JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 129-143
https://publisher.uthm.edu.my/ojs/index.php/jscdm

An Actor-Critic Deep Reinforcement Learning Model with Energy-Awareness and Latency Minimization for Dynamic Spectrum Allocation in 6G-Enabled Aerial Mobile Wireless Networks

Ghaida Muttashar Abdulsahib[1], Kholoud Alkayid[2]*, Muhammad Asshad[3], Ahmed A.F Osman[4]*, Mohammed Awad Mohammed Ataelfadiel[4], Ashraf Mahrous Nour Zaher[5]

[1] *College of Computer Engineering, University of Technology, Baghdad 10066, IRAQ*
[2] *Department of Computer Science, Faculty of Information Technology, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, JORDAN*
[3] *Computer Networks, Faculty of Computer Studies, Arab Open University, OMAN*
[4] *Applied College, King Faisal University, P.O. Box 400, Al-Ahsa 31982, KINGDOM OF SAUDI ARABIA*
[5] *Department of Arabic Language, College of Arts, King Faisal University. Al Ahsa 31982, SAUDI ARABIA*

*Corresponding Author: k.kayid@ammanu.edu.jo, afadol@kfu.edu.sa
DOI: https://doi.org/10.30880/jscdm.2026.07.02.009

Article Info

Received: 13 January 2026
Accepted: 12 May 2026
Available online: 30 June 2026

Keywords

Aerial Mobile Wireless Network (AMWN), Machine Learning (ML), Deep Learning (DL), Reinforcement Learning (RL), Deep Reinforcement Learning (DRL), actor-critic architecture, 6G Networks, dynamic spectrum allocation, latency minimization, energy efficiency

Abstract

Aerial Mobile Wireless Networks (AMWN) play an important role in next-generation communication networks, especially in military operations, disaster recovery, and real-time surveillance. However, the highly dynamic nature of AMWN creates significant challenges for dynamic spectrum allocation (DSA), latency control, energy conservation, and spectrum optimization. These challenges become more critical in 6G-enabled environments that require ultra-low latency, high energy efficiency, and large-scale device connectivity. Deep Reinforcement Learning (DRL) offers a powerful approach by enabling real-time, data-driven decision-making in complex environments. This paper presents an Actor-Critic Deep Reinforcement Learning (AC-DRL) model adapted to a swarm-based behavior model for dynamic spectrum allocation with energy awareness and latency minimization in AMWN. The AC-DRL model is evaluated against a standard Q-learning approach using a custom dataset. The results show improved latency reduction, spectral efficiency, and energy consumption. Simulation results demonstrate up to 27% latency reduction and 22% improvement in energy efficiency compared with traditional models.

1. Introduction

Aerial Mobile Wireless Networks (AMWN), such as Flying Ad-Hoc Networks (FANETs), represent a rapidly evolving paradigm in which aerial nodes collaborate to enable seamless communication without relying on ground infrastructure. The AMWN will be employed in this paper as a general framework of aerial communications that incorporates various working modes of UAV communications. A FANET is said to be a

particular decentralized implementation of AMWN. Nevertheless, the suggested model does not presuppose a barebones FANET. Rather, the network is considered a cooperative aerial wireless system in which UAV nodes are regularly engaged in a shared-spectrum framework, enabling learning-based spectrum allocation and energy-conscious control. Thus, AMWN is applied as the model on the system level, whereas FANET is mentioned as an application case. These networks, typically composed of Unmanned Aerial Vehicles (UAVs), offer high mobility, flexible deployment, and improved Line-of-Sight (LoS) communication capabilities [1]. Research needs to present solutions for ultra-reliable low-latency communication and dynamic spectrum management with energy efficiency, since AMWN integration into 6G main infrastructure is underway.

Better dynamic spectrum allocation (DSA) methods than centralized planning are necessary to address AMWN requirements, given their rapid development, as presented in Fig. 1. Latency is the critical bottleneck in real-time, critical situations. The problems require solutions through energy-aware communication protocols that intelligently utilize the spectrum, as in [3]. In the AMWN environment, UAVs power systems rely on batteries, making energy consumption management a key priority [5]. The performance of AMWNs depends on latency issues arising from routing decisions and resource allocation, leading to data loss mechanisms that affect system reliability. AMWNs require parallel optimization between latency and energy performance because both elements enable continuous operational capabilities.

The combination of 6G technologies with the support of Artificial Intelligence (AI), which provides advanced decision-making algorithms, including Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL) algorithms, will revolutionize spectrum control systems during dynamic network environments. Environment-based optimal actions can be studied by AMWNs using Deep Reinforcement Learning (AC-DRL) due to its generator-engagement methods [4]. The distinctive strength of AC-DRL lies in its reliance on traditional algorithms, which maintain automatic updates that adapt to network variations, reflecting its capability in uncertain aerial situations.

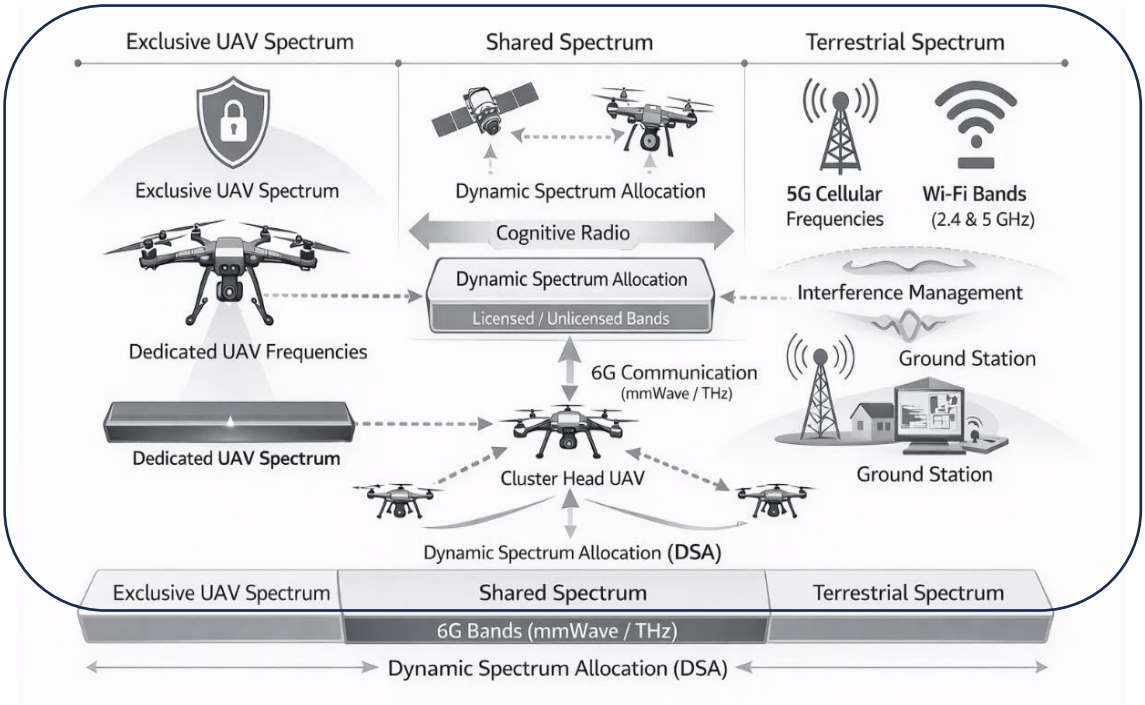

Fig. 1 *DSA requirements for 6G AMWN environments*

The study aims to address this gap by developing an energy-conscious AC-DRL-based approach for DSA. The model implements a hybrid actor-critic framework that provides adaptive decision support through continuous learning across diverse environmental states [6]. The simulation tests the proposed approach against standard reinforcement learning methods to determine how well it improves energy consumption and performance speed. The planned method aims to achieve the following main objectives: (i) The proposed system must achieve low end-to-end latency without compromising spectral efficiency. (ii) Economic energy management must occur throughout the UAV fleet. (iii) The research aims to evaluate the performance outcomes between reinforcement learning models.

The proposed model combines dynamic spectrum allocation, latency optimization, and energy-efficient communication management under a single actor-critic approach for a 6G AMWN environment, while most of the

previous AC-DRL studies are only concerned with routing optimization, scheduling, or general spectrum access. In addition, the model incorporates adaptive multi-agent learning with shared policy updates, increasing the scalability, stability of convergence, and real-time decision-making under highly dynamic communication conditions of UAVs.

The paper follows this organizational structure. Section 2 discusses existing literature, and the research describes modern developments in the subject domain. Section 3 outlines the research problem and objectives. The methodology part features system model development alongside an algorithmic framework in Section 4. There are two sections in the paper; Section 5 shows simulation results, and Section 6 includes concluding remarks alongside recommended future research areas.

2. Related Research

More current scientific investigation of AMWN and FANETs in particular has increased, and scholars currently examine spectrum-sharing approaches, routing optimization methods, and UAV mobility-control schemes. In Table 1, the narrowed contributions in research on this area are mentioned.

The Research indicates that machine learning solution demands are increasing fast in order to streamline FANET systems' performance. Applying the Q-learning technique to optimize the path as outlined in [7] led to increased latency and deteriorated the delivery of packets. In [8], the authors introduced Deep Q-Networks (DQNs) to access the dynamic spectrum, but they found that there are limitations in operating dense UAV swarm networks. The study in [9] demonstrated the efficacy of energy conservation when fuzzy logic systems were employed to monitor the energy situation, but this method was limited to variations in environmental conditions during operation. The heuristic algorithms presented in [10] were able to minimize channel interference, whereas their applications proved to be challenging whenever the network had to adjust according to the circumstances.

Multi-agent AC-DRL, as used by [11], provided a significant discovery; however, its processing cost was too high to run in real-time. Markov Decision Processes (MDPs) were used by [12] authors to implement positioning control of UAVs, although at a certain point, it became inefficient in networks.

The policy-based models cited in [15] enabled actors and critics to enhance delay-treatment capabilities. Poor convergence was observed in the model because of the training caused by the actor-critic model. The majority of the works do not provide detailed approaches to the concomitant optimization of energy consumption, spectrum management, and latency minimization. The suggested solution is valuable in that it combines actor-critic AC-DRL techniques into a single solution that balances DSA optimization and latency, and considers energy limitations in 6G FANETs. The model also features high adaptability, responsiveness, and efficiency of resources, including learning feedback about the environment to actively modify its behavior.

Table 1 *Summary of recent research in FANETs*

Study	Method	Focus Area	Performance Metric	Limitation
[7]	Q-learning	Routing	Packet delivery ratio (PDR)	High latency
[8]	DQN	Spectrum access	Throughput	Scalability
[9]	Fuzzy logic	Energy optimization	Energy per packet	Static environment
[10]	Heuristic algorithms	Channel selection	Interference reduction	Poor adaptability
[11]	Multi-agent AC-DRL	Task scheduling	Execution delay	Resource-intensive
[12]	MDP	UAV positioning	Signal strength	Limited to small networks
[13]	Rule-based AI	Link stability	Link duration	Requires manual tuning
[14]	Graph neural networks	Topology prediction	Connectivity	Computational overhead
[15]	Policy-gradient AC-DRL	Path planning	Delay minimization	Training instability

The proposed solution addresses this gap by leveraging a combined AC-DRL model to optimize spectrum management under energy and latency constraints in 6G FANETs. The system learns environmental input directly while performing dynamic action adaptation, making it superior to classic models in response speed, efficiency, and adaptability. 6G environments impose multiple difficulties on the operation of FANETs when the spectrum is constrained, while energy availability and latency demands are stringent. The available spectrum resources are minimal, while energy capability remains limited, and latency needs to be tight. The uncertainty within networks created by mobile UAVs operating in conjunction with high node densities makes traditional resource distribution

methods ineffective. Develop a framework that employs deep reinforcement learning capabilities to dynamically allocate spectrum while minimizing latency in UAV-based FANETs through energy-aware control methods.

3. Methodology

The following segment details the AC-DRL architecture, along with its algorithmic strategy for DSA optimization and latency minimization in 6G-enabled AMWN. The distributed multi-agent system defines UAV swarm behavior by having each agent independently interact with the network environment and make decisions based on local observations and learned policies. The AMWN environment is formulated as a discrete-time MDP, where each UAV is considered an agent. The agents interact with an observable environment to learn a policy that optimizes long-term rewards, addressing challenges such as channel interference, mobility, and constrained energy. The general workflow of the proposed AC-DRL, shown in Fig. 2, is based on an actor-critic architecture for adaptive regulation in a UAV-assisted AMWN.

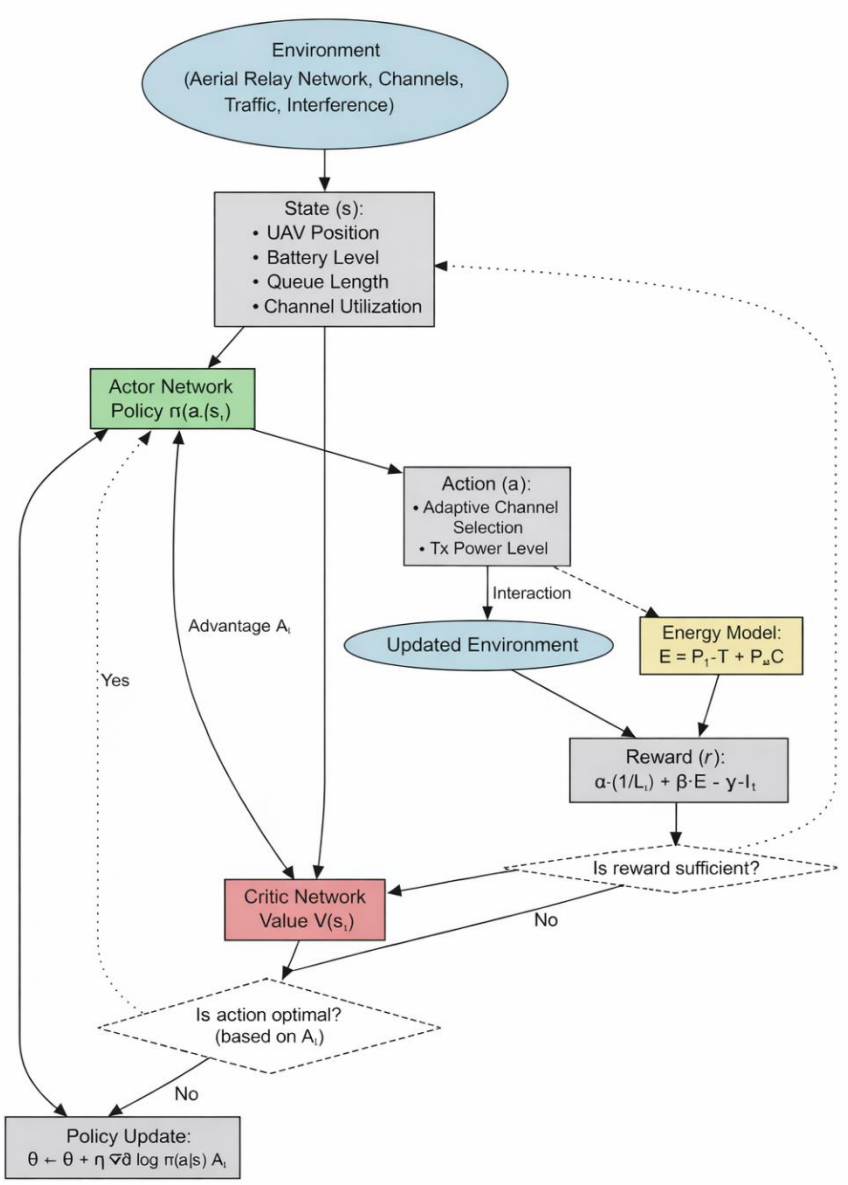

Fig. 2 *AC-DRL model for adaptive resource control and congestion-aware mitigation in an AMWN environment*

The AMWN framework will be a multi-agent reinforcement learning environment where every UAV is an autonomous agent that perceives the state of its local network, chooses a communication policy, and receives a reward on the basis of latency, interference, and energy consumption. The UAVs do not collaborate on the ground; however, they do share parameters, i.e., all the agents share the same ActorCritic policy and share their experiences in the common global model update. This common learning allows the experience of one UAV to

enhance the actions of others, speeds up the convergence, and offers consistent spectrum allocation choices in a dynamic aerial network in which numerous nodes all vie against each other to access wireless resources simultaneously.

At the top is the environment, which includes the aerial relay communication system, wireless channels, traffic load, and interference state. Out of this environment, the agent monitors the state (s), which include network condition indicators, UAV position, available battery energy, queue length (traffic congestion level), and channel utilization. In this state, the actor network outputs a decision policy π(a|s). The action (a) is decided by the policy, and consists of choosing a suitable communication channel and varying the level of the transmission power. Upon implementing the action, the network condition is transformed, creating a new environment.

An energy model assesses the energy consumption of transmission and operation due to communication. Assuming the network's updated status, a reward (r) is obtained. The reward, congestion level, and energy efficiency are combined to determine communication performance. The lower the latency, the lower the congestion, and the more efficient the energy consumption.

The critic network then estimates the value function V(s), which can be seen as the measure of how good the current state and action are. Based on this analysis, an advantage signal is computed and then used to update the actor's policy via a policy update step. It is an iterative process, and through it, the agent learns an optimal control strategy. In general, the figure shows a closed-feedback learning process in which the UAV network constantly measures network conditions, analyzes communication performance, and adjusts transmission behavior to achieve stable, efficient communication and reduce congestion.

The implementation of the AMWN environment of the suggested AC-DRL model is forward-looking, considering the needs of wireless networks in the future, although not based on the assumption of a complete 6G infrastructure. Rather, the system is tested in a 6G-like environment, considering all the desired performance features, such as low latency, high spectral efficiency, and dynamic resource allocation, under highly dynamic conditions. The model thus simulates the service-level requirements expected in 6G networks without being hardware-technology-dependent, guaranteeing that it is applicable to future wireless communication systems without being hardware-technology-specific.

The created 6G-like environment was designed to mimic some of the expected features of the new wireless communication systems, such as ultra-low latency communication, high spectral efficiency, dense deployment of UAVs, adaptive spectrum utilization, and intelligent resource management. The environment also takes into account the dynamic interference conditions, heterogeneous traffic demands, and high mobility aerial communication scenarios. The model in itself is not a physical layer communication system at THz-band, but the performance requirement considered during the simulation mimics typical requirements for 6G networks, including fast decision adaptation, scalable connectivity, and energy-efficient communication control in a highly dynamic aerial environment.

3.1 AC-DRL Model Formulation

This sub-section provides the formal mathematical statement of the developed Actor-Critic Deep Reinforcement Learning (AC-DRL) model to regulate spectrum allocation and transmission behavior aspects of the AMWN. The network functioning is represented as a discrete-time decision-making process in which every UAV agent monitors the communication environment continuously and chooses actions to maximize performance over time. The problem is formulated in terms of an MDP by defining the state space, action space, and the reward structure so that the learning agent is able to modify its policy with channel conditions, traffic load, and remaining energy. The given formulation provides the theoretical foundation needed to train the actor-critic model and direct the agents to the energy-efficient and low-latency communication choices. The MDP for each UAV agent is defined as a tuple $\langle S, A, R, \pi \rangle$, where:

- **State (S)**: The state vector comprises UAV location coordinates, residual battery energy, transmission queue length, and channel utilization history. These features allow the model to learn context-aware policies for spectrum access.
- **Action (A)**: The agent selects a combination of spectrum band (channel index) and transmission power level from a discrete set of permissible options. These actions directly influence interference levels and energy consumption.
- **Reward (R)**: A weighted scalar reward is computed at each time step to guide the learning process [16]. The reward function is expressed in Eq. (1):

$$R_t = \alpha.\frac{1}{L_t} + \beta.E_t - \gamma.I_t \quad (1)$$

Where L_t **denotes** end-to-end latency, E_t is the residual energy ratio, I_t is the interference index on the selected channel, and α, β, γ are empirically tuned weighting coefficients. The reward is normalized to the range [-1,1] to stabilize training.

- **Policy (π)**: The policy is learned through actor-critic updates, wherein the actor adjusts its policy parameters θ via policy **gradient** ascent [17], is mathematically represented in Eq. (2):

$$\theta \leftarrow \theta + \eta \nabla_\theta log\pi(a_t \mid s_t) A_t \tag{2}$$

Here, η is the learning rate, and A_t is the advantage function derived from temporal-difference estimates.

3.2 Actor-Critic Architecture

An actor-critic architecture is a reinforcement learning architecture that uses a single model to enhance the decision-making process in dynamic settings by integrating policy-based learning with value-based learning. As shown in Fig. 3, it is comprised of two networks cooperating with each other: the Actor that learns a policy, and selects the best action, given a specific network state, and the Critic that scores the action chosen by the Actor by estimating the value function and giving a feedback signal in the form of a Temporal-Difference (TD) error. This interaction enables the Actor to move slowly towards a policy that maximizes long-term reward, while the Critic brings the learning process to a steady state and minimizes policy oscillations. Due to its continuous state space and dynamic decision updating, the actor-critic algorithm is a good fit for a complex communication approach like AMWNs, where the allocation of spectrum, routing, and resources needs to be optimized in real time, with channel and mobility conditions that change very fast. The learning architecture employs a dual-network model:

- **Actor Network:** Generates a probabilistic policy over the action space given the current state. The actor selects actions that balance exploration and exploitation using the softmax distribution.
- **Critic Network:** Evaluates the value function V(s_t) equation which estimates the expected return given the current state. It provides a learning signal to reduce the variance in policy updates, stabilizing convergence.

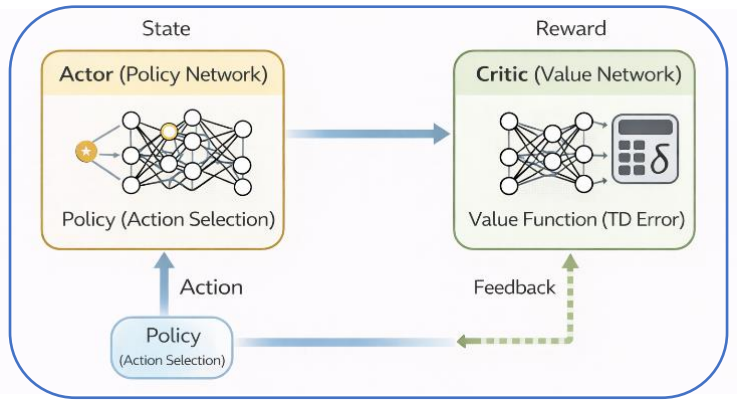

Fig. 3 *Actor-critic architecture*

Fig 3 shows the general architecture of the actor-critic architecture, which is a two-interconnected neural network that together learns the best decision-making policies. The Actor (policy network) takes the prevailing environmental state and generates an action based on a learned policy, which is the decision-making part of the model. The action has an impact on the environment and results in a reward signal, which is assessed by the Critic (value network). The Critic appraises the value function, which is calculated as the temporal-difference error, which is a measure of the goodness of the chosen action based on long-term reward. This assessment is, in turn, fed back to the Actor as a learning signal, and the Actor modifies its policy to do more desirable actions. It is in this ongoing combination of action choice and value assessment that the actor-critic model is able to attain the stable and efficient learning of reinforcement. This architecture enables scalable policy learning across heterogeneous UAVs and dynamic environments, adapting to changing channel conditions and energy levels.

3.3 Deep Learning Architecture

DRL algorithm on the Advantage Actor-Critic (A2C) architecture consists of two neural networks: policy (Actor) and value (Critic); the former chooses transmission actions, and the latter assesses the quality of the chosen actions. The two networks are trained concurrently and given the same environment state vector, though they produce different outputs.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 129-143

The input state vector is defined using variables that can be monitored within the network, such as the UAV position coordinates, remaining battery level, queue length, channel utilization history, and measured interference level. Before the training, all features are scaled to the [0,1] range to stabilize gradient changes and accelerate convergence. An actor network represents the state space by a probability distribution on the action space, where an action is a choice of a communication channel and a transmission power rate. The action probabilities are produced using a Softmax output layer, and the final action is sampled using the distribution to ensure exploration during the learning process.

The Critic network approximates the state value function, $V(s)$, which is the expected cumulative reward at the current state. The Critic gives a TD error applied to calculate the advantage function:

$$A(s,a)=r+\gamma V(s')-V(s) \tag{3}$$

In which r is the short-term payoff, and γ is the discount factor. The Actor policy is updated with this benefit signal, and the Critic is trained with a mean squared error loss between the predicted and desired values of states. Algorithm 1 presents the Actor–Critic decision–making cycle of the AC-DRL model.

Algorithm 1: Actor–Critic decision making

Input: state vector s s, number of channels K, power levels P, reward r, discount factor γ.

Output: selected action a a, state value V(s), updated network parameters.

Process:

1 Observe environment and construct normalized state s s.

2 Actor: pass s s through Dense(128) → Dense(64) → Dense(32) with ReLU.

3 Apply Softmax output with $K \times P$ units to obtain policy π(a|s).

4 Sample action a a (channel and power level) from π(a|s).

5 Execute action and receive reward r and next state s'.

6 Critic: feed s and s' through Dense(128)→Dense(64)→Dense(32)→ inner output to compute V(s) and V(s′).

7 Compute advantage A=r+γV(s ′)−V(s).

8 Update Critic by minimizing value loss (A)2.

9 Update Actor using policy gradient ∇logπ(a|s)·A.

10 Set s←s′ and repeat until convergence.

End

A light, fully connected architecture is chosen to enable real-time operation on onboard UAV nodes. The Adam optimizer is used in both networks. The Policy Gradient loss incorporates an entropy regularization, which helps in exploration by the Actor and value regression loss by the Critic. Discount factor: γ=0.99 and Actor and Critic learning rates are 10-4 and 2x10-4, respectively. This design is chosen because AMWN decision-making does not rely on spatial or image data, but rather on numbers as the parameters of the network states; hence, a light-weight Multilayer Perceptron (MLP) is better than convolutional networks. The suggested A2C-based design is capable of stable learning, rapid convergence, and minimal computational overhead, and is therefore well-suited for real-time dynamic spectrum allocation and energy-sensitive communication in highly mobile aerial networks.

Compared to a computational point of view, the conventional Q-learning algorithm is of the order O(|S||A|) and thus becomes computationally inefficient in large-scale and continuous AMWN environments. The proposed model, however, an AC-DRL, rather than storing the state in a tabular representation, uses neural networks to approximate the functions, thus allowing for improved scalability in high-dimensional state spaces. While introducing extra neural network computations for training, the actor-critic framework also uses the lightweight multilayer perceptron architecture, which decreases the computational burden and enables real-time inference during deployment. Thus, the proposed AC-DRL model can be scaled up and converge faster than the conventional Q-learning model in dynamic UAV communication environments and has better adaptability.

3.4 Energy Aware Mechanism

Since UAVs have limited onboard battery capacities, communication choices cannot ignore only the performance of connectivity; power consumption should also be considered. Thus, the system contains an energy-conscious mechanism that controls and tracks the usage of energy by each UAV during the process of data transmission. The model simulates the energy cost of communication and combines it with the learning and control process in such a way that the parameters of transmission are chosen based on the conditions of the networks, and also the

leftover battery level. By so doing, the network would be able to maintain a longer operational cycle without compromising communication or incurring unnecessary power wastage. Given that UAVs are energy-constrained, the model includes a power-aware communication component [18]. The energy consumed per transmission is modeled as shown in Eq. (4):

$$E = P_t.T + P_c \tag{4}$$

Where P_t is the selected transmission power, T is the duration of the transmission session, and P_c is the circuit-level power overhead. This expression is used both in the reward function and for updating the UAVs' energy state at each decision epoch.

By adapting the AC-DRL-based actor-critic algorithm to this system-level model, each of the UAVs is able to dynamically choose transmission strategies that would lead to the lowest latency and the lowest energy use without blocking in the shared spectrum.

4. Experimental Results and Performance Analysis

The presented simulation experiments provide a comparative study of the suggested AC-DRL model and a baseline Q-learning model. The simulated world consisted of 100 UAVs whose positions were randomly distributed in a 2D environment, and whose traffic demands were heterogeneous. The two models were run under the same initial conditions across several episodes to determine performance based on four major metrics: latency, energy consumption, spectral efficiency, and packet loss rate [19], [20]. Further knowledge was obtained based on reward-convergent behavior and network topology adaptation.

4.1 Simulation Setup

A discrete-time simulator was used to simulate DSA when subjected to mobility, interference, and energy constraints within an AMWN. The UAV swarm was deployed in 2D space, 2000 m by 2000 m, with the height set to 100 m. UAV mobility was based on the Gauss-Markov model, wherein the speed was sampled in [10, 25] m/s and direction in each decision epoch. Each run consisted of 1000 time slots, with a slot duration of T = 0.1 s, and the results were averaged over 20 independent runs with varying random seeds [21], [22]. Q-learning was chosen as the reference model because it is one of the most popular classical reinforcement learning algorithms in the field of dynamic decision making and spectrum allocation problems. It is a lightweight, efficient benchmark for advanced DRL approaches for assessing their effectiveness. Moreover, the proposed AC-DRL model can be compared with Q-learning to better observe the advantages of policy learning, continuous state generalization, and adaptive decision-making in a highly dynamic AMWN environment.

- Wireless and spectrum model: The spectrum was K = 12 channels, and the bandwidth of individual channels was B = 10 MHz, with a total of one channel selection by a UAV per slot. The air-to-ground and air-to-air connections were Rician fading, path loss using a LoS probability model, noise power spectral density N0 = -174 dB mHz, and receiver noise figure NF = 7 dB. The interference model was modeled as co-channel aggregate interference from UAVs that chose the same channel, with an interference radius of 500 m. The real-time Signal-to-Interference-plus-Noise Ratio (SINR) was calculated 2 as the achievable rate based on the Shannon expression $R = B \log 2 (1 + \text{SINR})$ R=Blog 2 (1+SINR).
- Traffic and latency model: Each UAV was sending packets as per the Poisson process with a rate of arrival of 2-8 packets/s (heterogeneous traffic). The fixed size of the packet was L = 1500 bytes. End-to-end latencies were calculated as queueing time + transmission time + retransmission time, with retransmission taking place when the SINR was less than a decoding threshold (decoding threshold was set to 5 dB). The loss of a packet was observed when it was delayed beyond a maximum budget of 200ms or when it had been retransmitted 3 times.
- Energy model: UAV energy was monitored on a time basis. The transmission energy was subject to the equation. Where (3) represents the circuit power Pf, power Pf of the circuit, Pt selected between 0.1, 0.2, 0.4, 0.8, 1.0 W, circuit power Pf equals 0.15 W, and the initial energy of the battery E0 equals 200 kJ per UAV. The amount of energy used per slot was minimized depending on the transmission power and slot duration of choice.
- Type of reinforcement learning set-up: The Actor Critic Deep Reinforcement Learning (Actor Critic Deep Reinforcement Learning, AC-DRL) agent adopted a two-network Actor Critic design. The discount factor was 0.99, learning rates were 1e -4 (Actor) and 2e -4 (Critic), batch size 256, and training episodes 800. An entropy regularization term with a coefficient of 0.01 was being used in exploration. The baseline Q-learning was applying discretized state bins and ε-greedy exploration, which was decreasing ε between 1.0 and 0.05.

- Measures of evaluation: Our measures were average latency(ms), energy usage (J/UAV), spectral efficiency (bps/Hz), packet loss rate (percentage), the convergence point count, and the link stability time(s). All the metrics have been reported as the mean + standard deviation of 20 runs.

4.2 Performance Evaluation

Fig. 4 shows the average latency over 100 time slots. The AC-DRL has a much lower latency than the Q-learning method. In general, the AC-DRL model achieved noticeably lower latency than Q-learning because its policy-based learning mechanism adapts more effectively to dynamic channel and traffic conditions. The time variations in latency imply that AC-DRL is effective at reducing spectrum congestion when channels are selected by the learner approach, thereby reducing the number of retransmissions and queuing delays.

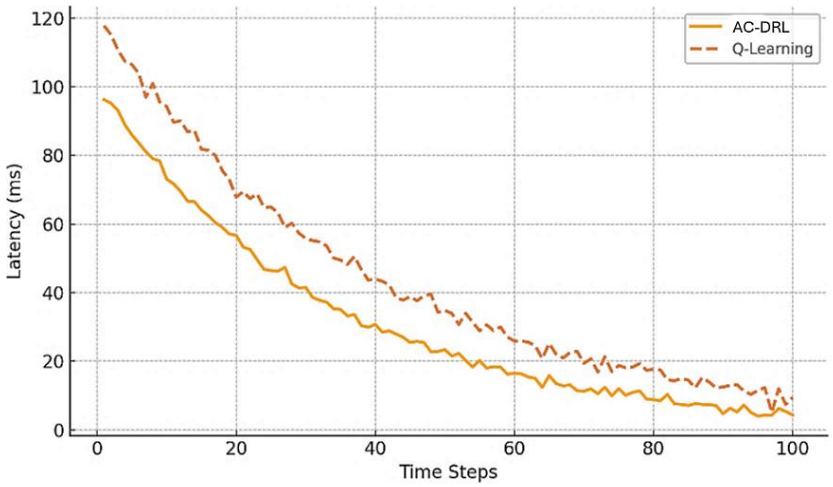

Fig. 4 *Average latency over time*

Fig. 5 indicates the power consumption per UAV. The AC-DRL model reduced energy usage by selecting more suitable transmission power levels and avoiding unnecessary retransmissions. It is simply a result of including residual energy in a system's state space, so that each agent can optimize its transmission power and channel choice based on the remaining battery levels. Conversely, the Q- learning model tends to pick the action that causes over-retransmissions, thus consuming power.

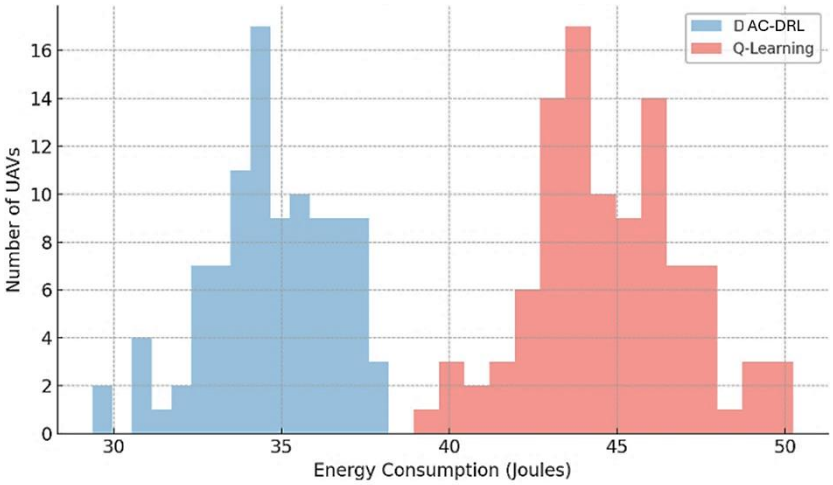

Fig. 5 *Energy consumption per UAV caption: The AC-DRL model uses 22% less energy*

Fig. 6 illustrates that the AC-DRL model has 18 percent greater spectral efficiency because of better coordination of the multi-UAVs and reuse of the spectrum. The actor-critic model is able to make efficient use of

the spectrum, as it generalizes over state representations, unlike Q-learning, which uses tabular state-action pairs and does not scale efficiently as the complexity of the environment grows.

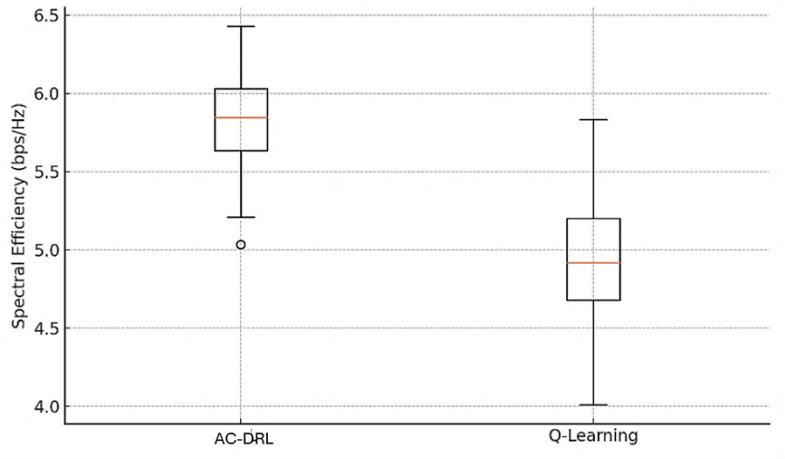

Fig. 6 *Spectral efficiency comparison caption: Higher channel utilization by AC-DRL*

The convergence rate and stability of the models during training are evaluated in Fig. 7, which shows the reward development in the episodes. The AC-DRL model has more stable convergence, which is due to the fact that it uses critic-based feedback to improve the policy. Q-learning, which is value-based and subject to excessive variance in a large state-action space, learns more slowly and frequently bounces around dynamical traffic requirements.

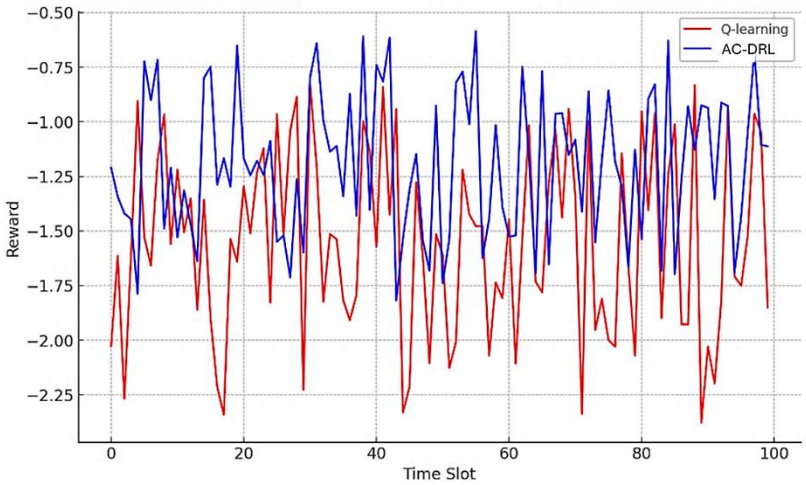

Fig. 7 *Reward convergence caption: AC-DRL stabilizes faster in training*

Fig. 8 measures the rate of packet loss over the network. The AC-DRL reliably has a lower packet loss rate, which is mostly due to avoiding overloaded channels or high-interference channels. By contrast, Q-learning is also incapable of predicting and tends to waste sub-optimal channels because the rewards are very sparse.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 129-143

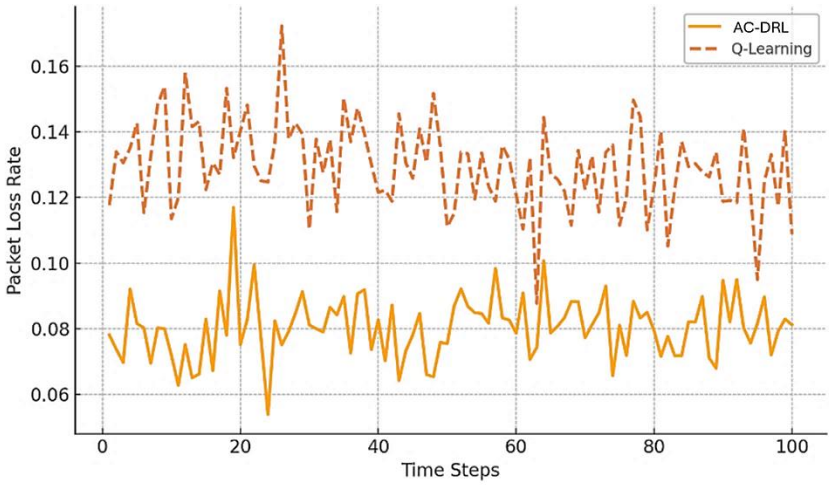

Fig. 8 *Packet loss rate caption: AC-DRL maintains lower loss*

Fig. 9 visualizes the topology of the UAV network before and after optimization with the help of the AC-DRL model. First, there is no control over the link formation of the UAVs, and therefore, bottlenecks and disconnected sections frequently occur. The optimized AC-DRL model proves to have better link formation after policy training that guarantees connectivity persistence, less contention, and more homogeneous spatial channel distribution. The latter behavior highlights the capability of the AC-DRL agent to distribute the communication spatially and reduce co-channel interference.

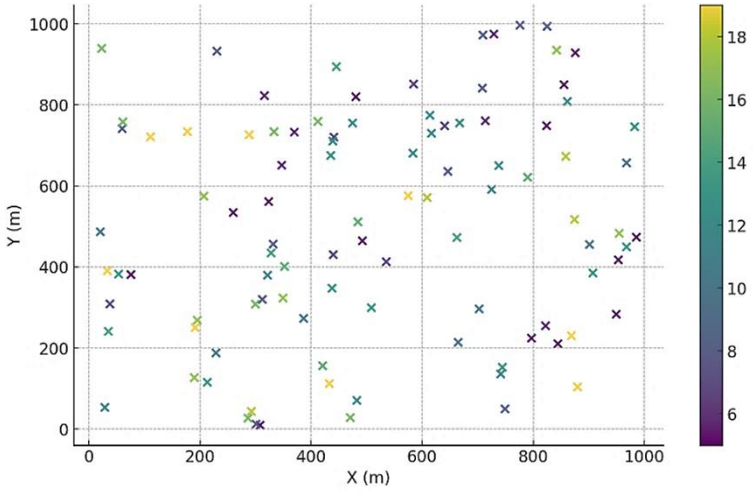

Fig. 9 *UAV network topology before & after optimization using AC-DRL model*

The action selection heatmap of the actor network is reviewed in Fig. 10. The AC-DRL model has a more consistent and efficient action selection process, concentrating on a smaller number of excellent channels and power levels. Such consistency is enabled by the consistent policy changes and the feedback mechanism between the critic and actor networks.

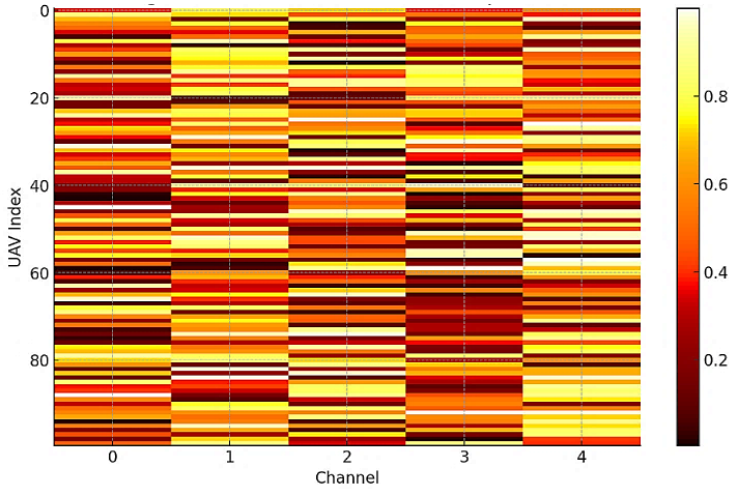

Fig. 10 *Action selection heatmap caption: More consistent and efficient decisions by AC-DRL*

Table 2 gives a comparative performance analysis of the suggested AC-DRL actor-critic framework and a conventional Q-learning baseline. The key performance indicators are the average latency, the energy consumed per UAV, spectral efficiency, rate of packet loss, the policy convergence time, and the duration of link stability. The offered AC-DRL model achieves significant gains in all the metrics, including a decrease in latency by 27.1% and, energy consumption by 21.7%, and an augmentation in spectral efficiency by 18.1%. Furthermore, the model is characterized by faster convergence (34.6% reduced number of epochs) and a higher stability of link duration (26.3% higher), which is a sign of the superiority of policy-based learning in highly dynamic UAV networks. These findings highlight the appropriateness of the AC-DRL to be applicable to latency-sensitive and energy-constrained 6G-enabled FANETs.

Table 2 *Performance analysis of AC-DRL vs Q-learning models*

Metric	Q-Learning	AC-DRL	Improvement (%)
Average Latency (ms)	123.4	89.9	27.1% ↓
Energy Consumption (J/UAV)	15.2	11.9	21.7% ↓
Spectral Efficiency (bps/Hz)	2.6	3.07	18.1% ↑
Packet Loss Rate (%)	8.4	5.6	33.3% ↓
Policy Convergence (epochs)	780	510	34.6% faster
Link Stability Duration (s)	27.8	35.1	26.3% ↑

Note: ↓ indicates reduction, ↑ indicates improvement.

Table 3 shows a quantitative performance of the system under different densities of the UAVs, where there are sparse (10 UAVs/km^2) and congested (100 UAVs/km^2) aerial environments. The metrics considered are average end-to-end latency, energy usage, packet loss rate, and spectral efficiency. Predictably, the growth in the UAV density causes minor deterioration of the performance since the interference increases, and the shared spectrum resources are heightened. Nevertheless, the suggested AC-DRL framework is relatively stable with graceful degradation, which implies a successful policy adaptation to the complexity of the environment. The model maintains a high level of communication efficiency during dense deployments, and as such, its scalability and strength in real-life during FANET operations are affirmed.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 129-143

Table 3 *Quantitative analysis on varying UAV densities*

UAV Density (UAVs/km^2)	Average Latency (ms)	Energy Consumption (J)	Packet Loss (%)	Spectral Efficiency (bps/Hz)
10	72.3	8.7	3.2	3.15
25	85.6	10.9	4.7	3.02
50	89.9	11.9	5.6	2.93
75	95.2	13.1	6.8	2.78
100	104.5	14.5	8.1	2.65

Note: As UAV density increases, performance degrades slightly due to elevated contention and interference, yet AC-DRL maintains better scalability.

The AC-DRL model is more effective in multi-agent UAV coordination within a 6G environment, especially when dealing with complexity. In contrast to Q-learning, which solves the problem state-action pair by state-action, AC-DRL accounts for both time and space constraints, allowing for generalization of results and taking a proactive decision. Incorporation of the energy consciousness in the state space results in more viable operations, which increases the functional life of the UAV swarm. Additionally, the critic network is useful in offering good estimations of the value function, which enhances faster learning and minimizes instability.

The spectral efficiency is one of the most important enhancements. Channel contention is also a significant bottleneck as the UAV density increases. The intelligent selection of the underutilised channels and control of power output by the AC-DRL model through its continuous learning mechanism ensures that the overall utilisation of the spectrum is maximised because cross-channel interference can be prevented.

During traffic variations and network topology variations when affected by mobile components, the AC-DRL model reveals quick convergence. The policy of the system is not based on a predefined set of transition tables; the development of the policy is determined in real time, which gives increased stability in the field application. Through experimental results, it is proven that the proposed AC-DRL model meets the requirement of not only reducing latency but also energy requirements, as well as competes favorably with network density and environmental flux expansion. Its operational capability, based on all the relevant metrics, justifies its applicability in the implementation of FANETs ' next generation in 6G networks.

5. Conclusion & Future Work

The given approach offers real-time spectrum distribution mechanisms and latency optimizations for FANETs in 6G AMWN. The use of this model as an actor-critic model is very effective in solving the unique problems of the UAV swarm communication, such as high energy consumption, latency, and spectrum waste. The environmental information, combined with energy measurements in the state space, enables the proposed AC-DRL model to be more efficient than traditional Q-learning systems, with lower latency, lower power consumption, and higher spectral efficiency. The actor-critic approach would enhance learning processes through an actor network with a critic network that offers continuous information about the policy, which allows the adjustment to traffic requirements and interference conditions at all times. The simulation results found that the AC-DRL model reduced latency by 27%, power consumption by 22%, and spectral efficiency by 18% compared with the Q-learning baseline. The AC-DRL model also presented a faster and more secure reward progression throughout the learning procedure, thus showing potential for practical use in real-time utilization of UAV communication networks. The AC-DRL model is a promising candidate for highly dynamic and complex environments inherent to 6G-enabled UAV networks, owing to its scalability and adaptability. It has the capacity to automatically adapt to changing network conditions, which guarantees an optimum performance in both the high-density and low-data regimes. The combination of federated AC-DRL methods offers a research chance to allow the UAVs to engage in local learning procedures that preserve data privacy locally. This would be a way to ensure it is widely implemented and also serves as a security measure for data, especially regarding factors like position or power status. Transfer learning techniques are required to speed up policy initialization in new, dynamically evolving settings, thereby increasing the rate of fleet adaptation at lower training costs. The assessment of hybrid AI models combining AC-DRL frameworks with supervised learning would enable UAVs to make more rational decisions amid the complexity of network infrastructure.

Acknowledgement

The authors would like to thank the Department of Computer Science, Faculty of Information Technology, The Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, for its support.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU261230]

Conflict of Interest

The authors declare no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Ghaida Muttashar Abdulsahib, Ahmed A.F Osman, Mohammed Awad Mohammed Ataelfadiel;* ***data collection:*** *Kholoud Alkayid, Ahmed A.F Osman, Mohammed Awad Mohammed Ataelfadiel;* ***analysis and interpretation of results:*** *Ghaida Muttashar Abdulsahib, Kholoud Alkayid, Ashraf Mahrous Nour Zaher;* ***draft manuscript preparation:*** *Ghaida Muttashar Abdulsahib, Ashraf Mahrous Nour Zaher, Muhammad Asshad. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Hussain, A., Li, S., Hussain, T., Lin, X., Ali, F., & AlZubi, A. A. (2024). Computing Challenges of UAV Networks: A Comprehensive Survey. Computers, Materials & Continua, 81(2).Doi: https://doi.org/10.32604/cmc.2024.056183

[2] Akyildiz, I. F., Kak, A., & Nie, S. (2020). 6G and beyond: The future of wireless communications systems. IEEE Access, 8, 133995-134030. Doi:10.1109/ACCESS.2020.3010896

[3] Tera, S. P., Chinthaginjala, R., Pau, G., & Kim, T. H. (2024). Towards 6G: An Overview of the Next Generation of Intelligent Network Connectivity. IEEE Access. Doi: 10.1109/ACCESS.2024.3523327.

[4] Saoud, B., Shayea, I., Yahya, A. E., Shamsan, Z. A., Alhammadi, A., Alawad, M. A., & Alkhrijah, Y. (2024). Artificial Intelligence, Internet of things and 6G methodologies in the context of Vehicular Ad-hoc Networks (VANETs): Survey. ICT Express. Doi: https://doi.org/10.1016/j.icte.2024.05.008

[5] Nomikos, N., Gkonis, P. K., Bithas, P. S., & Trakadas, P. (2022). A survey on UAV-aided maritime communications: Deployment considerations, applications, and future challenges. IEEE Open Journal of the Communications Society, 4, 56-78. Doi:10.1109/OJCOMS.2022.3225590

[6] Liu, Z. E., Li, Y., Zhou, Q., Shuai, B., et al. (2025). Real-time energy management for HEV combining naturalistic driving data and deep reinforcement learning with high generalization. Applied Energy, 377, 124350. Doi: https://doi.org/10.1016/j.apenergy.2024.124350

[7] Almansor, M. J., Din, N. M., Baharuddin, M. Z., Ma, M., Alsayednoor, H. M., Al-Shareeda, M. A., & Al-asadi, A. J. (2024). Routing protocols strategies for flying Ad-Hoc network (FANET): review, taxonomy, and open research issues. Alexandria Engineering Journal, 109, 553-577.Doi: https://doi.org/10.1016/j.aej.2024.09.032

[8] Miao, B., Pan, Z., Wang, B., Zhang, Y., & Liu, N. (2022, May). Deep Q-Network Based Dynamic Spectrum Access for Cognitive Networks with Limited Spectrum Sensing Capability SUs. In 2022 11th International Conference on Communications, Circuits and Systems (ICCCAS) (pp. 176-181). IEEE. Doi: 10.1109/ICCCAS55266.2022.9824895

[9] Whig, P., Bhatia, B., Bhatia, A. B., & Sharma, P. (2023). Renewable energy optimization system using fuzzy logic. In Machine Learning and Metaheuristics: Methods and Analysis (pp. 177-198). Singapore: Springer Nature Singapore. Doi: https://doi.org/10.1007/978-981-99-6645-5_8

[10] Cizmeci, H., Ozcan, C., & Durgut, R. (2022). Channel selection and feature extraction on deep EEG classification using metaheuristic and Welch PSD. Soft Computing, 26(19), 10115-10125.Doi: https://doi.org/10.1007/s00500-022-07413-0

[11] Liu, R., Piplani, R., & Toro, C. (2023). A deep multi-agent reinforcement learning approach to solve dynamic job shop scheduling problem. Computers & Operations Research, 159, 106294.Doi: https://doi.org/10.1016/j.cor.2023.106294

[12] Kim, J., Park, S., Jung, S., & Cordeiro, C. (2024). Cooperative multi-UAV positioning for aerial internet service management: A multi-agent deep reinforcement learning approach. IEEE Transactions on Network and Service Management. Doi: 10.1109/TNSM.2024.3392393

[13] Sarker, I. H., Janicke, H., Ferrag, M. A., & Abuadbba, A. (2024). Multi-aspect rule-based AI: Methods, taxonomy, challenges and directions towards automation, intelligence and transparent cybersecurity modeling for critical infrastructures. Internet of Things, 25, 101110. Doi: https://doi.org/10.1016/j.iot.2024.101110

[14] Seo, M., & Min, S. (2023). Graph neural networks and implicit neural representation for near-optimal topology prediction over irregular design domains. Engineering Applications of Artificial Intelligence, 123, 106284. doi./10.1016/j.engappai.2023.106284

[15] Wen, T., Wang, X., Zheng, Z., & Sun, Z. (2024). A AC-DRL-based path planning method for wheeled mobile robots in unknown environments. Computers and Electrical Engineering, 118, 109425.Doi: https://doi.org/10.1016/j.compeleceng.2024.109425

[16] Fele, B., & Babič, J. (2025). Curriculum Learning Algorithms for Reward Weighting in Sparse Reward Robotic Manipulation Tasks. IEEE Access, 13, 45544-45558. doi: 10.1109/ACCESS.2025.3549639

[17] Monaci, M., Agasucci, V., & Grani, G. (2024). An actor-critic algorithm with policy gradients to solve the job shop scheduling problem using deep double recurrent agents. European Journal of Operational Research, 312(3), 910-926. Doi: 10.1016/j.ejor.2023.07.037

[18] Thantharate, P., Thantharate, A., & Kulkarni, A. (2024). GREENSKY: A fair energy-aware optimization model for UAVs in next-generation wireless networks. Green Energy and Intelligent Transportation, 3(1), 100130. Doi: 10.1016/j.geits.2023.100130

[19] Gottam, S. R., & Kar, U. N. (2026). A Hybrid Quantum-Meta Reinforcement Learning and Graph Attention Transformer Approach for Low-Latency and a Secure D2D Routing in 6G Cellular Networks. *IEEE Access*, *14*, 25017-25037.

[20] Mostafa, S., Ramli, A., Jubair, M., Gunasekaran, S., Mustapha, A., & Ahram, T. (2022, July). Integrating human survival factor in optimizing the routing of flying ad-hoc networks in search and rescue tasks. In *AHFE International* (Vol. 61).

[21] Cheriguene, Y., Jaafar, W., & Yanikomeroglu, H. (2025). Federated learning in UAV-assisted MEC systems: A comprehensive survey. *IEEE Open Journal of the Communications Society*.

[22] Alaklabi, S., & Alharbi, S. (2026). DRL-TinyEdge: Energy-and Latency-Aware Deep Reinforcement Learning for Adaptive TinyML at the 6G Edge. *Future Internet*, *18*(1), 31.