

Comparing Bandwidth Allocation of Deep Reinforcement Learning and OFDMA in Wireless Networks: A Case of Wi-Fi 7 (IEEE 802.11be)

Daniel Ruiru^{1,2}[0000–0003–3197–8821], Nicolas Jouandeau²[0000–0001–6902–4324], and
Dickson Owuor¹[0000–0002–0968–5742]

¹ SCES, Strathmore University, Nairobi, Kenya

druiru@strathmore.edu & dowuor@strathmore.edu

² LIASD, Université Paris 8, Paris, France

nicolas.jouandeau@univ-paris8.fr & daniel.kiguru@etud.univ-paris8.fr

Abstract. This paper compares Deep Reinforcement Learning (DRL) agents with OFDMA in allocating bandwidth. OFDMA as used in Wi-Fi 7 (802.11be) is presented in this paper where its primarily characterized by flexible Resource Units (RUs) that can be allocated to different devices. To simulate realistic, real-world scenarios, dynamic and volatile user conditions are incorporated into the experiments. In the results, a Transformer based Deep Q-Network outperforms all other algorithms. This algorithm gets the best adaptability score (92%) and volatility handling capability score (89%). These values were the ultimate comparison methods following a series of evaluations that tested all the algorithms on metrics of throughput, fairness, delay, energy efficiency, and satisfaction. This paper concludes that DRL algorithms are ideal for dynamic and volatile environments while OFDMA techniques are good for static and predictable conditions.

Keywords: Deep Reinforcement Learning · OFDMA · Wi-Fi 7 · Bandwidth · Throughput.

1 Introduction

Wireless technology has improved significantly in the last few years. At the same time, demands for wireless technology have grown significantly to challenge the existing methods of service delivery. Today, the WiFi market is estimated to be worth over \$ 4.9 trillion owing to the advancements in the field trying to meet user demands [1]. This demand is particularly high in the fields of industrial automation where motion flexibility in devices like sensors, controllers, actuators and mobile users is needed [2]. Such use cases tend to have dense networks that require better quality of service. Additionally, they generate large volumes of data that is accompanied by heavy interferences [3]. These networks thus require efficient resource allocation and management to enable effective use.

The IEEE 802.11 standard tries to fulfill such demands by continuously implementing techniques and technologies to improve Wireless Local Area Networks (WLANs).

Introduced in 1997, the IEEE 802.11 standards as summarized in table 1 aimed to facilitate efficient device communication in wireless networks to improve user experiences [4]. Current standards (Wi-Fi 6 and 7), try to meet these original objectives by offering better speeds and smoother connections in dense networks. While they achieve their objectives through techniques like Orthogonal Frequency-Division Multiple Access (OFDMA) [5], there is still more to be done as the demands for connectivity grow. For one, there is need to develop better bandwidth allocation methods while minimizing the overhead costs such as advanced routers and switches. There is also a need to manage the ever growing demand for computing power to access every subsequent network generation [6].

Table 1: Wi-Fi generations

Name	IEEE standard	Frequency band(s)
Wi-Fi 1	IEEE 802.11b	2.4 GHz
Wi-Fi 2	IEEE 802.11a	5 GHz
Wi-Fi 3	IEEE 802.11g	2.4 GHz
Wi-Fi 4	IEEE 802.11n	2.4 GHz, 5 GHz
Wi-Fi 5	IEEE 802.11ac	5 GHz
Wi-Fi 6	IEEE 802.11ax	2.4 GHz, 5 GHz (6 GHz)
Wi-Fi 7	IEEE 802.11be	2.4 GHz, 5 GHz, 6 GHz

This paper contributes to these demands by comparing existing methods with Deep Reinforcement Learning (DRL). It compares OFDMA in Wi-Fi 7 with DRL algorithms in assigning bandwidth using the metrics of throughput, allocation fairness, delay and energy efficiency.

2 Related Works

2.1 The Resource Allocation Problem

Wi-Fi is one of the primary wireless technologies that carries IP traffic. IP traffic includes any data that is text, images, voice and videos which is most of the communication observed today [7]. While the modern Wi-Fi implementation dates back to 1997 after the introduction of IEEE 802.11, its existence can be traced back to 1971 with similar objectives as those of today i.e. flexibility and efficiency [8]. These objectives can be better considered with an effective management of the transmission process, a key resource allocation problem.

Wi-Fi 7 aims to manage this problem using OFDMA just like Wi-Fi 6 but with a few improvements in the allocation process. OFDMA works by splitting the transmission channel into subcarriers to connect multiple endpoints (devices) at the same time [2]. With a Wi-Fi 7 router, a single transmission can communicate with multiple devices without the devices waiting for turns as data is sent across networks [9].

2.2 OFDMA RU Allocation

The hallmark of Wi-Fi 7 is OFDMA which allocates time and frequency domain to multiple users. Cisco uses the analogy of a delivery truck to explain the technology in which prior to Wi-Fi 6 each truck carried only one user's cargo but now is loaded with multiple users' cargo [10]. OFDMA improves wireless access allocations by dividing the Wi-Fi channel into smaller sub-channels called Resource Units (RUs). These RUs are assigned to different devices simultaneously, instead of one device taking up the whole channel. The key difference between its application in Wi-Fi 6 and 7 is on the RUs and the size of the sub-channels. In the former, there are fixed RUs which result in rigid allocations. Due to the limited frequency band, the biggest partition of the sub-channel can only go to a maximum of 160 MHz [12]. Wi-Fi 7 has a flexible RU allocation which means devices can be assigned to multiple RUs (even noncontiguous ones) while also being able to skip occupied channels [11]. This generation also has wider channels of up-to 320 MHz (due to a bigger overall spectrum band).

The RUs are a relative concept as there are various sizes. A single RU can contain 26 (2 MHz), 52 (4 MHz wide each), 106 (8 MHz wide each), and up to 996 (78 MHz) tones, which means the maximum size of an RU depends of the channel bandwidth (20, 40, 80, 160 and up to 320 MHz for Wi-Fi 7). For the number of RUs, the more they are, the higher the multi-user processing efficiency as better throughput is provided to devices [11]. Fig. 1 summarizes this key concept of RU as used in OFDMA based on a 20 MHz channel¹.

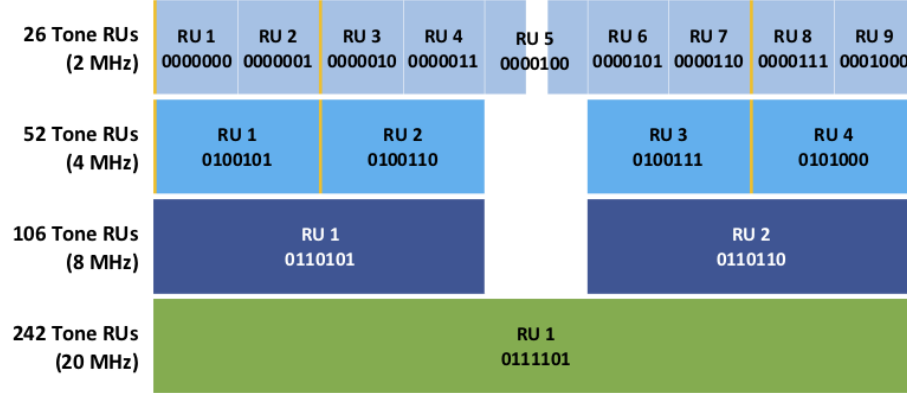


Fig. 1: Resource units distribution in 20 MHz channel

2.3 Deep Reinforcement Learning and Transformers

Classic reinforcement learning (RL) uses tabular methods (Fig.2). These methods excel in low-dimensional problems because state spaces are small and fully captured in compact feature sets [13]. Complex problems like resource allocation require techniques

¹ A 20 MHz channel can carry up to 9 concurrent users using 26 RUs.

that can sufficiently represent high-dimensional state spaces. Deep neural networks have excellent representation capacity [14]. Deep networks also can automatically discover features from high-dimensional inputs which make them ideal to combine with RL for complex problems.

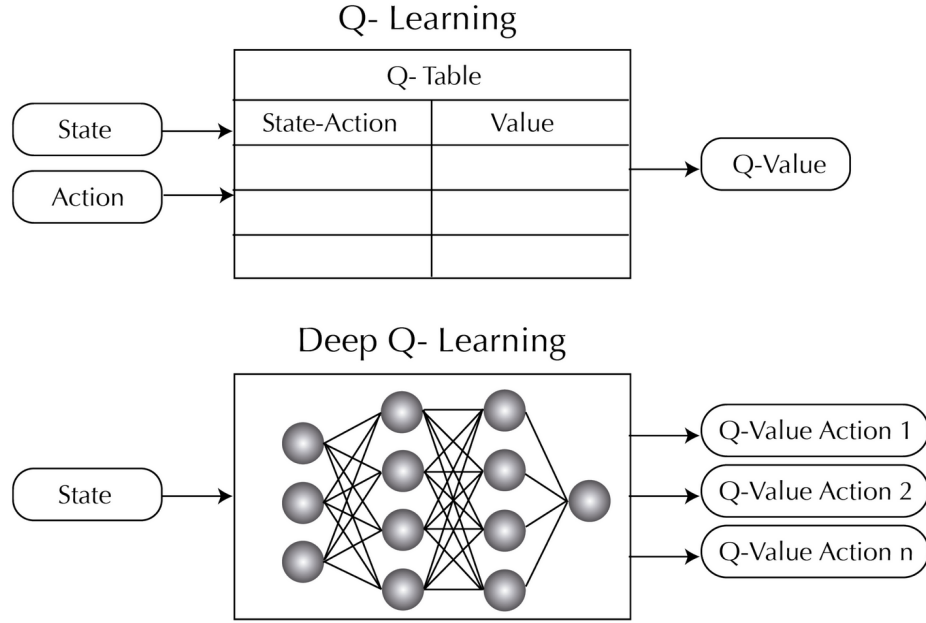


Fig. 2: Comparison of tabular Q-learning (classic RL) and deep Q-learning (DRL)

The key breakthrough for combining deep neural networks and RL came with the introduction of deep Q-networks (DQN). DQN demonstrated human-level performance on multiple Atari 2600 video games with only inputs of raw pixels and games scores [15]. Researchers rapidly expanded this new field of Deep RL (DRL) by introducing novel algorithms using techniques like policy gradient [16], actor-critic framework [17] and distributional RL [18].

Recent advances in DRL have incorporated the transformer architecture. Transformers, introduced by [19] offer a powerful alternative for state representation by modeling sequential and multiple dependencies using an attention mechanism [20]. Transformers leverage self-attention to capture temporal correlations and high-dimensional constraints. This gives better scalability, faster convergence and superior handling of dynamic state space fluctuations [20]. The combination of transformers and DRL (Fig.3) is effective in complex environments like communication networks where global attention, spatial awareness and dynamic handling of input variables is needed.

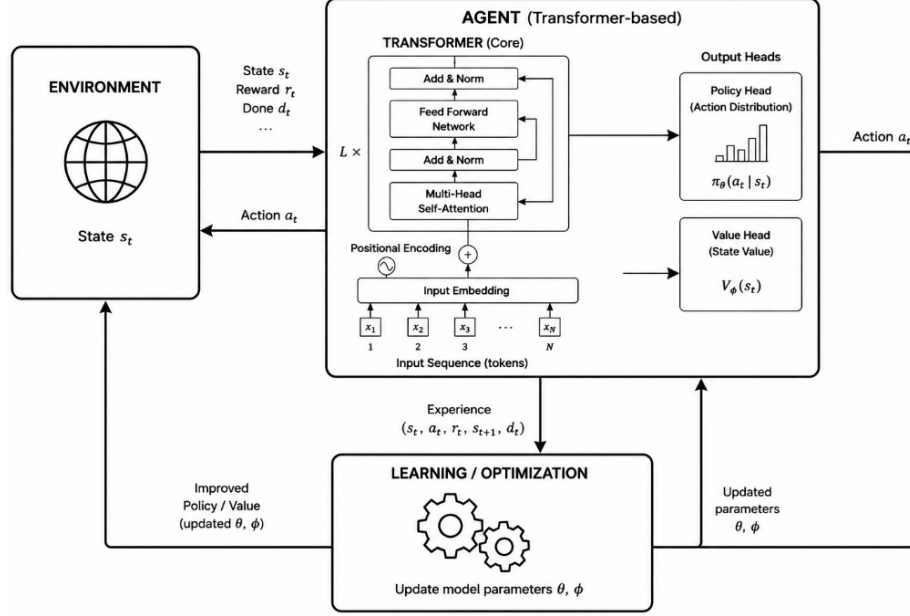


Fig. 3: Conceptual architecture illustrating the combination of transformers with deep reinforcement learning

2.4 DRL and Networking

Transformer based DRL is fairly new to the networking domain. Most of the existing studies still use conventional DRL methods such as policy gradient and actor-critic. For example, routing optimization has been done using self-learning methods [21]. In bandwidth allocation, DQN and policy based DRL have been used to allocate frequency and time resources in wireless networks [22]. [23] used a proximal policy optimization algorithm for dynamic power and bandwidth allocation in multi beam satellites. [24] used different learning techniques to train centralized RL agents on how to solve non-convex optimization problems in OFDMA. These agents controlled scheduling problems for down-link tasks. [25] used DRL agents like DQN and PPO for resource allocation in densely deployed medical sensors while trying to solve overlapping interference problems. While these classical DRL techniques work, they often fall short due to representation limitations and the need to handle dynamic state space changes. Transformers can be used to manage these drawbacks.

3 Methodology

3.1 Defining the Experiments

The proposed work aimed to compare Wi-Fi resource allocation methods. The experiments were setup in NS3Gym, an integrated framework that can simulate various net-

work problems [26]. Dynamic bandwidth requirement scenarios were emphasized to emulate realistic network conditions, with each strategy designed to improve throughput, fairness and energy efficiency across different user stations (devices). This dynamic nature meant that the experiments incorporated different bandwidth requirements and volatility modeling which led to the addition of other evaluation metrics, i.e. satisfaction, adaptability, and recovery speed. The experiments done tested for different traffic patterns (periodic, bursty, random and gradual) to provide deeper insight to how the proposed techniques adapt to fluctuating user demands in comparison to OFDMA.

3.2 DRL for Allocating Bandwidth

Four algorithms were implemented:

- **Baseline Round Robin** - Simple and fixed allocation (non-learning) for comparative analysis.
- **Standard DQN** - A vanilla Deep Q-Network with an experience replay [27]. An ϵ -greedy exploration policy with fully connected neural network. A 4-layer network was used having a configuration of 40, 64, 64 and 90. This configuration offers excellent dimensional expansion and high-capacity classification.
- **Improved DQN** - An enhanced DQN [27]² with Double DQN, Dueling Networks and Prioritized Experience Replay (whose experience sampling was varied). The Double DQN was used to reduce Q-value overestimation [28], while the Dueling architecture helped separate values and advantage streams [29]. The Prioritized Experience Replay had the goal of improving sampling from the replay buffer by picking important transitions more frequently [30].
- **Transformer DQN** - An attention based DQN having multi-head self attention for better state representation and positional encoding of temporal relationships [31]³. This agent also had an Actor-critic framework to separate policy from value network [32]. For better control, a continuous action space with 20 dimensions was used with this algorithm.

3.3 Environment Definition

All DRL algorithms were exposed to the same environment that started with 1 Access Points (AP) and 10 Stations (STA) (Only 1 AP because of the computational complexity of having 10 STAs with dynamic requirements) (Fig. 4 illustrates the environment). The state, action, transition and rewards were defined as follows:

² Mnih et al. [27] provide a clear description of the basic structure of DQN.

³ Vaswani et al. [31] give a detailed account of how transformers work.

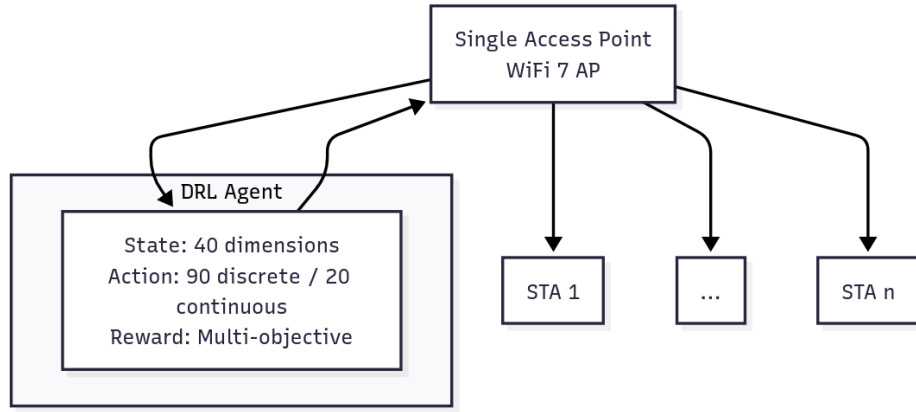


Fig. 4: High-level view of n-States

State Space For all the 10 stations, they were defined by; *throughput* (0-100 Mbps), *delay* (0 - 100 ms), *received signal strength (RSSI)* (-100 to -30 dBm) and *interference level* (0 to 50)⁴. Every agent received an observation vector of 40 numerical values leading to a state space of 40 dimensions.

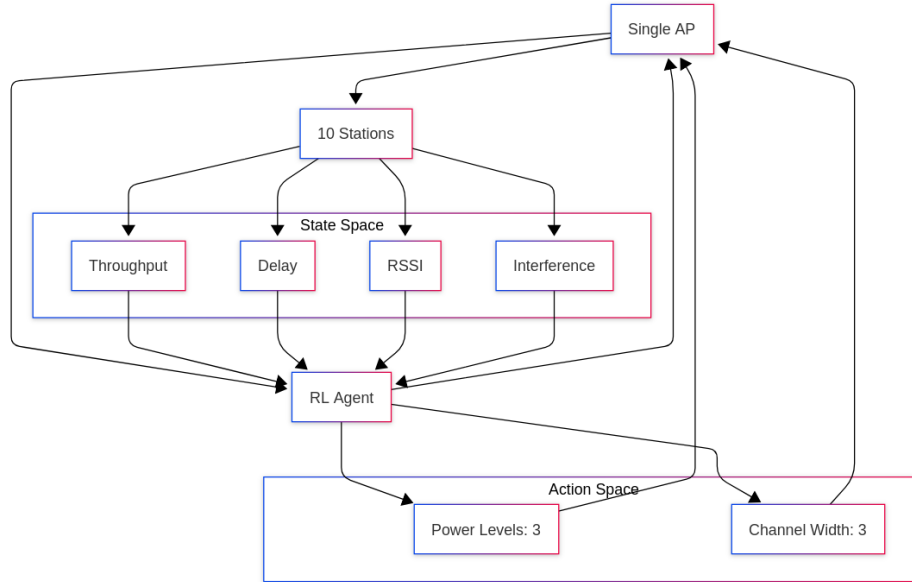


Fig. 5: State space and action space

⁴ 0 = no interference and 50 = heavy interference

Action Space Both discrete and continuous action spaces were available. The actions were subject to the power levels and channel width.

For the former, three power levels were provided. *Low power*⁵ (level 0), 10 - 15 dBm; *medium power*⁶ (level 1), 15 - 25 dBm and *high power*⁷ (level 2) 25 - 30 dBm.

Three channel widths were used. *Narrow channel*⁸ (level 0), 10 - 15 MHz; *medium channel*⁹ (level 1), 15 - 25 MHz and *wide channel*¹⁰ (level 2), 30 - 40 MHz.

With 10 STAs, an agent had access to 90 possible actions (Fig. 5). The goal of an agent was to learn trade-offs between power and width allocation.

Transition Probability NS3Gym provided the transition probability after an agent took an action from a given state. NS3Gym applied the allocation given and simulated a 20 ms real network scenario having; radio propagation model (how the signal traveled/interfered), packet traffic (real data flow between STAs and AP), channel conditions (noise, collisions and fading) and PHY/MAC layer behavior (based on extensions to 802.11ax simulations as 802.11be was not supported in NS3 at the time of the experiments).

Reward The reward was defined using six parameters. Throughput, fairness index¹¹, satisfaction, average delay, energy efficiency and an adaptability bonus. The reward was defined as shown in equation 7, using individual definitions, equations 1 to 6.

$$\text{Throughput} = \frac{\sum_i \text{throughput}_i}{1000} \quad (1)$$

$$\text{Fairness} = \frac{(\sum_i \text{throughput}_i)^2}{10 \times \sum_i \text{throughput}_i^2} \quad (2)$$

$$\text{Req_Satisfaction} = \frac{1}{10} \times \sum_i \left[\min \left(\frac{\text{throughput}_i}{\text{requirement}_i}, 1.0 \right) \right] \quad (3)$$

$$\text{Delay_Penalty} = - \frac{\text{average_delay}}{100} \quad (4)$$

$$\text{Energy_Reward} = \frac{\left(\frac{\text{total_throughput}}{\text{total_power}} \right)}{10} \quad (5)$$

$$\text{Adaptability_Bonus} = 0.1 \times \left(1 - \frac{\text{time_to_adapt}}{\text{max_adaptation_time}} \right) \quad (6)$$

⁵ Short range and energy efficient

⁶ Balanced range and power

⁷ Long range and high energy use

⁸ Low throughput but robust

⁹ Balanced for throughput and reliability

¹⁰ High throughput but susceptible to interferences

¹¹ Using Jain's fairness index.

$$\begin{aligned}
\text{Reward} = & 0.30 \times \left(\frac{\sum_i \text{throughput}_i}{1000} \right) \\
& + 0.30 \times \left(\frac{(\sum_i \text{throughput}_i)^2}{10 \times \sum_i \text{throughput}_i^2} \right) \\
& + 0.20 \times \left(\frac{1}{10} \times \sum_i \min \left(\frac{\text{throughput}_i}{\text{requirement}_i}, 1.0 \right) \right) \\
& + 0.10 \times \left(-\frac{\text{average_delay}}{100} \right) \\
& + 0.10 \times \left(\frac{\text{total_throughput}}{\text{total_power}} \right) \\
& + 0.10 \times \left(1 - \frac{\text{time_to_adapt}}{\text{max_adaptation_time}} \right)
\end{aligned} \tag{7}$$

4 Experiments

A circular STA distribution with variable distances of 10 to 20m radius was considered to simulate realistic signal attenuation. For traffic patterns, heterogeneous UDP traffic was used. This traffic had variable packet sizes of 1024 - 1792 bytes, diverse transmission intervals (2 - 6ms) and mixed loads (10000 - 20000 packets/STA).

The training procedure used 500 episodes with evaluations done after 20 episodes. 100 different and random experiments were performed having diverse volatility and dynamic patterns. On average, the training time was 80-160 hours for CPU (11th Gen Intel® Core™ i7-11850H × 16) and 20-80 hours for GPU (NVIDIA T600).

The results were compared with those of OFDMA. The OFDMA method was implemented with 3 strategies:

- **Proportional Fairness** - with an objective of balancing fairness with efficiency i.e. maximize the throughput.
- **Max-Min Fairness** - A scheduling technique that maximizes fairness over total throughput.
- **QoS Aware Allocation** - assigned resources based on service request by STAs.

The comparison between OFDMA and DRL algorithms was based on five performance metrics. Total throughput, fairness index, delay performance, energy efficiency and adaptability.

The final comparisons were meant to test the following hypothesis:

1. *On average, DRL algorithms would achieve higher total throughput than OFDMA methods across the 100 experiments.*
2. *DRL algorithms would demonstrate superior adaptability.*
3. *QoS aware OFDMA would outperform DRL in application-specific scenarios.*

5 Results & Discussion

5.1 DRL Results

Table 2 shows some of the results. The Improved DQN and Transformer DQN outperformed the other agents. On average, the Transformer DQN had a better cumulative reward than Improved DQN. But the latter regularly challenged the former, particularly in the early training episodes. This could be attributed to Transformer’s properties as they need a longer horizon and more experiences to get good results. The performance of these agents extended to other metrics with comparable outcomes (table 2 & 3).

Table 2: Performance comparison of DRL algorithms

Algorithm	Avg Reward	Max Reward	Avg TP (Mbps)	Max TP (Mbps)
Baseline	479.75 $\pm$ 10.00	539.75	968.88 $\pm$ 20.40	1068.15
DQN	477.81 $\pm$ 110.32	531.15	955.40 $\pm$ 220.45	1059.21
Improved DQN	498.78 $\pm$ 18.37	545.63	997.94 $\pm$ 37.23	1086.60
Transformer DQN	505.70 $\pm$ 13.10	546.46	1011.22 $\pm$ 25.88	1091.36

Table 3: Performance Comparison of DRL Algorithms

Algorithm	Avg Fairness	Avg Delay (ms)	Energy Efficiency
Baseline	0.837 $\pm$ 0.072	11.45 $\pm$ 4.15	4.968 $\pm$ 0.142
DQN	0.787 $\pm$ 0.196	13.20 $\pm$ 5.50	4.806 $\pm$ 1.114
Improved DQN	0.845 $\pm$ 0.063	13.72 $\pm$ 4.89	5.002 $\pm$ 0.192
Transformer DQN	0.840 $\pm$ 0.088	13.03 $\pm$ 4.08	5.033 $\pm$ 0.199

5.2 DRL vs OFDMA

Transformer DQN showed superior performance in adaptability scenarios. It had the highest adaptability (92%) and volatility handling capability (89%) score. Improved DQN was second with 85% and 79% respectively (Fig. 6 & 7). These results were not surprising as the two algorithms did have better throughput, delay progression and energy efficiency.

A result that stood out for the DRL algorithms in the comparison experiments was the poor performance of DQN. The vanilla DQN performance seemed to degrade with time during these tests. In the end, its results fell short, even below those of the baseline agent.

OFDMA strategies were better in terms of fairness, consistently having the highest values. An unexpected outcome of this consistency was its advantage in dealing with volatility conditions as all OFDMA methods did perform well in this test (Fig. 6).

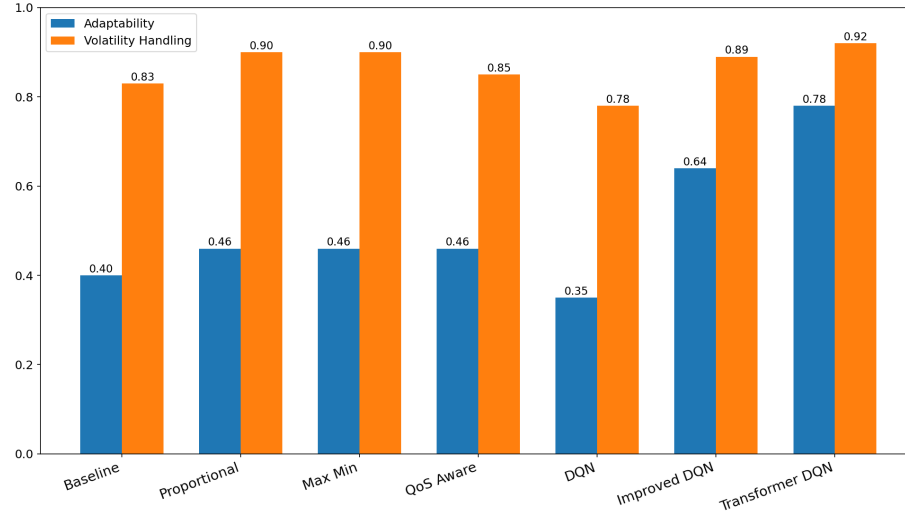


Fig. 6: Dynamic requirement handling score

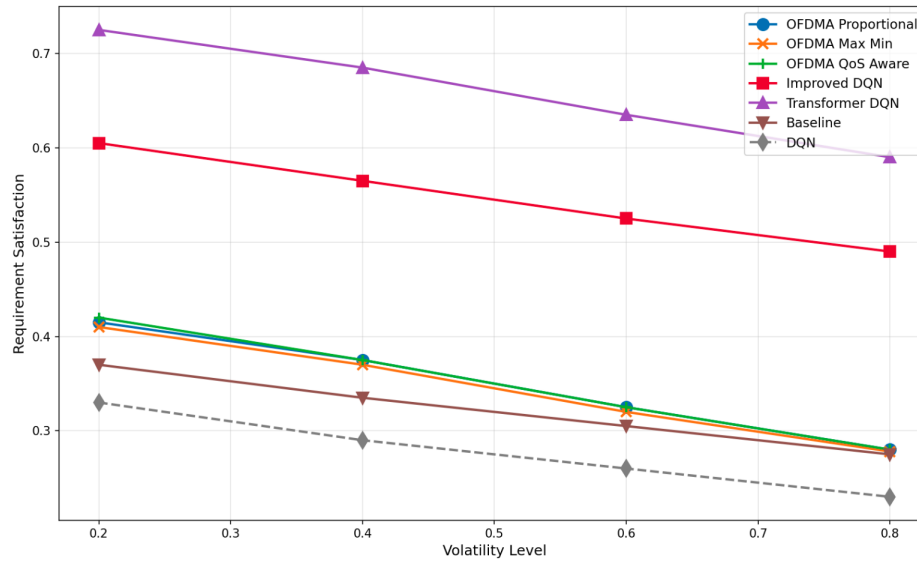


Fig. 7: Algorithms performance under different volatility levels

As for the hypothesis:

1. *Hypothesis 1 was validated as the DRL algorithms achieved higher throughput than OFDMA through the performance of Transformer DQN and Improved DQN. Transformer DQN had the best throughput over the 100 experiment (confidence level of 96%).*
2. *Hypothesis 2 was validated as Transformer DQN had the best adaptability score in volatile and dynamic environments.*
3. *Hypothesis 3 was validated as OFDMA methods outperformed DRL algorithms in some instance due to their consistent allocation that guaranteed a high fairness index.*

6 Conclusions

This paper compared DRL algorithms with OFDMA in allocating bandwidth in a Wi-Fi 7 scenario. From the initial experiments a Transformer based DQN outperformed all other DRL algorithms by getting better reward and higher values for other evaluation metrics. Improved DQN, followed closely with comparable results. In comparison tests, the DRL algorithms performed better than OFDMA methods in the dynamic and volatile conditions. This conclusion is based on the better performance in metrics of throughput progression, satisfaction, fairness, adaptability and efficiency. OFDMA was better in static conditions as its variations consistently gave a fair distribution of bandwidth.

Acknowledgment

This work was carried out using HPC resources from GENCI-IDRIS (Grant 2025-[AD010614071R2]). The researchers thank Strathmore University and LIASD Lab (Université Paris 8) for invaluable academic and institutional support during the research of this paper. The lead author gratefully acknowledges the French Embassy in Kenya for its generous support and facilitation of his doctoral studies.

References

1. Cisco, "The innovative past and brilliant future of Wi-Fi," 2024. [Online]. Available: <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2024/m11/the-innovative-past-and-brilliant-future-of-wi-fi.html>
2. E. Mozaffariahrar, F. Theoleyre, and M. Menth, "A survey of Wi-Fi 6: Technologies, advances, and challenges," *Future Internet*, vol. 14, no. 10, p. 293, 2022, doi: 10.3390/fi14100293.
3. D. Deng, S. Lien, J. Lee, et al., "On quality-of-service provisioning in IEEE 802.11ax WLANs," *IEEE Access*, vol. 4, pp. 6086–6104, 2016, doi: 10.1109/ACCESS.2016.2602522.
4. E. Oughton, W. Lehr, K. Katsaros, et al., "Revisiting wireless internet connectivity: 5G vs Wi-Fi 6," *Telecommun. Policy*, vol. 45, no. 5, pp. 102127, 2021, doi: 10.1016/j.telpol.2021.102127.

5. S. Kim, "Performance and simulation analysis of 802.11ax OFDMA in contention-driven scenarios," *PeerJ Comput. Sci.*, vol. 11, pp. e2687, 2025, doi: 10.7717/peerj-cs.2687.
6. C. Deng, X. Fang, X. Han, et al., "IEEE 802.11be Wi-Fi 7: New challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2136–2166, 2020, doi: 10.1109/COMST.2020.3012715.
7. K. Pahlavan and P. Krishnamurthy, "Evolution and impact of Wi-Fi technology and applications: A historical perspective," *Int. J. Wireless Inf. Netw.*, vol. 28, no. 1, pp. 3–19, 2021, doi: 10.1007/s10776-020-00501-8.
8. E. Firdus, R. Aghababayev, V. Aliyev, et al., "WiFi from past to today, consequences that can cause and measures of prevention from them, WiFi security protocols," in *Proc. E3S Web Conf.*, vol. 474, pp. 02004, 2024, doi: 10.1051/e3sconf/202447402004.
9. Intel, "What is Wi-Fi 6?," Sep. 2025. [Online]. Available: <https://www.intel.com/content/www/us/en/gaming/resources/wifi-6.html>
10. Cisco, "What is OFDMA?," Jun. 2025. [Online]. Available: <https://www.cisco.com/site/us/en/learn/topics/networking/what-is-ofdma.html>
11. X. Liu, Y. Dong, Y. Li, et al., "IEEE 802.11be Wi-Fi 7: Feature summary and performance evaluation," *arXiv*, 2023.
12. S. Tutelian, D. Bankov, D. Shmelkin, et al., "IEEE 802.11ax OFDMA resource allocation with frequency-selective fading," *Sensors*, vol. 21, no. 18, pp. 6099, 2021, doi: 10.3390/s21186099.
13. A. Barto, S. Bradtke, and S. P. Singh, "Learning to act using real-time dynamic programming," *Artificial Intelligence*, vol. 72, no. 1–2, pp. 81–138, 1995.
14. J. Terven, "Deep reinforcement learning: A chronological overview and methods," *AI*, vol. 6, no. 3, pp. 46, 2025.
15. V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Playing Atari with deep reinforcement learning," *arXiv*, 2013, doi: 10.48550/arXiv.1312.5602.
16. V. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Adv. Neural Inf. Process. Syst.*, vol. 12, 1999.
17. V. Mnih, A. Badia, M. Mirza, et al., "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1928–1937, 2016.
18. M. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 449–458, 2017.
19. C. Zhang and X. Li, "A transformer-based deep reinforcement learning method for controller parameter modulation in fault-tolerant control," *Mathematics*, vol. 14, no. 9, pp. 1409, 2026.
20. N. Luong, S. Feng, N. Hai, et al., "Transformer-enhanced reinforcement learning: Fundamentals and applications in communication networks," *arXiv*, 2026.
21. G. Lin, C. Hung, and C. Ke, "A deep reinforcement learning-based bandwidth demand-oriented routing in software-defined networking," *ICT Express*, 2025, doi: 10.1016/j.icte.2025.07.009.
22. R. Balakrishnan, K. Sankhe, V. Somayazulu, et al., "Deep reinforcement learning based traffic-and channel-aware OFDMA resource allocation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1–6, 2019.
23. J. Xu, Z. Zhao, L. Wang, et al., "A novel deep reinforcement learning architecture for dynamic power and bandwidth allocation in multibeam satellites," *Acta Astronaut.*, vol. 204, pp. 73–82, 2023, doi: 10.1016/j.actaastro.2022.12.018.
24. E. Bansbach, V. Eliachevitch, and L. Schmalen, "Deep reinforcement learning for wireless resource allocation using buffer state information," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1–6, 2021.
25. J. Huang, C. Yang, F. Yang, et al., "Deep reinforcement learning-based spectrum resource allocation for the web of healthcare things," *Digit. Commun. Netw.*, vol. 11, no. 3, pp. 671–680, 2025.

26. P. Gawłowicz and A. Zubow, “ns3-gym: Extending OpenAI Gym for networking research,” *arXiv*, 2018.
27. V. Mnih, K. Kavukcuoglu, D. Silver et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
28. H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-learning,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016.
29. Z. Wang, T. Schaul, M. Hessel et al., “Dueling network architectures for deep reinforcement learning,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1995–2003, 2016.
30. T. Schaul, J. Quan, I. Antonoglou, et al., “Prioritized experience replay,” *arXiv*, 2015.
31. A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.
32. V. Konda and J. N. Tsitsiklis, “Actor-critic algorithms,” in *Adv. Neural Inf. Process. Syst.*, vol. 12, 1999.