network

MDPI

Article

Reinforcement Learning for Resource Allocation in Energy-Harvesting Cooperative IoT Networks

Olumide Alamu *, Thomas O. Olwal and Munguakonkwa Emmanuel Migabo

Department of Electrical Engineering/F'SATI, Tshwane University of Technology, Pretoria 0001, South Africa; olwalto@tut.ac.za (T.O.O.); migabome@tut.ac.za (M.E.M.)

* Correspondence: alamuoa@tut.ac.za

Abstract

The internet of things (IoT) has rapidly evolved into a ubiquitous communication paradigm for enabling the deployment of autonomous wireless networks across diverse application domains. However, the limited energy storage capacity and computational resources of IoT devices (IoTDs) pose a serious concern to their long-term sustainability and the expected quality of service delivery. Moreover, in the foreseeable era of the internet of everything, centralised network resource management is likely to constrain network scalability. To tackle these challenges in the current and next-generation communication networks, the adoption of adaptive and lightweight computational frameworks coupled with energy-efficient transmission strategies is essential. To demonstrate this, we exploit the concept of cooperative communication and radio frequency-based energy-harvesting to improve the network throughput while maintaining power supply to the IoTDs. Furthermore, to intelligently and autonomously perform resource allocation, we employ the reinforcement learning frameworks, particularly state–action–reward–state–action (SARSA) and *Q*-learning. Based on key performance evaluation metrics, we compare our findings with the baseline methods, including the equal, random, and greedy power level selection schemes, with SARSA exhibiting the most favourable performance trade-offs.

Keywords: reinforcement learning; internet of things; energy-harvesting; cooperative communication; throughput

1. Introduction

Recently, the pervasive deployment of internet of things devices (IoTDs) for different applications has revolutionised various fields. These applications include environmental monitoring for weather forecasts in meteorology, health monitoring of buildings and bridges in structural engineering, precision farming and smart irrigation in agriculture, patient health monitoring in medicine, and so on [1,2]. Therefore, the importance of real-time data processing in internet of things (IoT) networks for enabling rapid decision-making cannot be overemphasised.

However, IoTDs are characterised by low computational capability and constrained energy resources. Consequently, they may be unable to efficiently support the local execution of complex signal processing algorithms and advanced security protocols. As a result, the computational tasks are often offloaded to the cloud or edge servers for processing [3,4]. However, this strategy introduces additional delay, bandwidth constraints, and energy consumption, particularly during data re-transmission due to link failure [5,6]. Moreover, IoTDs are often deployed in complex environments or remote areas where signal

Check for updates

Academic Editor: Luis Alonso

Received: 15 April 2026
Revised: 1 June 2026
Accepted: 10 June 2026
Published: 6 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

coverage may be limited. Consequently, the integration of intermediate transmitters in IoT networks, such as relay nodes, has been proposed to enhance link reliability between the serving transmitter and the IoTDs, an approach commonly referred to as cooperative IoT communication [7–9].

To support the scalability of cooperative IoT networks, battery-powered relays are deployed to overcome challenges associated with extending conventional power supply to harsh or remote environments. In line with the sustainable development goal of affordable and clean energy, recent studies have demonstrated the integration of energy-harvesting technologies into cooperative IoT networks. The primary objective is to harness ambient renewable energy sources (e.g., solar and radio frequency (RF)), which are subsequently converted into electrical energy to replenish relay batteries [10,11]. In this regard, energy harvesters, such as photovoltaic cells or rectennas, are integrated into relay receiver hardware to capture energy from renewable sources, thereby mitigating the challenge of manual battery recharging or replacement and enabling autonomous network operation [12–14].

Given the intermittent nature of renewable energy sources and the randomness of fading transmission channels, cooperative energy-harvesting IoT networks can be regarded as causal systems [15]. Consequently, efficient and adaptive resource management is essential for optimizing network performance. Traditional closed-form and iterative optimisation methods often rely on non-causal knowledge of network parameters, which may be impractical in real-world deployments. Furthermore, when complex mathematical models are involved, these solutions may not scale effectively, as they can become computationally intractable at each node in large-scale network scenarios [16].

To overcome this challenge, reward-oriented machine learning algorithms, particularly reinforcement learning, have attracted significant attention for solving various resource allocation problems in stochastic networks without requiring prior knowledge of network parameters [17]. The reinforcement learning (RL) framework enables a decision-making node within the network to learn an optimal policy that maximises an objective function through direct interaction with the network environment using a trial-and-error approach [18]. Furthermore, given the limited energy and computational resources of the IoTDs, RL algorithms are relatively lightweight and can be easily integrated at each node, in contrast to widely adopted deep learning algorithms that rely on computationally intensive frameworks [19]. Therefore, in this paper, our aim is to develop efficient and adaptive resource allocation policies for optimal operation of energy-harvesting cooperative IoT networks.

2. Related Works

Several resource allocation strategies have been proposed in the literature to address the diverse challenges associated with energy-harvesting cooperative IoT networks. These strategies aim to optimize the utilization of limited network resources, such as transmission power, bandwidth, time slots, and harvested energy, while satisfying various performance requirements. In [20], the authors propose subcarrier separation for energy and information transfer to minimise interference at the RF energy-harvesting nodes. The goal is to maximise the IoT system sum-rate, which is formulated as a joint power allocation, node selection, and subcarrier allocation problem. Due to the non-convex nature of the problem, a Lagrangian-based iterative algorithm is developed to obtain a global optimal solution. However, the study did not consider the impact of time or power partitioning ratios, protocols usually employed to separate the received RF signal for information decoding and energy-harvesting purposes. This limitation is addressed in [21]. In the study, the authors consider transmission power allocation, time and power splitting ratios, and relay selection constraints. These constraints are later introduced into the derived closed-form expressions

for evaluating the system outage probability, throughput, and energy efficiency. The results obtained suggest that optimising the allocated time or power ratios could further enhance the system's performance. In contrast to [21], where relay selection is performed based on the link quality between the source and destination nodes, the author of [22] further considers the harvested energy and buffer levels as key constraints that must be considered for an optimal relay selection process. The results obtained indicate that the introduced harvested energy and buffer level constraints provide additional improvements to the outage and throughput performance of the system. Some studies have also considered a closed-form power allocation scheme to maximise the achievable rate in scenarios where energy-harvesting unmanned aerial vehicle (UAV) relays are deployed to serve the ground terminal [23]. However, the closed-form solution approaches employed in the aforementioned studies often involve complex mathematical expansions and exhaustive iterations, which may not scale efficiently in large-scale network deployments.

Given the advantages of reinforcement learning (RL) algorithms over traditional optimisation methods, such as the ones adopted in [20–23], the study in [24] formulated the energy-per-packet minimisation problem as a multi-armed bandit (MAB) problem. The MAB algorithm aims to optimise the chosen spreading factor allocation and coding rate by the IoT node to maximise the reward function. However, compared with *Q*-learning and state–action–reward–state–action (SARSA) learning, the performance of MAB is generally limited due to its inability to model state changes, which is crucial for effective learning. The SARSA framework has also been proposed in [25] to develop a channel selection and BS association policy to maximise the sum-rate in the uplink. Furthermore, *Q*-learning has been widely applied to address various network optimisation problems, including latency reduction and data packet collision mitigation, through optimal channel and routing selection policies [26]. Other application areas include channel and transmission power selection for throughput maximisation [27], link selection for optimal reliable path prediction [28], relay and transmission power selection for energy efficiency maximisation [29], and adaptive sleep–awake scheduling for battery optimisation [30].

The performance of RL and energy-harvesting has also been jointly investigated in [31,32]. In [31], the channel allocation policy that minimises the outage probability of data transfer in the uplink channel is determined, while the paper in [32] proposes subchannel allocation and PS factor selection for EE maximisation using Q-learning. Early convergence of the algorithms is observed, indicating a less complex solution compared to the baseline exhaustive search algorithms. Furthermore, the convergence of cooperative communication with RL and energy-harvesting has been investigated in [33–35]. Papers [33,34] investigate the performance of SARSA for throughput maximisation under MDP and partially observable MDP, respectively. The authors of [35] develop a Q-learning-enabled duty cycle policy based on transmit–sleep scheduling of IoTDs. The goal is to minimise their energy consumption while satisfying the latency and packet delivery ratio requirements of the network.

Beyond the generic model of the solar-harvested energy considered in [33,34], these studies assume the non-existence of a direct path between the serving transmitter and the IoTDs, an assumption that may not hold in practical deployments. Although the solar energy-harvesting model is also adopted in [35], energy-harvesting is inherently limited to daytime, contrary to RF sources, which are readily available as long as a relay can maintain connectivity with a serving transmitter or access point. Moreover, data transfer in the proposed network requires several transmission hops to reach the destination node, an approach that could reduce network throughput, increase the network latency and energy consumption, and exacerbate security vulnerabilities. Motivated by these observations, and to the best of our knowledge, this paper, for the first time, investigates the integration

of RL and RF energy-harvesting in a cooperative IoT network. The core contributions of this paper are outlined as follows.

- We demonstrate the development of a two-hop cooperative IoT network over fading wireless channels in which the source and the destination IoT node can communicate via both a direct path and a non-line-of-sight path through a dedicated relay node. Furthermore, RF-based energy-harvesting is incorporated at the relay to prolong its operational lifetime. The objective is to maximise the overall network throughput.
- Using a Markov decision framework, we formulate the network objective as a power allocation problem aimed at throughput maximisation. To solve this problem, we develop RL-based resource allocation management schemes using Q-learning and SARSA to handle the complexity associated with resource allocation in the proposed IoT network. This is crucial due to random fading of the communication links, which impacts both energy and data arrival dynamics.
- Finally, we evaluate the effectiveness of the proposed RL algorithms using several metrics, including network throughput, energy efficiency, and latency. The simulation results demonstrate the superiority of the proposed schemes over the baseline methods, including equal, random, and greedy power allocation methods.

The rest of the article is organised as follows: The system model is developed in Section 3. Problem formulation is carried out in Section 4. The proposed solution is developed in Sections 5 and 6. Discussion of results and conclusions is provided in Sections 7 and 8, respectively.

3. System Model

The network scenario depicted in Figure 1 illustrates a cooperative IoT network with single antenna–antenna nodes. The main access point (AP) transmits data to the user equipment (UE), $U_k, k \in \{1,2\}$, in the network. Here, U_1 and U_2 are designated as the relay and the destination IoT devices, respectively. Owing to the near–far effect on the transmit signals, U_2 may experience deep fading. Thus, U_1 can be employed to serve as an intermediate transmitter for U_2. The transmission is scheduled in N orthogonal time slots with a total time duration of T. Furthermore, a quasi-static fading channel is assumed in the downlink transmission, where the channel gains between the AP and the UE remain fixed during each transmission block of duration T but vary independently between blocks.

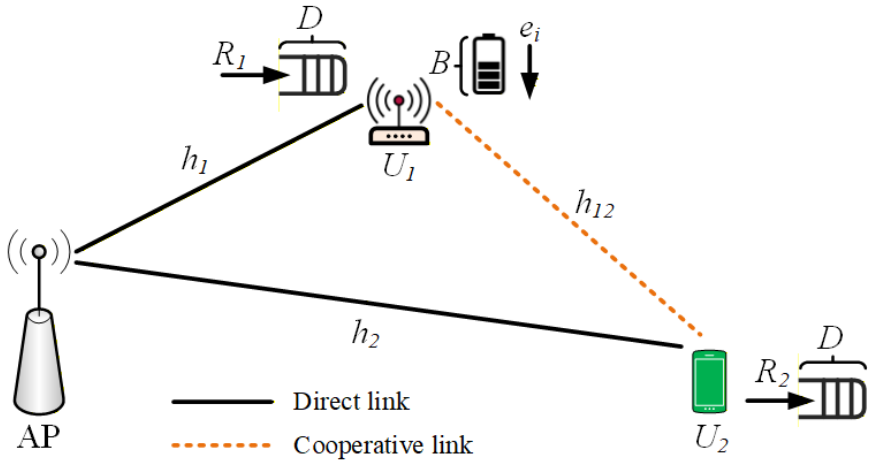


Figure 1. System model.

As depicted in Figure 2, each transmission block is partitioned into three time slots of duration τT, $(1-\tau)T/2$, and $(1-\tau)T/2$, for energy-harvesting at U_1, data transfer from AP to U_1 and U_2, and data transfer from U_1 to U_2, respectively, and τ is the fraction of time

allocated for each slot. Let $p_{0,i}$ denote the transmit power of the AP at the ith fading block; hence, the corresponding harvested energy by U_1 can be written as

$$e_i = \eta\tau T p_{0,i}|h_1|^2, \qquad i \in \{1, 2, \ldots, I\}, \tag{1}$$

where $0 < \eta < 1$ is the energy conversion efficiency and h_1 is the fading channel coefficient between AP and U_1. The harvested energy is stored in the U_1 battery; hence, we denote the battery level over a transmission block T as b_i having a finite capacity of B. Since U_1 also receives the broadcast signal from AP in the first $(1-\tau)T/2$ time slot, it uses its harvested energy to forward the received signal to U_2 in the second $(1-\tau)T/2$ using the decode-and-forward protocol. Thus, the achievable rate at U_k at a particular fading block is given as

$$R_{k,i}^{(1)} = \frac{1-\tau}{2}\log_2\left(1 + \frac{|h_{k,i}|^2 p_{0,i}}{\sigma^2}\right), \qquad k \in \{1, 2\} \tag{2}$$

where σ is the noise variance at $U_{k,i}$. Furthermore, U_1 can only use the harvested energy for transmission after it has been stored in the battery. Hence, the achievable rate in the second transmission time slot can be given as

$$R_{2,i}^{(2)} = \frac{1-\tau}{2}\log_2\left(1 + \frac{|h_{12,i}|^2 p_{1,i}}{\sigma^2}\right), \tag{3}$$

where $p_{1,i}$ is the harvested power allocated for transmitting the data of U_2. The harvested power can be computed from (1) as

$$p_{1,i} = \eta\tau p_{0,i}|h_1|^2 \tag{4}$$

Furthermore, we assume that the direct and cooperative links are aggregated at the U_2 based on a maximum ratio combining (MRC) technique, which can be given as

$$R_2^{MRC} = \frac{1-\tau}{2}\log_2\left(1 + \frac{1}{\sigma^2}\left(|h_{2,i}|^2 p_{0,i} + |h_{12,i}|^2 p_{1,i}\right)\right). \tag{5}$$

Further, to satisfy the energy overflow constraints at the relay, we have

$$b_i - \hat{e}_i + e_i \leq B_{\max}, \tag{6}$$

where b_i and $B_{\max}$ denote the instantaneous battery and maximum battery levels in the ith transmission block, and $\hat{e}_i = e_i\frac{(1-\tau)}{2}$ denotes the energy utilised for transmitting the data of U_2. Moreover, the battery state at the next transmission block can be expressed as

$$b_{i+1} = \min(b_i - \hat{e} + e_i, B_{\max,k}) \tag{7}$$

Similarly, U_1 must satisfy data transmission feasibility and overflow constraints as specified in (8) and (9), respectively.

$$R_{1,i}^{(1)} \leq D_{1,i} \tag{8}$$

$$d_{1,i+1} = \min(D_{1,i} - R_{1,i}, D_{\max,k}) \tag{9}$$

The constraint in (8) ensures that U_1 transmits only its stored data, while (9) discards incoming data once its maximum data buffer capacity is reached. $D_{1,i}$ and $D_{\max}$ denote the data buffer and maximum data buffer levels, respectively.

https://doi.org/10.3390/network6030049

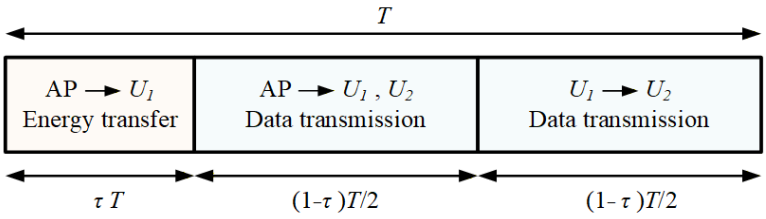


Figure 2. Energy and data transfer protocol.

4. Problem Formulation

Here, we consider a resource allocation approach that ensures successful data delivery to the destination IoT node within a specified time interval of a given transmission block. Since the direct AP-U_2 link may experience deep fading, the likelihood of data outage at the destination IoT node largely depends on the harvested energy and data availability at the relay node. Consequently, given (6)–(8), the problem can be formulated as a power allocation problem aimed at maximising throughput, as expressed in (10a).

$$\mathbf{P1}: \max_{p_{n,i},\, n \in \{0,1\}}, \quad \sum_{i=1}^{I} R_2^{MRC}, \tag{10a}$$

subject to:

$$\sum_{i=1}^{I} \hat{e}_i - \sum_{i=1}^{I} e_i \leq B_{\max,k}, \tag{10b}$$

$$\sum_{i=1}^{I} e_i \leq \sum_{i=1}^{I} \hat{e}_i, \tag{10c}$$

$$\sum_{i=1}^{I} R_{k,i} \leq \sum_{i=1}^{I} D_{max}, \tag{10d}$$

$$p_{n,i} \geq \frac{\sigma^2}{|h_{k,i}|^2} \psi, \tag{10e}$$

$$L'_{k,i} \geq \frac{L_{k,i}}{R_{th}} \tag{10f}$$

where the constraints (10b)–(10d) ensure that the battery capacity, harvested energy, and buffer capacity over the entire transmission block are not violated. The constraint in (10e) is the minimum transmit power required to ensure data throughput in the network, and $\psi = 2^{\frac{2R_{th}}{1-\tau}} - 1$ and R_{th} are the corresponding SNR and target data rate, respectively. The latency constraint is given in (10f), where $L_{k,i}$ denotes the transmitted packet size.

5. Proposed Solution

In problem (10), observe that the continuous-time random processes, including the energy arrival rate at U_1, data arrival rate at both U_1 and U_2, and the channel fading coefficients between the network nodes, are involved. Also, note that the time allocation for the energy and information transfer is defined over a discrete interval. Therefore, the problem can be formulated as a mixed-integer nonlinear optimisation problem, which is challenging to solve.

One could simplify the solution by applying traditional optimisation methods, such as the Lagrangian optimisation method, based on the assumption of global energy constraints. For instance, consider the transmit power constraints in (10e) having a Lagrangian expression of

$$\mathcal{L}_i = \frac{1-\tau}{2}\log_2\left(1+\frac{|h_{2,i}|^2 p_{0,i}+|h_{12,i}|^2 p_{1,i}}{\sigma^2}\right)-\lambda_n\left(p_{n,i}-\frac{\sigma^2}{|h_{k,i}|^2}\psi\right), \quad n\in\{0,1\}, \tag{11}$$

where λ_n denotes the Lagrange multiplier for the transmit power constraint. Using the Karush–Kuhn–Tucker conditions, the derivative with respect to $p_{0,i}$ and $p_{1,i}$ can be given as

$$\frac{\partial \mathcal{L}_i}{\partial p_{0,i}} = \frac{1-\tau}{2\ln 2}\frac{|h_{2,i}|^2}{\sigma^2+|h_{2,i}|^2 p_{0,i}+|h_{12}|^2 p_{1,i}} - \lambda_0 = 0 \tag{12}$$

and

$$\frac{\partial \mathcal{L}_i}{\partial p_{0,i}} = \frac{1-\tau}{2\ln 2}\frac{|h_{12,i}|^2}{\sigma^2+|h_{2,i}|^2 p_{0,i}+|h_{12}|^2 p_{1,i}} - \lambda_1 = 0, \tag{13}$$

respectively. The solutions to (12) and (13) admit a generalised water-filling closed-form solution, which can be written as

$$p^*_{0,i} = \left[\frac{1-\tau}{2\ln 2}\frac{1}{\lambda_0} - \frac{\sigma^2+|h_{12,i}|^2 p^*_{1,i}}{|h_{2,i}|^2}\right]^+, \tag{14}$$

and

$$p^*_{1,i} = \left[\frac{1-\tau}{2\ln 2}\frac{1}{\lambda_1} - \frac{\sigma^2+|h_{2,i}|^2 p^*_{0,i}}{|h_{12,i}|^2}\right]^+, \tag{15}$$

respectively.

However, aside from the oversimplified form of the solutions in (14) and (15), this approach and other similar closed-form optimisation methods would require non-causal knowledge of the aforementioned random processes, which is unrealistic in practice. Hence, by leveraging both present and past information on the decisions taken at the AP and the relay, including the selection of transmit power levels and operation of the relay in cooperative mode, the throughput of the IoT node can maximised, thereby minimizing transmission delay. Since these decisions are based on energy availability, channel conditions, and the levels of the energy and data buffers, the optimisation problem can be modelled as a Markov decision process (MDP) and RL can be employed to obtain an optimal solution.

MDP Preliminaries

An MDP provides the framework that enables decision-makers or agents (i.e., the AP and relay) to learn via direct interaction with the network environment. It can be defined by a tuple $(S,\ A,\ P,\ R)$, where S represents a finite set of states, A is the finite set of actions chosen by an agent, and $P : S \times S \times A \to [1,0]$ is the transition probability to a new state $s' \in S$ after an action $a \in A$ is taken in state $s \in S$. The notation $r : S \times A \to \mathbb{R}$ is the expected reward for executing the selected action. In the following, we map the described elements to the formulated optimisation problem.

(1) *States:* These consist of the battery and harvested energy levels of the relay, that is, $b_{1,i}$ and $e_{1,i}$, respectively, the buffer levels of the relay and the IoT devices $d_{k,i}$, and the channel coefficients between the nodes h_k and h_{12}. Hence, the state space can be expressed as $s_i \in \{b_{1,i}, e_{1,i}, d_{k,i}, h_k, h_{12}\}$.

(2) *Actions:* The set of actions, A, includes the transmit power values of the AP and the relay that can be selected to maximise the throughput of the destination IoT device. This can be expressed as $p_{n,i} \in A_n = \{0, \xi_n, 2\xi_n, \ldots, P_{n,\max}\}\ n \in \{0,1\}$, where ξ

denotes the step size. To relax the constraints in (10), we reconstruct the action space to enforce (10b) and (10e) such that

$$A_n = \begin{cases} p_{n,i} & p_{n,i} \geq \frac{\sigma^2}{|h_{k,i}|^2}\psi, e_{1,i} \leq b_{1,i} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

(3) *Reward:* This is related to the impact of the selected transmit power on the expected throughput of the IoT device, which can be formulated as

$$R_i = \mathbb{E}\left[\sum_{i=1}^{I} \gamma^i R_2^{MRC}\right], \tag{17}$$

where $0 \leq \gamma \leq 1$ is the discount factor.

Accordingly, we can reformulate (10) as

$$\begin{aligned} \mathbf{P2}: \max_{p_{n,i}\ n \in \{0,1\}} \quad & \mathbb{E}\left[\sum_{i=1}^{I} \gamma^i R_2^{MRC}\right] \\ & \text{subject to: (10b)} - \text{(10f)} \end{aligned} \tag{18}$$

6. RL Framework

Here, we propose the RL algorithms to solve the MDP problem in (18). Due to the dynamic and unpredictable nature of the state space elements coupled with the requirement of their non-causal knowledge in constraints (10b)–(10f) to solve the problem, it is appropriate to employ the model-free RL algorithm. Hence, we investigate the performance of SARSA and Q-learning algorithms to find the power allocation policies that maximise the system throughput.

The SARSA algorithm is based on learning the action-value function $Q(s, a)$. During the ith interval, the AP and U_1 are in state $s_{n,i}$ and they select from the set of A_n, comprising different transmit power levels. After the transmission is performed, they transition into a new state $s_{n,i+1}$, having an immediate reward of $R_{n,i+1}$. The expression for the update rule of the learned Q-value in the Q-table can be performed recursively as

$$\begin{aligned} Q(s_{n,i}, p_{n,i}) \leftarrow & Q(s_{n,i}, p_{n,i}) + \alpha\Big[R_{n,i+1} + \gamma Q(s_{n,i+1}, p_{n,i+1}) \\ & - Q(s_{n,i}, p_{n,i})\Big], \end{aligned} \tag{19}$$

where $\alpha \in [0, 1]$ is the learning rate. Furthermore, to ensure convergence, it is necessary to balance the trade-off between the exploration and exploitation of the agents' environment. Thus, we employ an ε-greedy strategy for the action selection at any state $s_{n,i}$, which can be expressed as follows:

$$p_{n,i} = \begin{cases} \underset{p_{n,i}}{\text{argmax}}\, Q(s_{n,i}, p_{n,i}) & \text{with probability } 1 - \varepsilon, \\ \text{random } p_{n,i} & \text{with probability } \varepsilon. \end{cases} \tag{20}$$

Note that the SARSA learning in (19) is regarded as an on-policy RL technique since the Q-values are updated by choosing the power values in the next state according to the ε-greedy policy [36]. On the other hand, the Q-learning technique updates its Q-values towards the greedy target policies by selecting the power value $p'_{n,i}$ that maximises the

Q-values in the next state [37,38]. Therefore, Q-learning is regarded as an off-policy RL technique, which can be modified as

$$Q(s_{n,i}, p_{n,i}) \leftarrow Q(s_{n,i}, p_{n,i}) + \alpha \Big[R_{n,i+1} + \gamma \max_{p'_{i,n}} Q(s_{n,i+1}, p'_{n,i}) - Q(s_{n,i}, p'_{n,i}) \Big]. \tag{21}$$

As stated in expressions (19) and (21), the RL-based algorithms for the throughput maximisation problem in (18) are summarised in Algorithms 1 and 2, respectively.

Algorithm 1 SARSA for Throughput Maximisation

Input parameters: α, γ, ε, ξ
Output: Optimised Q-values

Initialise $Q(s_{n,i}, p_{n,i}) = 0$ for all $(s_{n,i}, p_{n,i})$ pairs
for each episode **do**
  initialise $s_{n,i}$
  choose $p_{n,i}$ according to (20)
  for $i \in \{1, \dots I\}$ **do**
    use selected $p_{n,i}$ to transmit data
    get the reward $R_{n,i+1}$
    observe the next state $s_{n,i+1}$
    select the next action according to (20)
    get $Q(s_{n,i+1}, p_{n,i+1})$ from the Q-table
    update $Q(s_{n,i}, p_{n,i})$ by (19)
    set $s_{n,i} = s_{n,i+1}$; $p_{n,i} = p_{n,i+1}$
  end for
end for

Algorithm 2 Q-learning for Throughput Maximisation

Input parameters: α, γ, ε, ξ
Output: Optimised Q-values

Initialise $Q(s_{n,i}, p_{n,i}) = 0$ for all $(s_{n,i}, p_{n,i})$ pairs
for each episode **do**
  initialise $s_{n,i}$
  choose $p_{n,i}$ according to (20)
  for $i \in \{1, \dots I\}$ **do**
    using selected $p_{n,i}$ to transmit data
    get the reward $R_{n,i+1}$
    observe the next state $s_{n,i+1}$
    select the next action according to (20)
    get $\max_{p'_{i,n}} Q(s_{n,i+1}, p'_{n,i})$ from the Q-table
    update $Q(s_{n,i}, p_{n,i})$ by (21)
    set $s_{n,i} = s_{n,i+1}$
  end for
end for

Complexity

Regarding the computational complexities of the proposed algorithms, the analysis is performed in terms of the number of users in the network N, the number of episodes E, the number of iterations per episode I, and the action space A. In the case of SARSA, for each iteration with an episode, the action selection based on the ε-policy has the complexity of $\mathcal{O}(A)$. For Q-table lookup, reward computation and Q-value update, each process has a computation of $\mathcal{O}(1)$. Therefore, the complexity per iteration per user is $\mathcal{O}(A)$.

Consequently, the complexity per episode can be computed as $\mathcal{O}(NAIE)$. In the case of Q-learning, the ε-action selection strategy requires a computation of $O(A)$ and the maximum Q-value computation $\max_{p'_{i,n}} Q(s_{n,i+1}, p'_{n,i}) \rightarrow O(A)$. Thus, Q-learning exhibits the same asymptotic computational complexity as SARSA.

7. Results and Discussion

The network and simulation parameters used in this study are provided in Table 1.

Table 1. System simulation parameters.

Parameters	Values
Maximum AP transmit power (p_0)	3 W
Circuit power consumption of (p_0^c)	1 W
Maximum Relay transmit power (p_1^c)	1.5 W
Circuit power consumption of (p_1^c)	0.5 W
Energy conversion efficiency (η)	0.8
Packet size $(L_{k,i})$	30 B
Noise power (σ^2)	-30 dBm
Number of episodes	3000
Discount factor (γ)	0.9
Exploration probability (ε)	0.1
Learning rate (α)	0.2

The scripting and execution of the algorithms are performed and executed in MATLAB R2024b on Windows 11 with a 2.40 GHz i5-1135G7 processor and 16 GB RAM. The channel gains are estimated following the Rayliegh channel model $h_{k\in\{1,2\}} = d_n^{-\lambda}$ and $h_{12} = d_3^{-\lambda}$, where d_k and d_{12} represent the distance from the AP to U_1 and U_2, and the distance from U_1 to U_2, respectively. λ is the path loss exponent. We compare the results obtained with baseline scenarios where the chosen transmit power in each time slot is (i) equal, (ii) randomly selected, and (iii) the case where the maximum feasible transmit power is always selected, referred to as the greedy method.

In addition to evaluating the network throughput performance, we further analyze the energy efficiency and system latency, which are defined as follows.

$$\text{Energy Efficiency} = \frac{\frac{1-\tau}{2}\sum_{i=0}^{I}\log_2\left(1+\frac{1}{\sigma^2}\left(|h_{k,i}|^2 p_{0,i}+|h_{2,i}|^2 p_{0,i}+|h_{12,i}|^2 p_{1,i}\right)\right)}{p_0+p_n^c-p_1} \tag{22}$$

where P_0 , p_n^c and p_i denote the AP transmission power, circuit power consumption of the AP and the relay, and the harvest power at the relay.

$$\text{Latency} = \frac{L_{k,i}}{\max\left\{\frac{1-\tau}{2}\sum_{i=0}^{I}\log_2\left(1+\frac{|h_{k,i}|^2 p_{0,i}}{\sigma^2}\right), \frac{1-\tau}{2}\sum_{i=0}^{I}\log_2\left(1+\frac{1}{\sigma^2}\left(|h_{2,i}|^2 p_{0,i}+|h_{12,i}|^2 p_{1,i}\right)\right)\right\}}, \tag{23}$$

where $L_{k,i}$ denotes the transmitted packet size.

7.1. Throughput of Destination IoTD

Figure 3 depicts the throughput performance of the destination IoTD, U_2. We observe a significant transition to the convergence region at approximately 600 episodes. Furthermore, prior to the convergence phase, the proposed Q-learning and SARSA operate predominantly in the exploration–exploitation phase, indicating incomplete convergence of the policy. Beyond this phase, *Q*-learning achieves the highest instantaneous throughput performance, followed by SARSA. This gain can be attributed to its off-policy approach to power level selection, where the decision-making nodes greedily choose a higher transmit power level at the successive time slot.

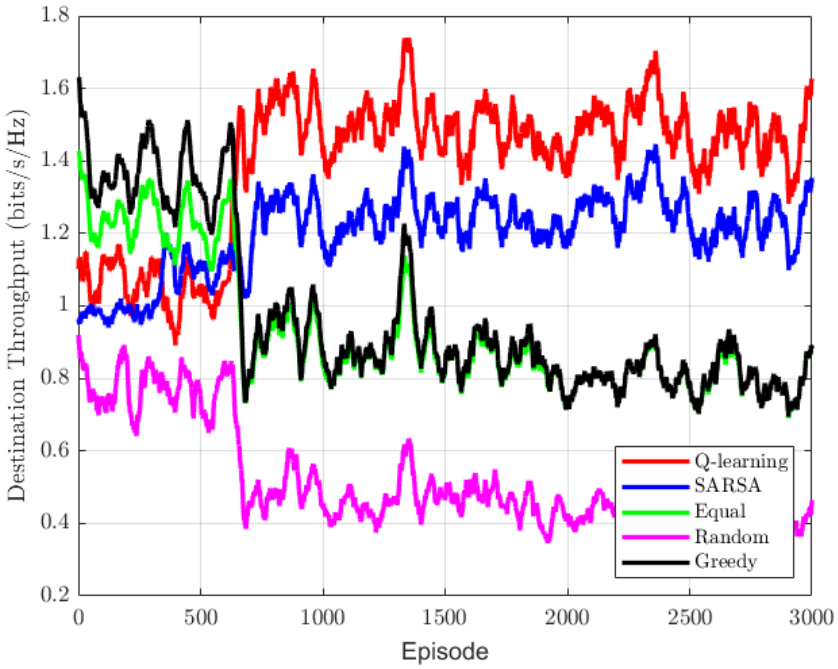


Figure 3. Convergence of throughput of destination IoTD $R_{th} = 0.5$ bps/Hz.

In contrast, although SARSA attains a lower throughput than Q-learning, its on-policy learning mechanism promotes more stable power allocation decisions. As a result, it can reduce large fluctuations in battery levels and energy consumption, thereby supporting more sustainable network operation. Among the baseline methods, the greedy algorithm initially demonstrates the best performance, followed by the equal transmit power selection. However, rapid degradation is later observed in both schemes, which can be attributed to the fast depletion of the U_1 battery, potentially limiting effective cooperation in the network. The degradation observed under equal power selection can be linked to the inability of the AP and U_1 to adapt to the energy, channel, and data causality. This lack of adaptability also explains the worst-case performance exhibited by the random power selection scheme.

7.2. Throughput of Destination IoTD Under Higher Rate Threshold

In Figure 4, we examine the throughput performance of U_2 when R_{th} is increased to 1 bps/Hz, in contrast to the scenario depicted in Figure 3, where $R_{th} = 0.5$ bps/Hz. For both SARSA and *Q*-learning, convergence is observed after approximately 300 episodes. This behaviour is due to the fact that as the rate threshold increases, the reward function in the algorithms will assign a lesser reward to power selection that yields insufficient throughput, thereby leading to a faster transition to the convergence region. Moreover, as the R_{th} increases, there is a high likelihood of the direct link not satisfying this value due to deep fading. This may lead to higher exploitation of the cooperative links, i.e., h_1

and h_{12}. Therefore, as the R_{th} increases asymptotically, we can infer that the throughput of the system will be dominated by the cooperative link. However, to satisfy a higher R_{th}, an increase in the transmit power of both the AP and U_1 may be required, which can adversely impact the system's energy efficiency.

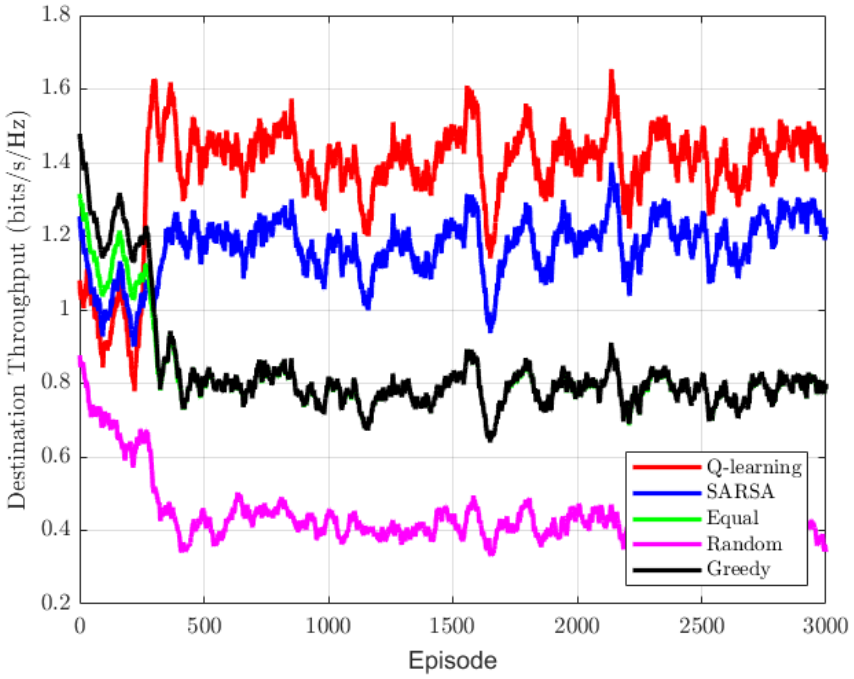


Figure 4. Convergence of throughput of destination IoTD $R_{th} = 1$ bps/Hz.

7.3. Sum-Throughput of the System

The outcome of the aggregation of the network sum-throughput based on the expressions in (2) and (5) is depicted in Figure 5. As expected, SARSA and *Q*-learning outperform the baseline methods. Moreover, an early convergence of both algorithms is observed, indicating fair resource allocation in the proposed solution. In the case of equal and greedy power selection schemes, though their performances remain below *Q*-learning and SARSA, a negligible difference is observed in both schemes. This is primarily due to the choice of power allocation scheme enforced in the simulation, which is $0.8P_{n\in\{0,1\}}$, while the greedy algorithms always aim for the selection of the highest feasible power level. As observed in the previous results, the random power selection scheme yields the poorest performance among all considered methods. Since transmit power levels are selected without considering the network state information, the scheme frequently allocates resources inefficiently, resulting in reduced cooperation efficiency and a significantly lower network sum-throughput.

7.4. Energy Efficiency

Figure 6 illustrates the energy efficiency of the developed system, which is computed in (22) as the ratio of the sum-throughput to total power consumption in the network. Both Q-learning and SARSA exhibit fluctuations in the first 500 episodes, which can be attributed to the exploration of different transmission power levels. Beyond this phase, both algorithms progressively converge, with *Q*-learning showing marginal improvement over SARSA. Although SARSA is expected to outperform Q-learning due to its more conservative selection of power levels, which leads to a relatively lower throughput, this constitutes the primary factor limiting its overall energy efficiency. Furthermore, the greedy scheme exhibits the worst performance as it consistently selects high transmit

power at every successive time slot. This observation validates the fact that higher power consumption is not desirable for energy-efficient system design.

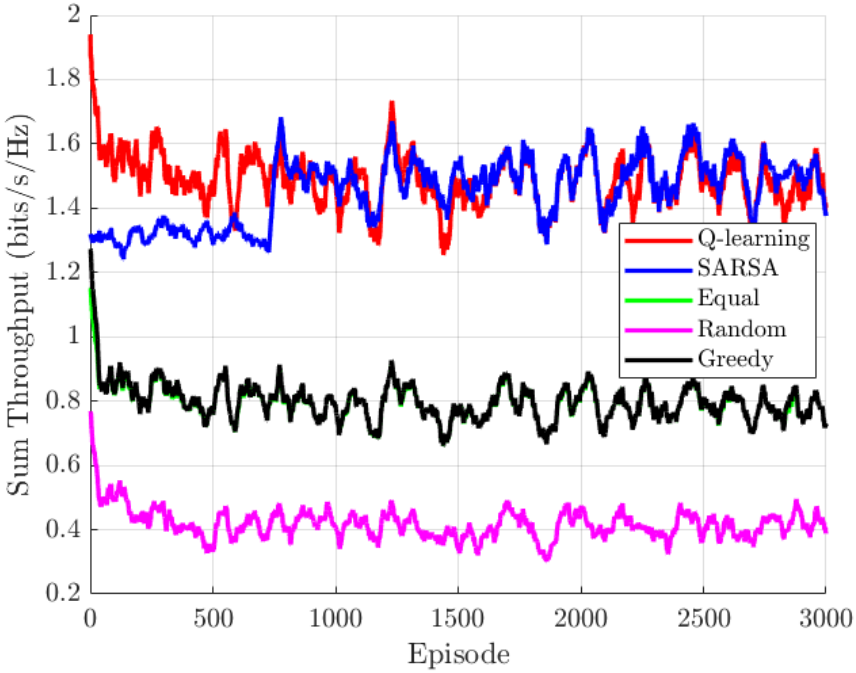


Figure 5. Convergence of system sum-throughput.

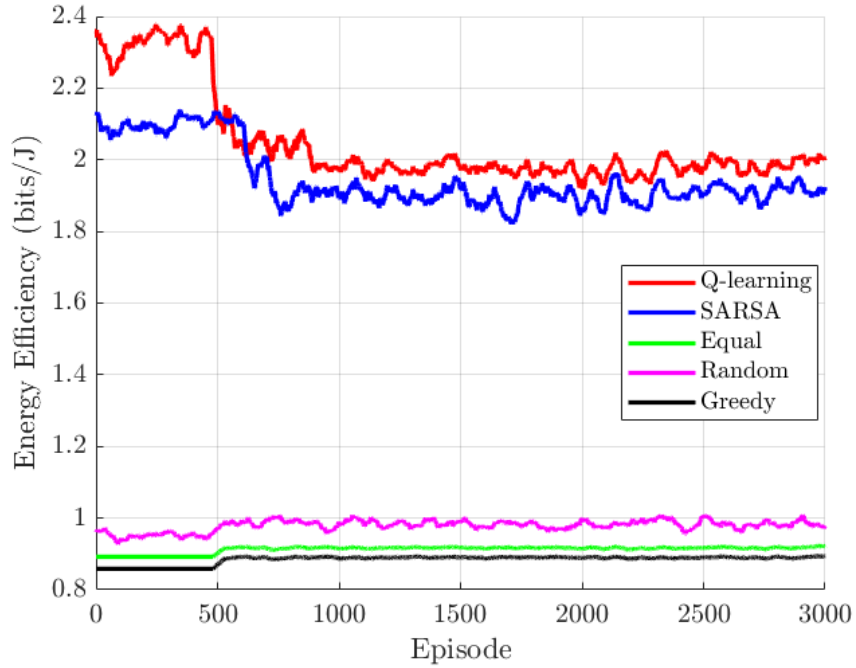


Figure 6. System energy efficiency.

7.5. System Latency

In Figure 7, we examine the latency performance of the system. We observe a slight improvement in the latency of *Q*-learning compared with SARSA, where the convergence is observed at mean values of 4 s and 5 s, respectively. Q-learning rapidly converges to a policy that maximises the system throughput due to its aggressive reward structure, which encourages the selection of transmit power levels that maximise the system throughput. Also, we observe that *Q*-learning and SARSA exhibit marginal performance with the baseline schemes, except for the random scheme with the poorest performance. This observation

suggests that while the baseline schemes can achieve acceptable latency performance under certain conditions, they are unable to consistently adapt their transmission decisions to the dynamic network environment. In contrast, the learning-based approaches continuously adjust their actions according to variations in channel quality, battery state, harvested energy, and queue conditions, enabling more efficient utilisation of network resources.

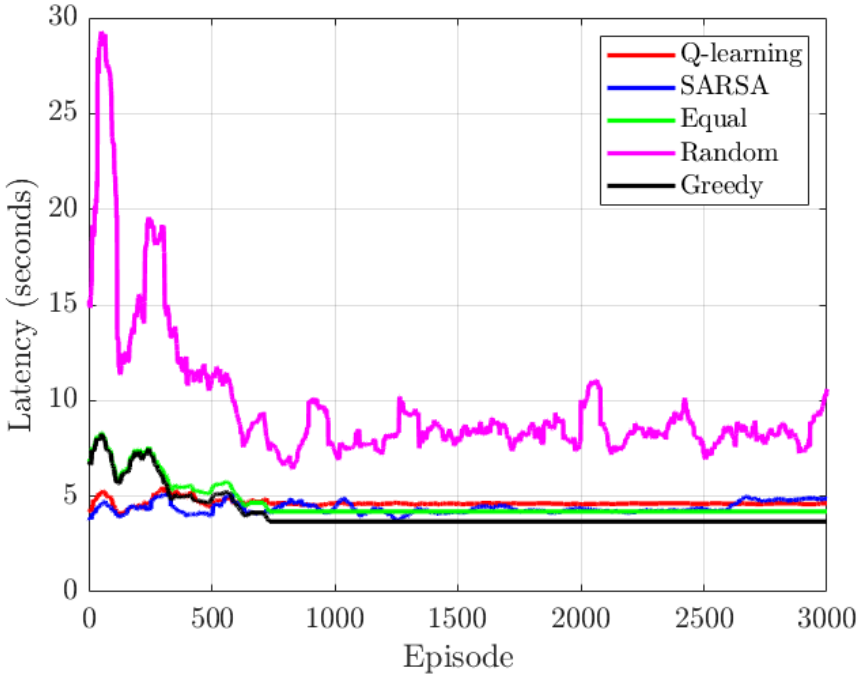


Figure 7. System latency.

8. Conclusions

In this paper, the development of an RL framework for resource allocation in a cooperative IoT network is demonstrated. Moreover, an RF-based energy-harvesting technique is employed to replenish the battery of the relay node. The resource allocation problem is formulated such that the throughput of the destination IoTD is maximised. Owing to the unavailability of non-causal or statistical information regarding harvested energy, channel conditions, and data arrival rates, we adopt model-free RL approaches, namely SARSA and Q-learning. Based on the simulation results, SARSA shows the most favourable performance trade-offs in comparison with Q-learning. In addition, our findings outperform the baseline methods such as the equal, random, and greedy transmission power level schemes, with the random power level selection approach exhibiting the worst performance. In our future work, we intend to incorporate an edge computing framework within large-scale IoT networks to enhance the computational performance of the developed system. Moreover, we shall also introduce the concept of post-decision state to accelerate the convergence rate of the proposed SARSA and Q-learning.

Author Contributions: Conceptualization; O.A., T.O.O. and M.E.M.; methodology, O.A.; validation, T.O.O., M.E.M. and O.A.; formal analysis. O.A.; writing—original draft preparation, O.A.; writing—review and editing, T.O.O. and M.E.M.; resources, T.O.O. and M.E.M.; supervision, T.O.O. and M.E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Tshwane University of Technology, South Africa.

Data Availability Statement: The dataset used is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ullah, I.; Bilal, H.; Sharafian, A.; Betalo, M.L.; Samy, S.A.; Bai, X. The Internet of Nature Things (IoNT): Pioneering a New Frontier in Environmental Monitoring and Sustainable Ecosystem Management. *IEEE Internet Things J.* **2026**, *13*, 4018–4045. [CrossRef]
2. Sharma, B.P.; Peelam, M.S.; Gupta, A.; Shekhar, C.; Chamola, V. A Comprehensive Survey on Data Converters for IoT Applications: Scope, Issues, and Future Directions. *IEEE Internet Things J.* **2025**, *12*, 18993–19017. [CrossRef]
3. Ribeiro, A.R.; Andrade, A.C.S.; dos Santos, G.L.; Marcondes Lopes, G.D.; Oliveira, E.M.d.; Souza, A.D.d.; Machado, J.B. Green Scheduling and Task Offloading in Edge Computing: A Systematic Review. *Network* **2026**, *6*, 17. [CrossRef]
4. Kumar, A.; Pal, A. A Survey on Computation Offloading and Current Trends. *IEEE Access* **2025**, *13*, 185318–185356. [CrossRef]
5. Schulz, J.; Seeling, P.; Reisslein, M.; Fitzek, F.H.P. No Further Delay: Making Time an Ally of Edge Computation of AI Workloads. *IEEE Internet Things Mag.* **2026**, *9*, 33–40. [CrossRef]
6. Liu, J.; Du, Y.; Yang, K.; Wu, J.; Wang, Y.; Hu, X.; Wang, Z.; Liu, Y.; Sun, P.; Boukerche, A.; et al. Edge-Cloud Collaborative Computing on Distributed Intelligence and Model Optimization: A Survey. *IEEE Commun. Surv. Tutor.* **2026**, *28*, 5049–5080. [CrossRef]
7. Cheikh, I.; Roy, S.; Sabir, E.; Aouami, R. Energy, Scalability, Data, and Security in Massive IoT: Current Landscape and Future Directions. *IEEE Internet Things J.* **2026**, *13*, 10127–10163. [CrossRef]
8. Zhang, Z.; Atapattu, S.; Ren, B.; Wang, Y.; Di Renzo, M. Optimizing Energy-Efficient Cooperative MAC Strategies for Data Collection in IoT Networks with Terrestrial and Nonterrestrial Relays. *IEEE Internet Things J.* **2025**, *12*, 35556–35576. [CrossRef]
9. Bagheri, A. Toward Next-Generation Connectivity: Analyzing Cell Edge User Connectivity in Cooperative NOMA-Based Cellular Networks Over Correlated Rayleigh Channels for B5G and 6G. *IEEE Trans. Commun.* **2025**, *73*, 12533–12544. [CrossRef]
10. Hammoud, K.; Landivar, J.F.; Häkkinen, J.; Volskiy, V.; Schreurs, D.; Pollin, S.; Visser, H.J. Efficient Light Energy Harvesting and RF Distribution for Dense IoT Networks. *IEEE Internet Things J.* **2026**, *13*, 15005–15019. [CrossRef]
11. Qasim Alazzawi, M.; Sánchez-Aarnoutse, J.C.; Martínez-Sala, A.S.; Cano, M.D. Green IoT: Energy Efficiency, Renewable Integration, and Security Implications. *IET Netw.* **2025**, *14*, e70003. [CrossRef]
12. Safaei, B.; Peiravian, M.; Siamaki, M. Eco-Friendly IoT: Leveraging Energy Harvesting for a Sustainable Future. *IEEE Sens. Rev.* **2025**, *2*, 32–75. [CrossRef]
13. Balouch, Z.A.; Sampe, J.; Misran, N.; Abdullah, H.; Rhazali, Z.A.; Yunus, N.H.M. Rectennas in RF energy harvesting: A comparative analysis on design, performance and challenges for low power devices. *Sustain. Energy Technol. Assess.* **2026**, *87*, 104918. [CrossRef]
14. Li, X.; Jiang, J.; Wang, H.; Han, C.; Chen, G.; Du, J.; Hu, C.; Mumtaz, S. Physical Layer Security for Wireless-Powered Ambient Backscatter Cooperative Communication Networks. *IEEE Trans. Cogn. Commun. Netw.* **2023**, *9*, 927–939. [CrossRef]
15. Alamu, O.; Olwal, T.O.; Migabo, E.M. Machine learning applications in energy harvesting internet of things networks: A review. *IEEE Access* **2025**, *13*, 4235–4266. [CrossRef]
16. Dulaj, K.; Alhammadi, A.; Shayea, I.; El-Saleh, A.A.; Alnakhli, M. Harnessing Machine Learning for Intelligent Networking in 5G Technology and Beyond: Advancements, Applications and Challenges. *IEEE Open J. Intell. Transp. Syst.* **2025**, *6*, 605–633. [CrossRef]
17. Hady, M.A.; Hu, S.; Pratama, M.; Cao, Z.; Kowalczyk, R. Multi-agent reinforcement learning for resources allocation optimization: A survey. *Artif. Intell. Rev.* **2025**, *58*, 354. [CrossRef]
18. Frikha, M.S.; Gammar, S.M.; Lahmadi, A.; Andrey, L. Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey. *Comput. Commun.* **2021**, *178*, 98–113. [CrossRef]
19. Xi, M.; Dai, H.; He, J.; Li, W.; Wen, J.; Xiao, S.; Yang, J. A lightweight reinforcement-learning-based real-time path-planning method for unmanned aerial vehicles. *IEEE Internet Things J.* **2024**, *11*, 21061–21071. [CrossRef]
20. An, H.; Park, H. Energy-Balancing Resource Allocation for Wireless Cooperative IoT Networks with SWIPT. *IEEE Internet Things J.* **2022**, *9*, 12258–12271. [CrossRef]
21. Vu, T.H.; Nguyen, T.V.; Kim, S. Cooperative NOMA-Enabled SWIPT IoT Networks with Imperfect SIC: Performance Analysis and Deep Learning Evaluation. *IEEE Internet Things J.* **2022**, *9*, 2253–2266. [CrossRef]
22. Alkhawatrah, M. Energy-Harvesting Cooperative NOMA in IOT Networks. *Model. Simul. Eng.* **2024**, *2024*, 1043973. [CrossRef]
23. Alharbi, T.E. Resource Allocation and Energy Harvesting in UAV-Assisted Full-Duplex Cooperative NOMA Systems. *Mathematics* **2025**, *13*, 3544. [CrossRef]
24. Minhaj, S.U.; Mahmood, A.; Abedin, S.F.; Hassan, S.A.; Bhatti, M.T.; Ali, S.H.; Gidlund, M. Intelligent resource allocation in LoRaWAN using machine learning techniques. *IEEE Access* **2023**, *11*, 10092–10106. [CrossRef]
25. Ahsan, W.; Yi, W.; Qin, Z.; Liu, Y.; Nallanathan, A. Resource allocation in uplink NOMA-IoT networks: A reinforcement-learning approach. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 5083–5098. [CrossRef]
26. Malik, T.S.; Malik, K.R.; Afzal, A.; Ibrar, M.; Wang, L.; Song, H.; Shah, N. RL-IoT: Reinforcement learning-based routing approach for cognitive radio-enabled IoT communications. *IEEE Internet Things J.* **2022**, *10*, 1836–1847. [CrossRef]

27. Jayakumar, S.; Nandakumar, S. Distributed resource optimisation using the Q-learning algorithm, in device-to-device communication: A reinforcement learning paradigm. *Results Eng.* **2024**, *23*, 102462. [CrossRef]
28. Jisi, C.; Roh, B.h.; Ali, J. Reliable paths prediction with intelligent data plane monitoring enabled reinforcement learning in SD-IoT. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102006. [CrossRef]
29. Chen, X.; Mao, Y.; Xu, Y.; Yang, W.; Chen, C.; Lei, B. Energy-efficient multi-hop LoRa broadcasting with reinforcement learning for IoT networks. *Ad. Hoc. Netw.* **2025**, *169*, 103729. [CrossRef]
30. Nishat, T.A.H.; Jeong, J.H.; Jo, H.; Xia, S.; Liu, J. Reinforcement learning for adaptive battery management of structural health monitoring IoT sensor network. *Appl. Energy* **2025**, *390*, 125731. [CrossRef]
31. Kang, J.M. Reinforcement learning based adaptive resource allocation for wireless powered communication systems. *IEEE Commun. Lett.* **2020**, *24*, 1752–1756. [CrossRef]
32. Omidkar, A.; Khalili, A.; Nguyen, H.H.; Shafiei, H. Reinforcement-learning-based resource allocation for energy-harvesting-aided D2D communications in IoT networks. *IEEE Internet Things J.* **2022**, *9*, 16521–16531. [CrossRef]
33. Ortiz, A.; Al-Shatri, H.; Li, X.; Weber, T.; Klein, A. Reinforcement learning for energy harvesting decode-and-forward two-hop communications. *IEEE Trans. Green Commun. Netw.* **2017**, *1*, 309–319. [CrossRef]
34. Ortiz, A.; Weber, T.; Klein, A. Multi-agent reinforcement learning for energy harvesting two-hop communications with a partially observable system state. *IEEE Trans. Green Commun. Netw.* **2021**, *5*, 442–456. [CrossRef]
35. Dutta, H.; Bhuyan, A.K.; Biswas, S. Cooperative Reinforcement Learning for Energy Management in Multi-Hop Networks with Energy Harvesting. *IEEE Trans. Green Commun. Netw.* **2025**, *9*, 1783–1793. [CrossRef]
36. Rummery, G.A.; Niranjan, M. *On-Line Q-Learning Using Connectionist Systems*; University of Cambridge, Department of Engineering: Cambridge, UK, 1994; Volume 37.
37. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
38. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, UK, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.