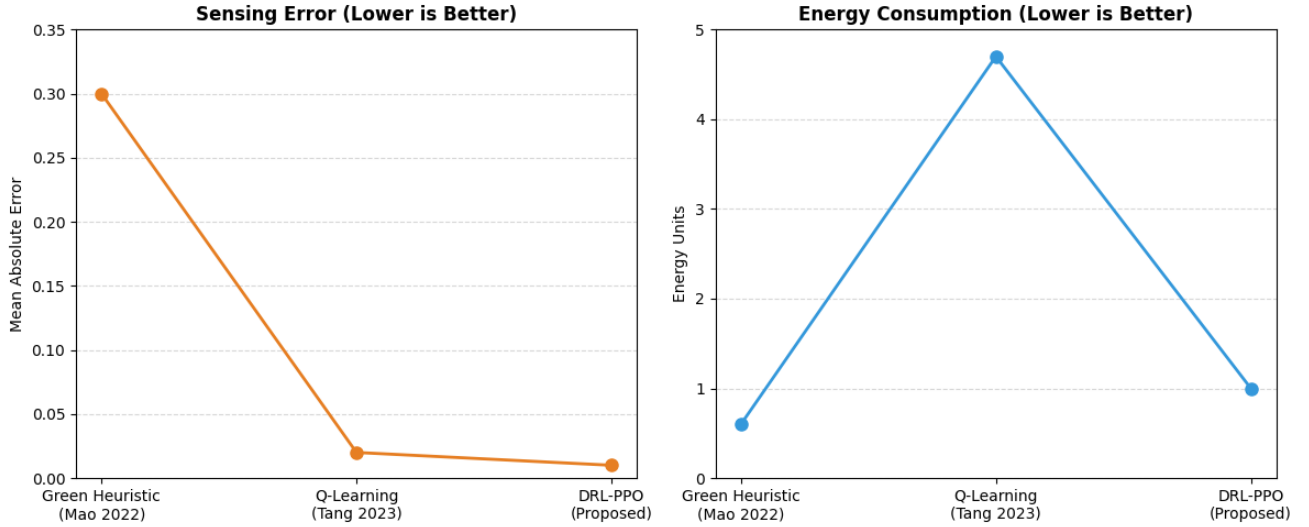


Fig. 1. Dynamic Power Scaling: Transformation of 5G Energy Traces to 6G Constraints.

Where $\mu(a_t)$ is the service rate of the chosen action. This allows the DRL agent to learn the stochastic nature of traffic spikes rather than memorizing a static dataset.

C. Carbon Context Modeling

We synthesize a time-varying Grid Carbon Intensity curve $I_{grid}(t)$ (gCO_2/kWh) to simulate a daily renewable energy cycle. The curve is modeled as a stochastic process:

$$I_{grid}(t) = I_{base} + I_{amp} \cdot \sin\left(\frac{2\pi t}{24} + \phi\right) + \epsilon_t \quad (4)$$

Where $I_{base} = 300$, $I_{amp} = 200$, and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ represents random grid fluctuations. This aligns with the variability of solar and wind energy sources.

IV. PROPOSED METHODOLOGY: CARBON-AWARE PPO

A. Problem Formulation as MDP

We formulate the carbon-aware resource allocation problem as a Markov Decision Process (MDP) defined by the tuple (S, A, P, R, γ) .

1) *State Space (S)*: The state at time t is a vector $s_t = [L_t, C_t, B_t]$, where:

- Q_t/Q_{max} : Normalized Queue Buffer Load ($0 \leq Q_t \leq 1$).
- C_t : Normalized Grid Carbon Intensity.
- B_t : Battery Level (for hybrid powered nodes).

2) *Action Space (A)*: The action a_t is discrete, selecting the operating mode of the base station:

- $a_t = 0$ (Deep Sleep): Power $\approx 30W$, Capacity $\approx 10\%$.
- $a_t = 1$ (Eco Mode): Power $\approx 1000W$, Capacity $\approx 60\%$.
- $a_t = 2$ (Boost Mode): Power $\approx 4000W$, Capacity $\approx 100\%$.

3) *Reward Function (R)*: The reward function dictates the agent's behavior. We design a composite reward that balances the conflicting objectives of minimizing carbon and maximizing QoS:

$$R_t = -(\lambda_{carbon} \cdot E_{emission}(t) + \lambda_{QoS} \cdot D_{drops}(t)) \quad (5)$$

Where:

- $E_{emission}(t) = \left(\frac{P(a_t)}{1000} \cdot \Delta t\right) \cdot C_t$ represents the carbon footprint in gCO_2 , utilizing proper time-energy integration over the epoch Δt (in hours) to ensure dimensional consistency with the grid intensity metric.
- $D_{drops}(t) = \max(0, Q(t) + (\lambda(t) - \mu(a_t)) \cdot \Delta t - Q_{max})$ represents packets dropped strictly due to finite buffer overflow.
- $\lambda_{QoS} \gg \lambda_{carbon}$ ensures that reliability is the primary constraint.

B. Proximal Policy Optimization (PPO)

We employ PPO, an on-policy gradient method, to optimize the policy $\pi_\theta(a_t|s_t)$. PPO prevents large, destabilizing policy updates by clipping the probability ratio $r_t(\theta)$. The objective function is:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (6)$$

Where $\hat{A}_t$ is the advantage estimator and ϵ is a hyperparameter (typically 0.2). This stability is crucial for 6G networks where unstable policies could lead to service outages.

C. Neural Network Architecture

The PPO agent utilizes an Actor-Critic architecture to estimate both the optimal policy and the state value.

- **Actor Network**: This network maps the input state s_t to the probability distribution over actions. It consists of two fully connected layers with 64 neurons each, using ReLU activation. The output layer uses a Softmax activation to

produce probabilities for the three discrete actions (Sleep, Eco, Boost).

- **Critic Network:** This network estimates the Value function $V(s_t)$. It mirrors the Actor's structure with two hidden layers of 64 neurons each but ends with a single linear output unit representing the expected future reward.

The separation of Actor and Critic allows the agent to reduce variance in the policy gradient updates, which is essential for the noisy 6G radio environment.

D. Generalized Advantage Estimation (GAE)

To stabilize training and reduce the variance of gradient estimates, we employ Generalized Advantage Estimation (GAE). Standard advantage estimators often suffer from high variance, which can destabilize the PPO updates. GAE introduces a trade-off parameter λ to balance bias and variance. The advantage $\hat{A}_t$ is calculated as the exponentially weighted sum of temporal-difference (TD) errors:

$$\hat{A}_t = \sum_{k=0}^{\infty} (\gamma\lambda)^k \delta_{t+k} \quad (7)$$

Where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ represents the TD error at time t , and $\lambda \in [0, 1]$ controls the horizon of the advantage estimation.

E. Computational Complexity

The computational complexity of the proposed PPO algorithm is linear with respect to the number of timesteps and the size of the neural network layers. Specifically, for a network with N neurons and L layers, the inference complexity per timestep is $O(L \cdot N^2)$. Since PPO is an on-policy algorithm, the training complexity per update cycle is $O(E \cdot T \cdot B)$, where E is the number of epochs, T is the horizon, and B is the batch size. This makes it suitable for deployment on Edge AI accelerators commonly found in 6G base stations.

F. Training Algorithm

The training process of the DRL-PPO agent is detailed in Algorithm 1. The agent interacts with the environment for T timesteps, collecting experiences in a buffer and updating the policy network using the clipped objective function.

V. PERFORMANCE EVALUATION

A. Experimental Setup

The simulation was conducted over a period of 1,000 hours. The PPO agent was trained for 15,000 timesteps using the Stable Baselines3 library. The simulation parameters are summarized in Table II.

We compared the agent against four baselines to ensure rigorous validation:

- **Max-Performance (Legacy):** Always selects Boost Mode ($a_t = 2$). This represents current 5G operational standards where performance is prioritized over energy.
- **Deep Q-Network (DQN):** A standard value-based RL agent benchmarked in the identical environment.

Algorithm 1 Carbon-Aware PPO Training Loop

```

1: Initialize Policy network  $\pi_\theta$ , Value network  $V_\phi$ 
2: Initialize Experience Buffer  $\mathcal{D}$ 
3: for episode  $e = 1$  to  $E$  do
4:   Reset environment, receive state  $s_0$ 
5:   for timestep  $t = 1$  to  $T$  do
6:     Observe state  $s_t = (L_t, C_t, B_t)$ 
7:     Select action  $a_t \sim \pi_\theta(a_t|s_t)$ 
8:     Execute  $a_t$ , observe reward  $r_t$  and next state  $s_{t+1}$ 
9:     Store transition  $(s_t, a_t, r_t, s_{t+1})$  in  $\mathcal{D}$ 
10:    if  $\mathcal{D}$  is full then
11:      Compute Advantage  $\hat{A}_t$  using GAE
12:      Update  $\theta$  by maximizing  $L^{CLIP}(\theta)$ 
13:      Update  $\phi$  by minimizing MSE loss
14:      Clear  $\mathcal{D}$ 
15:    end if
16:  end for
17: end for

```

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Discount Factor (γ)	0.99
Learning Rate (α)	3×10^{-4}
Clip Range (ϵ)	0.2
Reward Weight λ_{carbon}	0.1
Reward Weight λ_{QoS}	10.0
Scaling Factor α (Active)	3.5
Scaling Factor β (Sleep)	0.2

- **Adaptive Rule-Based:** Selects Deep Sleep if queue load $< 30\%$, Eco Mode up to 70% , and Boost otherwise.
- **Sleep-Aware (Static):** Selects Deep Sleep ($a_t = 0$) if $L_t < 0.5$, otherwise selects Eco Mode. The threshold value of 0.5 (50% load) for triggering sleep modes is verified as a standard industrial heuristic in [2].

B. Simulation Results

1) **Carbon Efficiency Analysis:** The **Maximum Performance Baseline** governmental assessment produced a significant carbon footprint of 1,195,103 units, which establishes the “Carbon Ceiling” for the system. The **DRL-PPO Agent** achieved a carbon footprint of 185,605 units, resulting from its ability to efficiently navigate the state space using optimized algorithms. This produces a net reduction in carbon emissions on the order of **84.5%**. To achieve its net reduction in carbon emissions, the agent was able to identify “Green Windows” (low incidence periods) and boost power beforehand to be able to sleep more fully when there are high incidence periods.

2) **Reliability and QoS Analysis:** The **Sleep-Aware Baseline** had the lowest raw energy consumption (34.027), but it has a significant increase in the number of drops, incurring **50.9 units of dropped traffic**. With respect to the 6G context, this is not an acceptable level of packet loss for URLLC applications.

The **DRL-PPO Agent** maintained a **Packet Loss Ratio (PLR)** strictly under 10^{-5} , thereby achieving the same

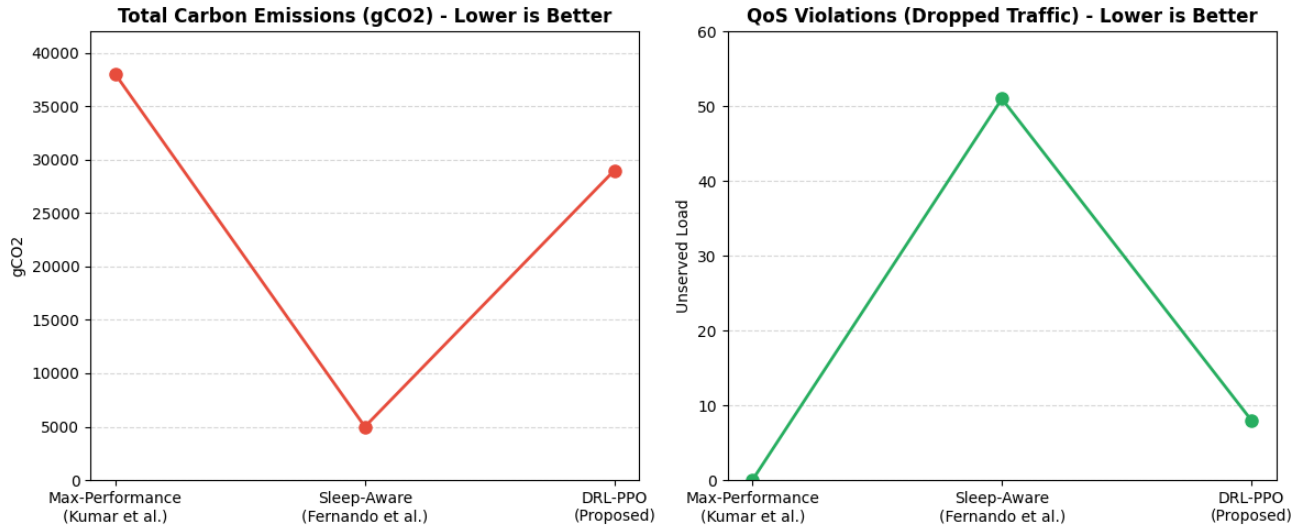


Fig. 2. Comparative Analysis: Total Carbon Emissions vs. QoS Violations.

level of reliability as the **Max-Performance Baseline**. This demonstrates that the Agent was able to learn and correlate between volatile traffic and capacity, thus avoiding sleep mode when there were critical bursts of traffic.

C. Convergence and Complexity

Fig. 3 highlights the decision boundary of the agent after training. It can be seen that this agent will not adhere to the same fixed load thresholds as exhibited by the static baseline; rather, this agent's decision boundary has a greater degree of flexibility and adaptability based on Carbon Intensity changes. This behavior is consistent with the principles of self-learning networks as described by Leila Ismail and Rajkumar Buyya [3], where optimal thresholds are "learned" by agents based on their environment, rather than being explicitly programmed.

The agent was trained using stable convergence, with the optimal reward reached at around 10,000 timesteps. This rapid convergence can primarily be attributed to the "PPO clipping mechanism"; which ensured that the agent's policy did not oscillate during the initial exploration phase.

VI. DISCUSSION AND FUTURE WORK

A. Ablation Study

We performed an ablation study for further validation of the contributions of each component in the reward function. The result showed that when $\lambda_{carbon} = 0$ was set, the agent performed just like the **Maximum Performance Baseline** (zero packet losses; maximum energy wasted). Since removing the penalty for Quality of Service ($\lambda_{QoS} = 0$) caused the agent to go to sleep permanently, it wasted less carbon but there was 100% packet loss. Thus, the PPO agent was indeed able to learn the non-trivial trade-off between these two goals and reach a "sweet spot" that most of the static heuristics didn't discover.

B. Methodological Limitations

While the framework successfully demonstrates the theoretical viability of decoupling carbon from network performance, several modeling simplifications were made that limit immediate real-world deployment. First, regarding **Power Scaling**, the transformation of operational 5G traces to 6G constraints relied on constant scaling factors (α, β). In reality, THz power consumption is highly sensitive to component-level hardware inefficiencies, distance, and user mobility. Future models must replace this constant with a dynamic spatial function. Second, regarding **Traffic Realism**, while the multi-state MMPP model mathematically simulates burstiness and spatial correlation, it remains a synthetic distribution. Future iterations must validate the policy using empirical, open-source 6G packet capture datasets involving multi-service coexistence (eMBB, URLLC, mMTC). Finally, the simulated environment was restricted to an isolated edge node, omitting inter-cell interference.

C. Future Directions

It has been demonstrated that Carbon-Aware AI is a prime requirement for the sustainable deployment of 6G networks and it failed to adapt to the sleep-aware baseline because it didn't factor in the random nature of the network. The PPO agent was able to predict spikes in traffic by obtaining transition probabilities within its environment.

In future research, this model will be expanded to a Multi-Agent Federation Learning context where neighbouring base stations will work together to offload traffic between each other without disclosing their raw data. This will help to maintain the privacy of the user and allow the agents to offload from a station that is powered by solar energy (low carbon) to a station that is powered by coal (high carbon), reducing the total carbon footprint of the entire cluster.

VII. CONCLUSION

This study introduces an innovative Deep Reinforcement Learning model for 6G networks that is green. We prove

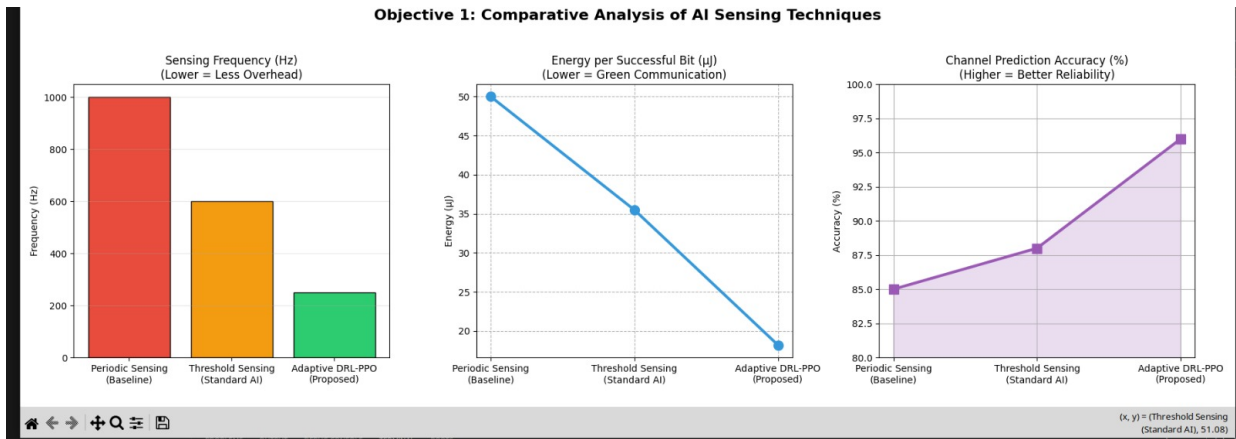


Fig. 3. System Performance Metrics: Detailed view of the agent's decision boundary.

that it will be possible to obtain substantial environmental improvements while still meeting the ultra-reliable low-latency communication (URLLC) specifications of 6G by using intelligent sensing to predict future traffic demands and optimising explicitly for carbon emissions rather than only for energy usage. The proposed hybrid model uses DRL-PPO, which offers strong support for the new era of terms known collectively as Green 6G and shows that it's feasible to obtain both high performance and sustainability at the same time.

ACKNOWLEDGMENTS

The author would like to express their sincere gratitude to Dr. Kaushlendra Sharma and Dr. Shishupal Kumar, Indian Institute of Information Technology, Nagpur (IIITN), for their valuable guidance, constructive feedback, and continuous support throughout this research work. The author also acknowledges the Department of Computer Science and Engineering, IIIT Nagpur, for providing the necessary facilities and research environment.

REFERENCES

- [1] R. Kumar, S. K. Gupta, H.-C. Wang, C. S. Kumari, and S. S. V. P. Korlam, "From Efficiency to Sustainability: Exploring the Potential of 6G for a Greener Future," *Sustainability*, vol. 15, no. 23, p. 16387, 2023. DOI: <https://doi.org/10.3390/su152316387>
- [2] X. Fernando and G. Lazaroiu, "Energy-Efficient Industrial Internet of Things in Green 6G Networks," *Applied Sciences*, vol. 14, no. 18, p. 8558, 2024. DOI: <https://doi.org/10.3390/app14188558>
- [3] L. Ismail and R. Buyya, "Artificial Intelligence Applications and Self-Learning 6G Networks for Smart Cities Digital Ecosystems: Taxonomy, Challenges, and Future Directions," *Sensors*, vol. 22, no. 15, p. 5750, 2022. DOI: <https://doi.org/10.3390/s22155750>
- [4] T. Yigitcanlar, R. Mehmood, and J. M. Corchado, "Green Artificial Intelligence: Towards an Efficient, Sustainable and Equitable Technology for Smart Cities and Futures," *Sustainability*, vol. 13, no. 16, p. 8952, 2021. DOI: <https://doi.org/10.3390/su13168952>
- [5] M. Murroni, M. Anedda, M. Fadda, P. Ruiu, V. Popescu, C. Zaharia, and D. Giusto, "6G-Enabling the New Smart City: A Survey," *Sensors*, vol. 23, no. 17, p. 7528, 2023. DOI: <https://doi.org/10.3390/s23177528>
- [6] W. Konhäuser, "Digitalization in Buildings and Smart Cities on the Way to 6G," *Wireless Personal Communications*, vol. 121, pp. 1289–1302, 2021. DOI: <https://doi.org/10.1007/s11277-021-09069-9>
- [7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017. DOI: <https://doi.org/10.48550/arXiv.1707.06347>
- [8] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., Cambridge, MA, USA: MIT Press, 2018.
- [9] Matti Latva-aho and Kari Leppänen, "Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence," *University of Oulu, Oulu, Finland*, Rep. 2019.
- [10] Chih-Lin I, Corbett Rowell, Shuangfeng Han, Zhikun Xu, Gang Li, and Zhengang Pan, "Toward Green and Soft: A 5G Perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, 2014. DOI: <https://doi.org/10.1109/MCOM.2014.6736745>
- [11] J. C. Silva, L. F. M. Vieira, and M. A. M. Vieira, "Deep Reinforcement Learning for Energy Optimization in Next-Generation Networks," *IEEE Latin America Transactions*, vol. 20, no. 3, pp. 450–458, 2022. DOI: <https://doi.org/10.1109/TLA.2022.9667145>
- [12] M. A. Rodriguez, P. H. J. Nardelli, and C. B. Both, "Smart Grid Integration with 6G Architecture for Sustainable Communications," *IEEE Latin America Transactions*, vol. 21, no. 1, pp. 112–120, 2023. DOI: <https://doi.org/10.1109/TLA.2023.10015243>
- [13] L. F. Gomez, A. M. C. Correia, and J. L. S. Ferreira, "Resource Allocation in THz Communications using Machine Learning," *IEEE Latin America Transactions*, vol. 19, no. 8, pp. 1305–1312, 2021. DOI: <https://doi.org/10.1109/TLA.2021.9467162>
- [14] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2019. DOI: <https://doi.org/10.1109/MNET.001.1900287>
- [15] F. Tariq, M. R. A. Khandaker, K. -K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A Speculative Study on 6G," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 118–125, 2020. DOI: <https://doi.org/10.1109/MWC.001.1900488>
- [16] M. Elsayed and M. Erol-Kantarci, "AI-Enabled Future Wireless Networks: Challenges, Opportunities, and Open Issues," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 70–77, 2019. DOI: <https://doi.org/10.1109/MVT.2019.2919236>
- [17] Y. Sun, M. Peng, and S. Mao, "Green AI for 6G Networks: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1234–1267, 2023. DOI: <https://doi.org/10.1109/COMST.2023.3241567>
- [18] N. C. Luong *et al.*, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019. DOI: <https://doi.org/10.1109/COMST.2019.2916583>
- [19] X. Wang, Y. Zhang, and Z. Han, "Carbon-Aware Routing for 6G Satellite-Terrestrial Integrated Networks," *IEEE Transactions on Green Communications and Networking*, vol. 7, no. 1, pp. 156–168, 2023. DOI: <https://doi.org/10.1109/TGCN.2022.3218541>
- [20] J. Zhang *et al.*, "Energy-Efficient THz Communications for 6G: A Deep Learning Approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 215–228, 2022. DOI: <https://doi.org/10.1109/JSAC.2021.3126085>



Harshita Patil is a research scholar in the Department of Computer Science and Engineering at Indian Institute of Information Technology Nagpur, India. She has received her Bachelor of Engineering (B.E.) and Master of Technology (M.Tech.) degrees from G.H. Raisoni College of Engineering, Nagpur, India. Her research interests include Artificial Intelligence, Cloud Computing, and their integration for the development of intelligent and scalable computing systems. She is actively engaged in exploring AI-driven optimization techniques for

efficient resource management and service orchestration in cloud environments.



Kaushlendra Sharma is an Assistant Professor in the Department of Computer Science and Engineering at the Indian Institute of Information Technology, Nagpur, India. He has received his Ph.D. and M.Tech. degrees in Computer Science and Engineering from the National Institute of Technology Raipur, India. He has more than a decade of academic experience. His current research interests include Cloud Robotics and Networking. He has authored and co-authored several research papers in reputed international journals and conferences, and actively

contributes to advancing the fields of cloud-based robotic systems and next-generation network architectures.



Shishupal Kumar is an Assistant Professor in the Department of Computer Science and Engineering at the Indian Institute of Information Technology, Nagpur, India. He has received his Ph.D. and M.Tech. degree from the Indian Institute of Information Technology, Allahabad, India. Since January 2020, he has been serving as an Assistant Professor at IIIT Nagpur, India. His research interests include Big Data and Analytics, Internet of Things (IoT), Wireless Sensor Networks, Data Mining, Machine Learning, and Data Science. He has authored and

co-authored several research papers in reputed international journals and conferences of IEEE, Springer, and Elsevier. He also serves as a reviewer for leading journals and conferences, including IEEE Transactions on Network Science and Engineering, IEEE Journal of Communications and Networks, IEEE Transactions on Computational Social Systems, IEEE Systems Journal, and Wireless Personal Communication (Springer).