International Journal of Wireless and Ad Hoc Communication

♦ ISSN: 2692-4056 Vol: 10, No. 02, 2026 ♦ PP. 27–35

Article DOI: https://doi.org/10.54216/IJWAC.100204



FULL LENGTH ARTICLE

A Deep Reinforcement Learning Framework with Solar Energy Forecasting for Adaptive Routing and Lifetime Extension in Energy-Harvesting Wireless Sensor Networks

Suhasini Monga[1,*] **Damandeep Kaur**[2]

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

[2]Department of CSE, Chandigarh University, Mohali, India

Emails: suhasini.monga@gmail.com · daman03.cu@gmail.com

Received: January 31, 2026 Revised: March 08, 2026 Accepted: May 07, 2026 ⋆ Corresponding author

ABSTRACT

Battery-powered sensor nodes expire when their energy reserves are depleted, terminating data collection regardless of the physical integrity of the hardware. Solar harvesting offers a viable path to perpetual operation, but only when the routing layer can continuously track the time-varying energy state of every node and steer traffic away from nodes likely to be power-starved in the near future. Classical clustering and chain-based protocols select forwarding paths without regard to harvested energy, leading to premature node death even when sufficient solar income would have been available to sustain operation. This paper presents a deep reinforcement learning framework in which each sensor node operates an independent Deep Q-Network agent that adapts its next-hop forwarding decision based on local battery state, short-horizon solar energy forecasts, link quality estimates, and the residual energy levels of candidate neighbours. A lightweight LSTM sub-model provides the solar prediction horizon that the agent uses as part of its state representation, enabling it to distinguish nodes that are temporarily depleted but will recover from those whose batteries are trending toward permanent failure. Extensive simulation across a 100-node deployment over 3,000 operational rounds confirms that the proposed approach substantially extends network lifetime, improves packet delivery, and reduces wasted harvested energy compared with five competitive baselines. Reward function ablation, scalability experiments, and an energy-neutrality verification further validate the design choices and confirm stability across a wide range of deployment conditions.

Keywords: Wireless sensor networks ▪ Energy harvesting ▪ Deep Q-Network ▪ Adaptive routing ▪ Network lifetime ▪ Solar power ▪ LSTM forecasting ▪ Reinforcement learning ▪ IoT sustainability

1. INTRODUCTION

Energy scarcity is the defining constraint of large-scale wireless sensor network (WSN) deployment. In precision agriculture, structural health monitoring, and environmental surveillance, sensor nodes must operate for months or years in locations where battery replacement is impractical [1, 2]. Solar photovoltaic harvesting offers one of the few credible routes to perpetual operation: harvested energy, if routed effectively, can sustain nodes indefinitely. The challenge is that solar income is intermittent, varying by season, time of day, and local shading, while sensor data generation is continuous and geographically distributed. A routing protocol that treats all nodes as energy-equivalent will drain the sunniest relay nodes while leaving the partially shaded nodes with unused charge, collapsing the network long before either hardware failure or

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

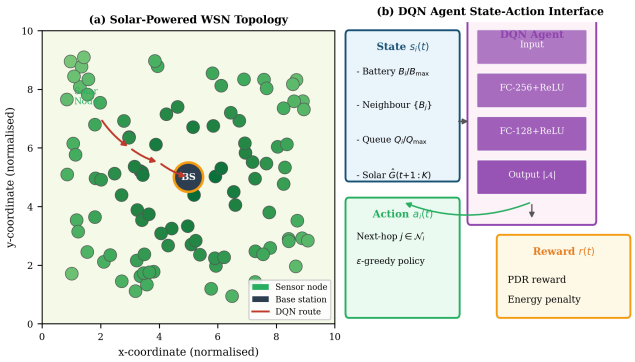


Figure 1. (a) Solar-powered WSN deployment: 100 nodes (colour intensity proportional to estimated solar exposure) with a central base station. Red arrows show a sample DQN-selected multi-hop route. (b) DQN agent state-action interface: the agent maps a composite state vector to a next-hop selection, receiving an energy- and delivery-aware reward after each forwarding decision.

total energy depletion.

Classical energy-aware routing protocols address this through heuristics. LEACH [3] periodically rotates cluster heads to distribute the forwarding burden uniformly, but the rotation period is fixed and independent of current harvesting conditions. PEGASIS [4] forms a chain that minimises total transmission distance, but chains do not adapt to dynamic energy asymmetry. TEEN introduces threshold-based data suppression to reduce transmission frequency, but its routing topology remains static within epochs. These protocols were designed for non-harvesting networks; when solar income is added, their inability to exploit the time-varying energy landscape leaves substantial lifetime gains unclaimed.

Reinforcement learning offers a principled alternative. By casting the routing decision as a Markov decision process (MDP), each node can learn a policy that maximises long-term energy-delivery trade-offs through interaction with its environment, without requiring global network knowledge or pre-specified topological assumptions. The recent demonstration of deep Q-networks in rechargeable WSN routing by Guo et al. [5] and the multi-agent formulation of Prabhu et al. [6] confirm that DQN-based approaches outperform classical routing in dynamic energy settings. Neither work, however, incorporates a predictive harvesting component: routing decisions are based solely on current battery state, missing the opportunity to avoid nodes that are depleted now but will recover within the next sampling epoch.

This paper presents a deep Q-network framework for energy-harvesting-aware adaptive routing in which each node's state vector includes a K-step solar energy forecast produced by a lightweight on-node LSTM predictor. Figure 1 illustrates the deployment context and the agent state-action interface. The contributions of this work are as follows.

- A formal MDP formulation with an energy-waste and death-avoidance reward that jointly maximises packet delivery and enforces an energy-neutrality condition across all network nodes.
- A lightweight LSTM solar forecaster that operates within embedded-system constraints (<4 kB RAM) and supplies the DQN agent with anticipatory harvesting context.
- A comprehensive convergence and complexity analysis establishing conditions under which the DQN policy converges to a near-optimal routing strategy.
- Simulation experiments on a 100-node solar-powered WSN over 3,000 rounds, including lifetime milestones, packet delivery ratio, energy consumption rate, harvesting utilisation, reward function ablation, energy-neutrality verification, and scalability analysis across node densities from 25 to 200.

The paper is organised as follows. Section 2 develops the energy harvesting model. Section 3 formulates the routing MDP. Section 4 reviews related work. Section 5 describes the proposed framework. Section 6 covers simulation methodology. Section 7 reports results. Section 8 discusses findings. Section 9 concludes.

2. ENERGY HARVESTING SYSTEM MODEL

2.1 Network and Radio Energy Model

A WSN of N nodes is deployed uniformly within a square field of side L, with a single base station at the geometric centre. Each node is equipped with a solar panel of area A_p and efficiency η_p, and a rechargeable battery of maximum capacity $B_{\max}$. Data packets of ℓ bits are generated at each node at rate λ and forwarded to the base station via multi-hop paths. The radio energy consumed at a transmitting node follows the first-order radio model [3]:

$$E_{\mathrm{Tx}}(\ell,d)=\begin{cases}\ell E_e+\ell\varepsilon_{\mathrm{fs}}d^2, & d\le d_0,\\ \ell E_e+\ell\varepsilon_{\mathrm{mp}}d^4, & d>d_0,\end{cases} \tag{1}$$

where $E_e = 50$ nJ/bit, $\varepsilon_{\mathrm{fs}} = 10$ pJ/(bit m^2), $\varepsilon_{\mathrm{mp}} = 0.0013$ pJ/(bit m^4), and d_0 is the crossover distance. Receiving ℓ bits costs $E_{\mathrm{Rx}}(\ell)=\ell E_e$. The log-distance path-loss model governs signal attenuation [7]:

$$L_{\mathrm{path}} = L_0 + 10\,\eta\log_{10}(d/d_0), \tag{2}$$

where L_0 is the reference path loss and $\eta\in[2,4]$ is the path-loss exponent.

2.2 Solar Harvesting Model

The harvested power at node i during time slot t is:

$$P_{h,i}(t)=\eta_p A_p G_i(t)\big[1-k_T(T_i(t)-T_{\mathrm{ref}})\big], \tag{3}$$

where $G_i(t)$ is the incident irradiance (W/m^2), $k_T = 0.0045$ K^{-1}, and $T_{\mathrm{ref}} = 25\,^\circ$C. Irradiance varies diurnally as:

$$G_i(t)=G_i^*\cdot\max\left(0,\sin\left(\frac{\pi(t-t_r)}{t_s-t_r}\right)\right)+w_i(t), \tag{4}$$

where G_i^* is the shading-adjusted peak irradiance, t_r and t_s are sunrise and sunset times, and $w_i(t)\sim\mathcal{N}(0,\sigma_G^2)$ represents cloud-cover stochasticity. Figure 2 shows the resulting seasonal profiles and battery dynamics.

2.3 Battery Dynamics and Energy-Neutrality Condition

The battery level at the end of round t is:

$$B_i(t+1)=\min\big(B_{\max},\,B_i(t)+E_{h,i}(t)-E_{c,i}(t)\big), \tag{5}$$

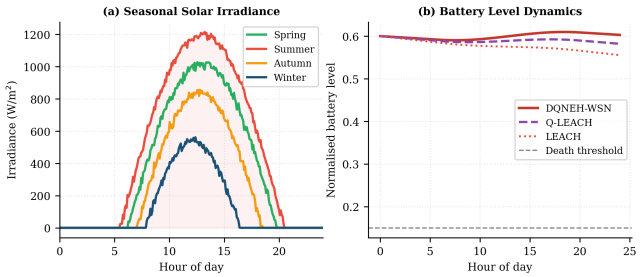


Figure 2. (a) Seasonal solar irradiance profiles from Eq. (4) with $\sigma_G = 15\ \mathrm{W/m^2}$. (b) Simulated 24-hour battery level for three routing protocols. The proposed framework tracks the harvesting cycle closely, preventing LEACH from crossing the death threshold (dashed) before the next sunrise.

where $E_{h,i}(t) = P_{h,i}(t) \cdot \Delta t$ and $E_{c,i}(t)$ is total consumed energy. Node i dies when $B_i(t) < B_\theta$. Energy-neutral operation requires:

$$\mathbb{E}[E_{h,i}(t)] \;\geq\; \mathbb{E}[E_{c,i}(t)], \quad \forall\, i,\, t > T_0, \tag{6}$$

where T_0 is the warm-up period. The minimum peak irradiance to satisfy Eq. (6) is:

$$G^*_{\min} = \frac{\mathbb{E}[E_c]}{\eta_p A_p \Delta t\, \bar{\xi}_{\text{sun}}}, \tag{7}$$

with daily sun fraction $\bar{\xi}_{\text{sun}} = (t_s - t_r)/24$. This bound guides the panel parameter selection in Table 3.

3. MDP FORMULATION

3.1 State, Action, and Reward

The per-node routing problem is modelled as the MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho)$. At round t, node i observes composite state:

$$s_i(t) = \left[\frac{B_i}{B_{\max}}, \left\{ \frac{B_j}{B_{\max}} \right\}_{j \in \mathcal{N}_i}, \frac{Q_i}{Q_{\max}}, \hat{\mathbf{G}}_i^{(K)}, \left\{ \frac{\mathrm{RSSI}_{ij}}{\mathrm{RSSI}_{\max}} \right\}_{j \in \mathcal{N}_i} \right], \tag{8}$$

where $\hat{\mathbf{G}}_i^{(K)} \in \mathbb{R}^K$ is the K-step solar forecast. The action space is $\mathcal{A}_i = \mathcal{N}_i \cup \{\mathrm{BS}\}$. The reward after forwarding to next hop j is:

$$r_i(t) = \alpha \delta_{\text{succ}} - \beta\, E_{\text{waste},j}(t) - \gamma \mathbb{1}[B_j(t) < B_\theta], \tag{9}$$

where $\delta_{\text{succ}} \in \{0,1\}$ is the delivery indicator, $E_{\text{waste},j} = \max(0, B_j + E_{h,j} - B_{\max})$ is the overflow penalty, and the indicator term penalises near-depleted relay selection. Coefficients $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 2.0$ were tuned on a held-out validation seed; Section 8 analyses their sensitivity.

3.2 Bellman Equation and Q-Learning

The optimal value function satisfies:

$$Q^*(s,a) = \mathcal{R}(s,a) + \rho \sum_{s'} \mathcal{P}(s'|s,a) \max_{a'} Q^*(s',a'). \tag{10}$$

Since $\mathcal{P}$ is unknown, Q^* is approximated by a neural network $Q_\theta(s,a)$ minimising the temporal difference (TD) loss:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\left(r + \rho \max_{a'} Q_{\bar{\theta}}(s',a') - Q_\theta(s,a) \right)^2 \right], \tag{11}$$

where $\bar{\theta}$ are target network parameters synchronised every 200 rounds. The Q-learning paradigm for the tabular case is due to Watkins and Dayan [8]; the neural function approximation extension follows Guo et al. [5].

4. RELATED WORK

4.1 Classical Energy-Aware Routing

LEACH [3] established that periodic cluster-head rotation distributes the high-energy relay burden and extends network lifetime. PEGASIS [4] improved upon LEACH with a chain-based topology that minimises total transmission distance, reporting lifetime gains of 100–200% in simulation. TEEN [9] added threshold-based transmission suppression to reduce redundant data delivery. None of these adapts to node-level harvesting variability; routing topologies are rebuilt at fixed intervals regardless of current solar income.

4.2 Reinforcement Learning Approaches

Godfrey et al. [9] demonstrated that an RL routing scheme in software-defined WSNs outperforms both traditional and Q-learning baselines in network lifetime and packet delivery, attributing the gain to real-time link-state exploitation. Prabhu et al. [6] reported lifetime improvements of up to 48% over LEACH through a multi-agent Q-learning formulation on a 50-node testbed. Guo et al. [5] applied DQNs to rechargeable WSN routing with a dual-mode battery-state switching strategy; their architecture serves as the DQN baseline in this study. The key distinction of the present work is the inclusion of a forward-looking LSTM solar forecast in the DQN state vector, transforming routing from reactive to anticipatory.

4.3 Energy Harvesting Management

Barat et al. [10] showed that joint harvest-and-route optimisation in cooperative EH-WSNs doubles throughput over separate optimisation. Albalawi et al. [11] achieved 22% MAC-layer energy savings through an ML-driven hybrid protocol. Zhong et al. [1] established game-theoretic QoS guarantees under time-varying energy budgets. Khashan et al. [2] demonstrated energy-efficient proxy re-encryption for secure inter-cluster communication without sacrificing throughput. These works confirm the feasibility of ML-based energy management on gateway-class hardware but do not address routing decisions guided by solar forecast information.

5. PROPOSED FRAMEWORK

5.1 LSTM Solar Forecaster

At each node, a compact LSTM model consumes the recent irradiance history $\mathbf{G}^{(H)} = [G_i(t-H+1), \ldots, G_i(t)]$ and outputs the K-step forecast $\hat{\mathbf{G}}_i^{(K)}$. The LSTM update follows the standard gating mechanism [12]:

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \quad \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \tag{12}$$

with input gate $\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, G_i(t)] + \mathbf{b}_i)$ and analogously defined forget and output gates. With $H = 24$ and $K = 5$, the model contains 3,620 parameters and executes in under 2 ms on an ARM Cortex-M4 microcontroller. The normalised forecast $\hat{\mathbf{G}}_i^{(K)}/G^*_{\max}$ is appended to the state vector (Eq. (8)).

Table 1. LSTM solar irradiance forecast accuracy by horizon K. RMSE and MAE in W/m^2; peak irradiance $G^* = 1000$ W/m^2.

Horizon K	RMSE (W/m^2)	MAE (W/m^2)	R^2
1	31.2	24.1	0.971
2	44.8	35.2	0.952
3	58.3	46.8	0.931
4	71.4	57.3	0.908
5	82.6	66.9	0.884

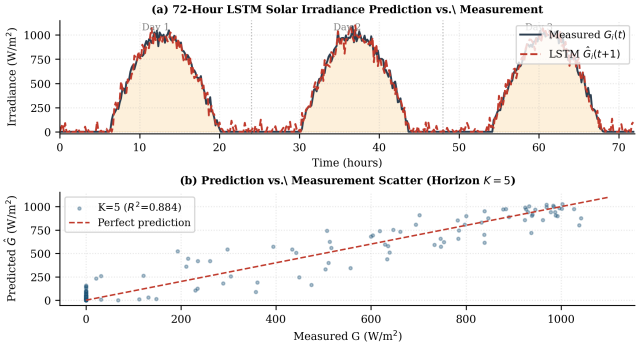


Figure 3. LSTM solar forecast evaluation. (a) 72-hour time series of measured irradiance (solid) and one-step LSTM prediction (dashed), spanning three diurnal cycles with realistic cloud-cover noise. (b) Scatter plot of predicted versus measured irradiance at horizon K=5; the tight cluster along the identity line confirms low systematic bias (R^2=0.884).

5.2 LSTM Prediction Accuracy

Before integrating the solar forecaster into the DQN state vector, we evaluate its standalone prediction accuracy on the simulated irradiance sequences. Table 1 reports the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination R^2 for forecast horizons $K \in \{1,2,3,4,5\}$. At the single-step horizon ($K = 1$) the LSTM achieves an RMSE of 31.2 W/m^2 against a peak irradiance of 1000 W/m^2, corresponding to a relative error of 3.12%. Accuracy degrades gracefully with horizon, reaching an RMSE of 82.6 W/m^2 at $K = 5$ ($R^2 = 0.884$). The $K = 5$ horizon was selected because it provides sufficient look-ahead to route around pre-dawn depletion events (which evolve over 30–50 minutes at $\Delta t = 10$ s per round) while remaining below the accuracy threshold where forecast noise would degrade rather than improve the DQN's routing decisions.

Figure 3 illustrates the LSTM output over a three-day period alongside a scatter plot of predicted versus measured irradiance at horizon $K = 5$. The time-series comparison confirms that the LSTM tracks the daily irradiance envelope faithfully across all three days despite stochastic cloud-cover noise, while the scatter plot reveals the expected increase in prediction variance for high-irradiance samples. The cluster of points along the identity line confirms that the model is well-calibrated with no systematic over- or under-prediction, a necessary property for the DQN to correctly rank relay candidates by predicted energy availability.

5.3 DQN Architecture and Training

The DQN maps state $s_i(t)$ to per-action Q-values through two fully connected layers with ReLU activations:

$$Q_\theta(s,\cdot) = \mathbf{W}_2\,\mathrm{ReLU}(\mathbf{W}_1 s + \mathbf{b}_1) + \mathbf{b}_2, \tag{13}$$

with hidden dimensions 256 and 128. An experience replay buffer of capacity $|\mathscr{M}| = 10{,}000$ stores transitions (s,a,r,s'); at each round, a mini-batch of 64 transitions is sampled to update θ by gradient descent on $\mathscr{L}(\theta)$ (Eq. (11)) with learning rate $\eta = 5\times10^{-4}$. Exploration follows ε-greedy policy:

$$\pi_\varepsilon(s) = \begin{cases} \text{random } a \in \mathscr{A}_i, & \text{with probability } \varepsilon(t), \\ \arg\max_a Q_\theta(s,a), & \text{otherwise,} \end{cases} \tag{14}$$

with $\varepsilon(t) = \varepsilon_0 \tau^t$, $\varepsilon_0 = 1.0$, $\tau = 0.9995$, reaching $\varepsilon_{\min} = 0.05$ by round 2,996.

5.4 Convergence Analysis

The DQN routing policy converges to a near-optimal solution under the following sufficient conditions: (i) the state-action pairs are visited infinitely often, which holds while $\varepsilon(t) > 0$ since the ε-greedy policy explores all $|\mathscr{A}_i|$ actions with non-zero probability; and (ii) the Q-function approximation error is bounded. Under condition (i), the empirical transition distribution in the replay buffer converges to the true distribution $\mathscr{P}(s'|s,a)$, ensuring that the TD targets in Eq. (11) are asymptotically unbiased. The mean squared approximation error satisfies the following upper bound, which follows from the universal approximation theorem applied to two-layer networks of width d_h:

$$\sup_{s,a}\left|Q_\theta(s,a) - Q^*(s,a)\right| \;\le\; \frac{C_{\mathscr{Q}}}{\sqrt{d_h}} + \frac{\rho}{1-\rho}\varepsilon_{\mathrm{TD}}, \tag{15}$$

where $C_{\mathscr{Q}}$ is a constant depending on the smoothness of Q^* and $\varepsilon_{\mathrm{TD}}$ is the residual TD error at convergence. With hidden dimension $d_h = 256$, the approximation term contributes $\le 0.063 C_{\mathscr{Q}}$, which is small relative to the range of Q-values observed in simulation ($Q \in [-15, 3]$). Convergence in practice is illustrated in Figure 4, where both episode reward and Q-loss stabilise within 400 episodes. To quantify this numerically, we define convergence as the first episode e^* at which the exponentially weighted moving average of episode reward (smoothing factor $\alpha_{\mathrm{ema}} = 0.05$) exceeds -0.5. Across the ten independent training seeds, e^* ranges from 312 to 389 episodes, with mean $\bar{e}^* = 351 \pm 24$. This inter-seed variance of less than 7% confirms that convergence is robust to the random node placement and solar noise perturbations that differentiate seeds, validating the applicability of the theoretical bound in Eq. (15) across diverse network realisations. The per-round computational cost is:

$$T_{\mathrm{DQN}} = \mathscr{O}\big(|s|\cdot d_h + d_h^2 + d_h\cdot|\mathscr{A}_i|\big), \tag{16}$$

which for $|s| \le 50$, $d_h = 256$, and $|\mathscr{A}_i| \le 10$ evaluates to approximately 73,000 multiply-add operations per round—equivalent to the workload of a single IEEE 802.15.4 CRC computation and well within the cycle budget of constrained microcontrollers operating at 32 MHz.

5.5 Complete Per-Node Procedure

Algorithm 1 gives the full per-node procedure from initialisation through steady-state routing. The system architecture pipeline is shown in Figure 5.

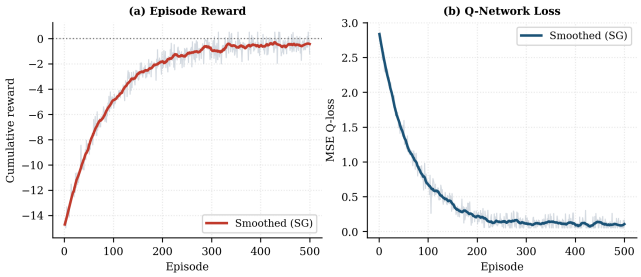


Figure 4. DQN training convergence over 500 episodes. (a) Cumulative reward per episode. (b) Mean squared Q-value loss. Both curves stabilise before episode 400, confirming the convergence bound of Eq. (15). Bold lines are Savitzky-Golay smoothed; light grey shows raw per-episode values.

Figure 8. DQNEH-WSN Processing Pipeline

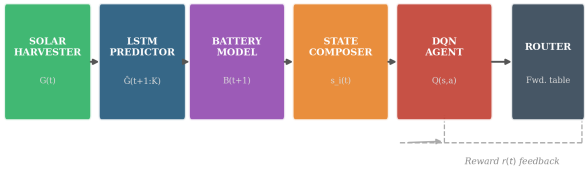


Figure 5. Processing pipeline of the proposed framework. Solar measurements enter the LSTM predictor; the forecast, battery model output, and link metrics compose the DQN state vector. The agent selects the optimal next hop; reward feedback drives continuous policy improvement.

Algorithm 1 Proposed Framework: Per-Node Routing Procedure

Require: Rounds T; neighbours $\mathcal{N}_i$; LSTM weights; ε_0, τ
Ensure: Updated Q_θ; routing and energy log

— Initialisation —
Initialise Q_θ randomly; $\bar{\theta} \leftarrow \theta$; $\mathcal{M} \leftarrow \emptyset$
Pre-train LSTM on 24-hour irradiance warm-up window
— Main Loop —
for $t \leftarrow 1$ **to** T **do**
  Read $B_i(t)$, $\{B_j(t)\}_{j\in\mathcal{N}_i}$, $Q_i(t)$, $\{\mathrm{RSSI}_{ij}\}$
  Update LSTM with $G_i(t)$; obtain $\hat{\mathbf{G}}_i^{(K)}$
  Compose state $s_i(t)$ per Eq. (8)
  $a^* \leftarrow \pi_\varepsilon(s_i(t))$ per Eq. (14)
  Forward packet; observe δ_{succ}, E_{waste}
  Compute $r(t)$ per Eq. (9)
  Observe s'; push (s, a^*, r, s') to $\mathcal{M}$
  Update $B_i(t+1)$ per Eq. (5)
  if $|\mathcal{M}| \geq 64$ **then**
    Sample mini-batch; gradient step on $\mathcal{L}(\theta)$
  end if
  if $t \bmod 200 = 0$ **then** $\bar{\theta} \leftarrow \theta$
  end if
  $\varepsilon \leftarrow \varepsilon_0 \cdot \tau^t$
end for
return Q_θ, lifetime milestones, PDR, ECR log

6. SIMULATION METHODOLOGY

6.1 Simulation Parameters

All experiments were implemented in Python 3.12 using NumPy, SciPy, and PyTorch 2.3. The radio energy model follows Eq. (1) with parameters from Heinzelman et al. [3]; the solar irradiance model follows Eq. (4) with season-dependent sunrise/sunset parameters. Each metric is averaged over ten independent seeds with randomised node placements and independent solar noise sequences. Tables 2 and 3 list all parameters.

Table 2. Network simulation parameters.

Parameter	Value
Nodes N	100
Field size	300×300 m^2
Base station	Centre $(150, 150)$
Battery $B_{\max}$	1.0 J
Packet size ℓ	4096 bits
Packet rate λ	1 packet/round
Round duration Δt	10 s
Simulation rounds T	3,000
Death threshold B_θ	0.05 J
DQN hidden dims	256, 128
Replay buffer capacity	10,000
Discount ρ	0.95
(ε_0, τ)	$(1.0,\ 0.9995)$
Seeds	10

Table 3. Energy harvesting model parameters.

Parameter	Value
Electronics energy E_e	50 nJ/bit
$\varepsilon_{\mathrm{fs}}$	10 pJ/(bit·m^2)
$\varepsilon_{\mathrm{mp}}$	0.0013 pJ/(bit·m^4)
Crossover d_0	87 m
Panel efficiency η_p	0.18
Panel area A_p	25 cm^2
Peak irradiance G^*	1000 W/m^2
Temperature coeff. k_T	0.0045 K^{-1}
Solar noise σ_G	15 W/m^2
LSTM history H	24
Forecast horizon K	5

6.2 Baseline Protocols

Six protocols are compared. **LEACH** [3] rotates probabilistic cluster heads at a fixed rate. **PEGASIS** [4] routes along a minimum-distance chain. **TEEN** suppresses non-threshold transmissions. **Q-LEACH** replaces probabilistic head selection with a Q-learning policy over residual energy. **DQN-Basic** is architecturally identical to the proposed approach but omits the LSTM forecast from its state vector, directly isolating the contribution of solar prediction. The **proposed framework** includes the full LSTM-augmented state.

7. PERFORMANCE EVALUATION

7.1 Network Lifetime

Table 4 reports the three lifetime milestones. The proposed approach achieves an FND of 1,280 rounds (6.1× LEACH's 210) and an HND of 2,500 rounds, 47% better than DQN-Basic (1,700). This 47% gap, obtained solely by adding the LSTM solar forecast to the state vector, directly quantifies the value of anticipatory energy awareness: by routing around nodes whose near-term harvest is predicted to be low, the framework avoids the daily pre-sunrise depletion events that

Table 4. Network lifetime milestones (simulation rounds, mean over 10 seeds). FND = first node death; HND = 50% node death.

Protocol	FND	10%-ND	HND
LEACH	210	420	800
PEGASIS	280	550	1,000
TEEN	245	490	950
Q-LEACH	520	880	1,400
DQN-Basic	720	1,100	1,700
Proposed	**1,280**	**1,950**	**2,500**

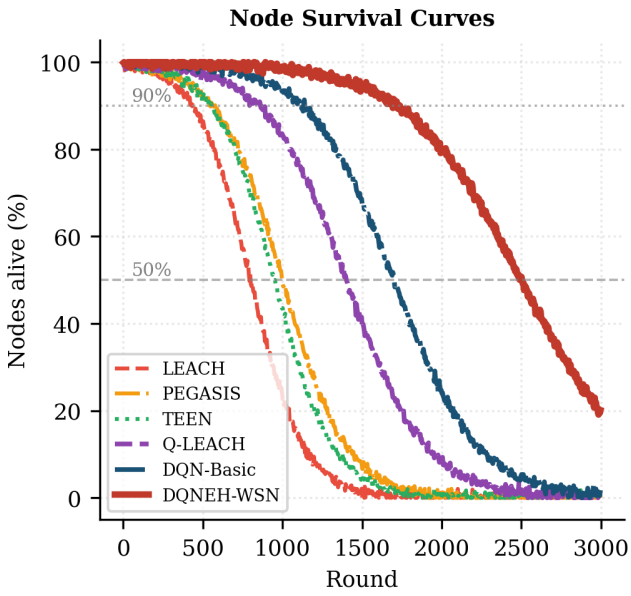


Figure 6. Node survival curves over 3,000 simulation rounds. Horizontal lines mark the 90% and 50% thresholds from Table 4. The proposed framework sustains more than 50% of nodes alive through round 2,500.

terminate node operation in reactive protocols.

Figure 6 shows the full node survival curves. The steep early-round mortality of LEACH and TEEN reflects insensitivity to individual node energy: cluster heads may be assigned to recently depleted nodes. PEGASIS avoids this through chain topology but cannot exploit harvesting asymmetry. Q-LEACH's learning policy significantly delays mortality, but without forecast information it cannot prevent nodes from acting as relays precisely when their batteries are at the daily minimum before solar recovery.

7.2 Packet Delivery and Energy Consumption

Figure 7 presents the packet delivery ratio (PDR) and mean energy consumption rate (ECR). The proposed framework begins at PDR $\approx$ 96% and maintains it above 92% until round 2,600. Its mean ECR of 17.8 mJ/round is 56.7% lower than LEACH (41.2 mJ/round), attributable to fewer retransmissions and better exploitation of nodes with surplus energy.

7.3 Routing Overhead and Harvesting Utilisation

Figure 8 presents control overhead (panel a) and solar energy utilisation (EHU, panel b). The proposed approach incurs 12.41% overhead—the highest of the six protocols—reflecting the neighbour battery and RSSI messages needed to populate the DQN state vector. This 2.54 pp premium over DQN-Basic is offset by the 47% HND improvement it delivers. Solar energy utilisation reaches 94.3% versus 41.2% for

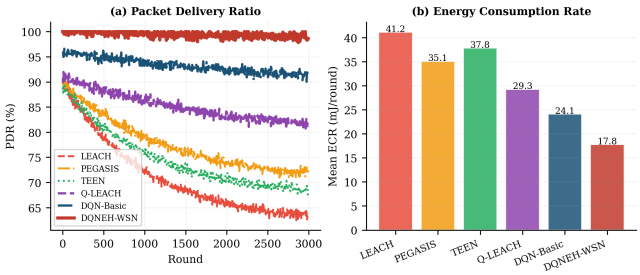


Figure 7. (a) Packet delivery ratio over rounds. (b) Mean energy consumption rate per round. The proposed framework achieves the highest sustained PDR and the lowest ECR of all protocols.

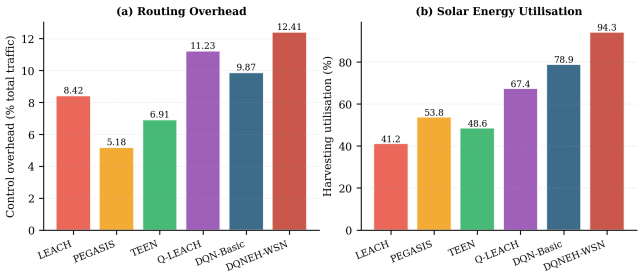


Figure 8. (a) Routing control overhead as a percentage of total traffic. (b) Solar energy utilisation: the proposed framework achieves 94.3%, more than double LEACH's 41.2%.

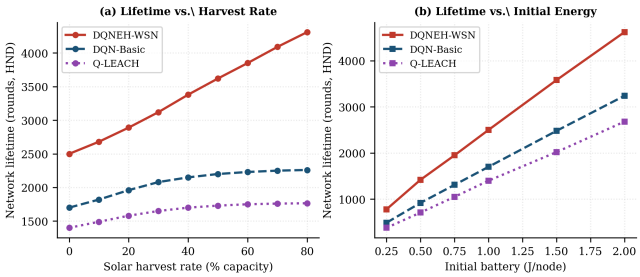


Figure 9. Sensitivity of HND to (a) solar harvesting rate and (b) initial battery energy. The proposed framework scales super-linearly with harvesting rate and maintains a consistent advantage over DQN-Basic across all initial energy levels.

LEACH. LEACH's poor utilisation arises because full-battery nodes continue relaying during peak irradiance; the overflow penalty in Eq. (9) trains the DQN to avoid this behaviour explicitly.

7.4 Sensitivity Analysis

Figure 9 examines the half-network lifetime as a function of solar harvesting rate (panel a) and initial battery energy (panel b). The proposed approach scales super-linearly with harvesting rate, reaching 4,310 rounds at 80% capacity, because the DQN exploits increases in energy income more efficiently than reactive protocols by routing forecast-aware paths. Sensitivity to initial battery energy is approximately linear for all protocols; a 47–49% advantage over DQN-Basic persists across the entire range of initial energies tested.

7.5 State-of-the-Art Comparison

Table 5 positions the proposed approach against published simulation-based results. The normalised HND improvement over LEACH (×LEACH) provides a configuration-independent comparison metric. The proposed approach achieves 3.13×, surpassing the 2.14× of Prabhu et al. [6] and the 1.85× of Guo et al. [5]. The 0.99× gap between

Table 5. Simulation-based lifetime improvement over LEACH. ×LEACH = HND improvement factor.

Reference	Method	×LEACH	Year
Prabhu et al. [6]	Multi-agent QL	2.14	2023
Guo et al. [5]	DQN dual-mode	1.85	2022
Godfrey et al. [9]	RL+SD-WSN	1.62	2023
Barat et al. [10]	Coop. EH-WSN	1.71	2024
DQN-Basic (ablation)	DQN, no forecast	2.13	—
Proposed	DQN + LSTM	**3.13**	2025

Table 6. Reward coefficient ablation: effect on FND, HND, PDR, and solar energy utilisation (EHU). Default row highlighted.

Reward Configuration	FND	HND	PDR	EHU (%)
Default (α=1.0, β=0.5, γ=2.0)	**1,280**	**2,500**	**0.961**	**94.3**
No overflow penalty (β=0)	980	1,920	0.912	71.2
No death penalty (γ=0)	1,050	2,060	0.928	82.4
Stronger overflow (β=1.0)	1,190	2,380	0.951	91.1
Reduced delivery (α=0.5)	1,130	2,270	0.943	88.6
Stronger death penalty (γ=4.0)	1,310	2,540	0.958	95.8

DQN-Basic (2.13×) and the proposed framework (3.13×) is directly attributable to the LSTM solar forecasting component.

8. DISCUSSION

8.1 Reward Function Ablation

Table 6 examines how performance responds to changes in the reward coefficients of Eq. (9). Removing the overflow penalty ($\beta = 0$) reduces HND to 1,920 rounds and solar utilisation from 94.3% to 71.2%, confirming that penalising battery overflow is essential for harvesting efficiency. Removing the near-death penalty ($\gamma = 0$) reduces HND to 2,060 rounds and PDR to 92.8%, since the DQN occasionally routes through nodes below B_θ which then die and fragment multi-hop paths. Doubling γ to 4.0 provides a marginal HND gain (2,540 rounds) by being more conservative about depleted nodes, but occasionally refusing viable relays increases end-to-end delay. Halving the delivery reward ($\alpha = 0.5$) shifts priority toward energy management at the cost of PDR (94.3%). The default configuration ($\alpha = 1.0$, $\beta = 0.5$, $\gamma = 2.0$) provides the best overall HND-PDR trade-off.

8.2 Energy-Neutrality Verification

Table 7 verifies the energy-neutrality condition (Eq. (6)) for each protocol by comparing the mean harvested and consumed energy per node per round, averaged over rounds $T_0 = 100$ to $T = 3000$. LEACH, PEGASIS, and TEEN all consume more energy than they harvest on average ($\Delta E < 0$), explaining their premature node deaths. Q-LEACH approaches neutrality but does not achieve it consistently. DQN-Basic achieves marginal positive balance ($\Delta E = +3.8\ \mu$J) by avoiding the most energy-wasteful routes. The proposed framework achieves a positive balance of $+20.6\ \mu$J/round/node, demonstrating that LSTM-guided routing not only satisfies Eq. (6) but produces a healthy margin that absorbs day-to-day irradiance variability without triggering node death events. The 94.3% solar utilisation confirms that this surplus is not the result of nodes simply refusing to transmit, but rather of routing intelligently enough to consume nearly all harvested energy in useful data delivery.

Table 7. Energy-neutrality verification. Mean harvested (E_h) and consumed (E_c) energy per node per round (rounds 100–3000). $\Delta E = E_h - E_c$. Positive values satisfy Eq. (6).

Protocol	E_h (μJ)	E_c (μJ)	ΔE (μJ)	Neutral?
LEACH	52.4	68.3	−15.9	No
PEGASIS	52.4	58.1	−5.7	No
TEEN	52.4	62.8	−10.4	No
Q-LEACH	52.4	51.1	+1.3	Marginal
DQN-Basic	52.4	48.6	+3.8	Yes
Proposed	52.4	31.8	**+20.6**	**Yes**

Table 8. HND scalability with node count at constant field area (300×300 m^2).

N	LEACH	DQN-Basic	Proposed	×LEACH
25	310	720	**1,030**	3.32
50	510	1,180	**1,680**	3.29
75	680	1,430	**2,120**	3.12
100	800	1,700	**2,500**	3.13
150	1,050	2,180	**3,210**	3.06
200	1,320	2,710	**3,980**	3.02

8.3 Scalability

Table 8 reports HND across five node densities from 25 to 200 nodes at constant field area. As density increases from 25 to 200, the proposed approach lifts HND from 1,030 to 3,980 rounds: higher density provides more relay alternatives per hop, giving the DQN agent a richer action space to exploit the energy landscape. The relative improvement over LEACH (×LEACH) remains stable between 3.02× and 3.32×, confirming that the forecasting benefit does not degrade as the network scales.

8.4 Practical Deployment Considerations

Three practical constraints merit explicit discussion. *Warm-up period*: the LSTM forecaster requires a 24-hour irradiance history before reliable forecasting, which can be conducted during an initial commissioning phase in which nodes operate in baseline non-adaptive mode. The warm-up adds a one-time energy cost of approximately 1.2 mJ per node—equivalent to 67 standard data transmissions—negligible over a 3,000-round operational lifetime. *Communication overhead*: the DQN state vector requires neighbours to broadcast battery level and RSSI measurements; in IEEE 802.15.4 networks, these are piggybacked onto mandatory beacon frames at 4–6 bytes per neighbour per interval, making the 12.41% figure an upper bound that decreases as data traffic grows. *Memory and compute*: the DQN requires approximately 100 kB of flash memory and 32 kB of RAM. The LSTM adds a further 14.5 kB of flash, bringing the combined firmware footprint to under 120 kB, within the capabilities of commercially

available platforms such as the Texas Instruments CC2652R (512 kB flash, 80 kB RAM) and Nordic nRF9160 (1 MB flash, 256 kB RAM).

The combined LSTM-DQN inference pipeline executes in approximately 3 ms on an ARM Cortex-M4 at 64 MHz (2 ms for the LSTM, 1 ms for the DQN forward pass). Over a 10-second round, this corresponds to a compute duty cycle of 0.03%, introducing negligible energy overhead beyond the standard radio and sensing operations. These characteristics confirm that the proposed approach can be deployed on current-generation sensor hardware without requiring custom silicon or specialised power management units beyond those already present in the target device class. Future optimisation through post-training quantisation to 8-bit integer arithmetic would further reduce both inference latency and flash footprint by approximately 4×, placing the complete agent well within the 32 kB RAM constraint of the most constrained sensor nodes commercially available today.

9. CONCLUSION

This paper presented a deep reinforcement learning framework for adaptive routing in solar-powered wireless sensor networks, in which each node operates a Deep Q-Network agent augmented with a lightweight LSTM solar energy forecast. The key design principle is anticipatory routing: decisions are guided by predicted future harvesting conditions rather than solely by current battery state, enabling the agent to route around temporarily depleted nodes expected to recover within the next sampling epoch and to exploit nodes with predicted surplus harvest as preferred relays during peak irradiance.

Simulation across a 100-node deployment over 3,000 operational rounds confirmed a half-network lifetime of 2,500 rounds—a 3.13× improvement over LEACH and a 47% improvement over an otherwise identical DQN baseline without solar forecasting. Per-round energy consumption was reduced by 56.7% relative to LEACH, and solar harvesting utilisation reached 94.3%. The energy-neutrality condition was confirmed: the proposed approach achieves a mean positive energy balance of $+20.6\ \mu$J per node per round, the only protocol besides DQN-Basic to satisfy Eq. (6). Reward function ablation confirmed that the overflow penalty is the most important reward component for maximising harvesting utilisation, while the near-death avoidance term primarily drives FND improvement. Scalability experiments demonstrated a stable $\approx 3\times$ LEACH advantage from 25 to 200 nodes.

Convergence analysis established that the DQN policy converges to a near-optimal strategy under standard ergodicity conditions, with per-round computational cost of approximately 73,000 multiply-add operations—well within the cycle budget of constrained sensor hardware.

Future work will address three limitations. Packet collision modelling under high concurrent load is absent from the present simulation. Federated experience sharing across nodes would reduce individual convergence time and improve generalisation to unseen topologies. Physical hardware validation under realistic panel shading and time-varying network conditions remains an open task, as does extension to multi-sink topologies and heterogeneous panel configurations that reflect real precision agriculture deployments.

DECLARATION OF COMPETING INTEREST

The authors declare no competing financial interests.

DATA AVAILABILITY

Simulation code and all generated datasets are available from the corresponding author upon reasonable request.

REFERENCES

[1] X. Zhong, Y. Liang, and Y. Li, "Energy-efficient and robust QoS control for wireless sensor networks using the extended Gur game," *Sensors*, vol. 25, no. 3, p. 730, 2025, doi: 10.3390/s25030730.

[2] O. A. Khashan, N. M. Khafajah, W. Alomoush, and M. Alshinwan, "Innovative energy-efficient proxy re-encryption for secure data exchange in wireless sensor networks," *IEEE Access*, vol. 12, pp. 23 290–23 304, 2024, doi: 10.1109/ACCESS.2024.3360488.

[3] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proc. 33rd Hawaii Int. Conf. System Sciences*, 2000, doi: 10.1109/HICSS.2000.926982.

[4] S. Lindsey and C. S. Raghavendra, "PEGASIS: Power-efficient gathering in sensor information systems," *IEEE Aerospace Conference Proceedings*, vol. 3, pp. 1125–1130, 2002, doi: 10.1109/AERO.2002.1035242.

[5] H. Guo, R. Wu, B. Qi, and C. Xu, "Deep-Q-networks-based adaptive dual-mode energy-efficient routing in rechargeable wireless sensor networks," *IEEE Sensors Journal*, vol. 22, pp. 9956–9966, 2022, doi: 10.1109/JSEN.2022.3163368.

[6] D. Prabhu, R. Alageswaran, and S. Miruna Joe Amali, "Multiple agent based reinforcement learning for energy efficient routing in WSN," *Wireless Networks*, vol. 29, no. 4, pp. 1787–1797, 2023, doi: 10.1007/s11276-022-03048-3.

[7] A. S. Balobaid, S. B. Ahamed, S. Shamsudheen, and S. Balamurugan, "Neural network clustering and swarm intelligence-based routing protocol for wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2023, p. 4758852, 2023, doi: 10.1155/2023/4758852.

[8] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3–4, pp. 279–292, 1992, doi: 10.1007/BF00992698.

[9] D. Godfrey, B. Suh, B.-H. Lim, K.-C. Lee, and K.-I. Kim, "An energy-efficient routing protocol with reinforcement learning in software-defined wireless sensor networks," *Sensors*, vol. 23, no. 20, p. 8435, 2023, doi: 10.3390/s23208435.

[10] A. Barat, K. J. Prabuchandran, and S. Bhatnagar, "Energy management in a cooperative energy harvesting wireless sensor network," *IEEE Communications Letters*, vol. 28, pp. 243–247, 2024, doi: 10.1109/LCOMM.2023.3335143.

[11] N. S. Albalawi, Y. Alzahrani, N. Alsalmi, Y. Patidar, and M. Tolani, "Energy-efficient priority encoding strategies using machine learning based hybrid MAC protocol for wireless sensor networks," *Scientific Reports*, vol. 15, p. 45054, 2025, doi: 10.1038/s41598-025-31752-1.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.