International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 467-480
https://doi.org/10.70917/ijcisim-2026-4430
Article

Proximal Policy Optimization for Adaptive Cluster Head Selection in Wireless Sensor Networks

Rabinarayan Panda[1], Sahana Edwin[2], John P[3], Thalaimalaichamy M[4], Madhawa Surendar[5*]

[1,2,3,5] Department of Computer Science and Engineering, Garden City University, Bengaluru, India.
[4]Department of Electronics and Communication Engineering, SASTRA Deemed University, Tamil Nadu, India.
m.surendar@gmail.com

Abstract: The energy-efficient operations of WSNs have become important due to the limited battery life of the sensor nodes. WSNs are increasingly used increasingly for mission-critical applications like environmental monitoring, precision agriculture, and industrial automation. Choosing the correct cluster head is one of the core deficiencies in the implantation of any clustering algorithm which greatly affect network longevity and energy consumption. Protocols designed to achieve LEACH, iLEACH, SEP, TEEN, and Sim-TEEN use static heuristics or probability models. They do not adapt to dynamic variations of remaining energy and changing topology. These lead to early energy depletion of nodes. They apply unequal utilization of nodes. To address these limitations, a CH selection method based on Proximal Policy Optimization (PPO) using reinforcement learning is proposed in this paper. A WSN (wireless sensor network) environment, compatible with OpenAI Gym, is created to simulate a field deployment with 200 sensor nodes in a 300×300 m² area. The PPO agent's policy runs for 100,000 timesteps for a single episode of 1000 rounds which helps the policy learn about temporal patterns and long-term energy. The method adaptively chooses CHs based on energy thresholds, node distribution and communication range constraints while minimizing energy loss and ensuring fairness. PPO significantly outperforms conventional algorithms in terms of savings of residual energy as well as node survival, as shown by the simulation results. The article compares the performance of a novel protocol for wireless sensor networks with previous methods. The results indicate the robustness, flexibility and applicability of PPO in energy efficient WSN protocol design with a scalable and intelligent solution to the traditional CH selection scheme.

Keywords: -Proximal Policy Optimization (PPO), Cluster Heads (CHs), WSN (wireless sensor network), OpenAI Gym-compatible, energy-efficient

1. Introduction

WSNs are already a key technology for modern data-driven monitoring systems and are a key component of many applications from environmental sensing and precision farming to industrial automation and smart cities. WSNs consist of spatially distributed sensor nodes that have sensing, computing and communication capabilities to report the environment to centralized sinks or base stations. not done yet Sensor nodes usually run on low capacity batteries, and in several situations, like hostile environment, underwater use, or disaster zones, it's not possible or downright impossible to change or recharge the batteries. As pointed out by Akyildiz et al. [1], 70% of WSNs' failure is due to exhaustion of energy, in turn affecting coverage, correctness of data and reliability. To guarantee energy-conserving operation, one well-known method is clustering, whereby some nodes act as Cluster Heads (CHs) to gather and forward data to the base station. Optimistic CH selection can considerably reduce energy consumption, minimize communication overhead, and distribute the workload evenly among the nodes. Various classical protocols such as LEACH (Low-Energy Adaptive Clustering Hierarchy) [2], iLEACH [3], SEP (Stable Election Protocol) [4], and TEEN (Threshold-sensitive Energy Efficient sensor Network protocol) [5] have been extensively used to solve this issue. These approaches rely mostly on static heuristics or probability models with predetermined parameters for CH election.

While effective in static environments, these strategies find it difficult to dynamically respond to real-world deployments where network conditions change as a result of node mobility, uneven energy expenditure, or changes in topology. As the research of Heinzelman et al. [6], Smaragdakis et al. [7], and Manjeshwar and Agrawal [8] points out, this inability to adapt results in premature battery depletion, uneven energy distribution, and a reduced total network lifetime.

Whereas CH selection has been the subject of much research, temporal context-awareness is largely absent from most protocols, and time-evolving factors such as residual energy behaviors or adaptive learning are not appropriately included. This necessitates context-aware and learning-based models that can perform in time-evolving WSNs and enhance the lifespan of the network.

To tackle these issues, the major research question raised is: "How can Proximal Policy Optimization (PPO), a reinforcement learning algorithm, be used to adaptively and optimally select CH in dynamic WSN environments to enhance energy efficiency and network lifetime?" To investigate this, the current research employs the Proximal Policy Optimization (PPO), which is a powerful and reliable reinforcement learning (RL) algorithm, to solve the CH selection problem. The WSN is modelled as a Markov Decision Process (MDP), where the RL agent perceives the network state containing parameters such as residual energy, node position, and past CH involvement. Feedback through the reward function is also transmitted to the agent. The reward function encapsulates energy fairness, network lifetime, and balance. An environment is developed for simulating 200-node WSN in a region of 300×300 m² compatible with OpenAI gym.

The PPO agent is trained for more than 100,000 timesteps in a single long episode consisting of 1000 rounds, which allows it to learn long-term temporal correlations and energy-efficient decision-making methods. In contrast to the conventional method of resetting CH selection every round using current state only, the PPO agent constructs a cumulative picture of energy dynamics, which allows for more intelligent and sustainable CH selections with time.

1.1. Objectives of the Research

The main aims of this work are:

- To cast CH selection as a sequential decision-making problem and reduce it to a reinforcement learning problem.
- To develop an OpenAI Gym-compatible simulation platform to accurately simulate a 200-node WSN in a 300×300 m² region.
- To develop and train a PPO agent with Stable-Baselines3 for the best CH selection according to node-level energy behaviours.
- To compare the novel PPO-based CH selection strategy with conventional protocols like LEACH, iLEACH, SEP, and TEEN.

1.2. Methodology

We created an appropriate OpenAI Gym compatible simulation that mimicked real-world WSN scenarios, like exhaustion of node energy, backhaul space constraints, and the presence of realistic overheads. The reinforcement learning agent based on PPO was taught for longer than 100,000 timesteps equating to one long episode with 1000 rounds of communication. Through each round, the agent observes node level features such as residual energy, spatial location, etc., and dynamically selects suitable CHs, that minimizes the energy loss and enhances the lifetime. The assessment of the adaptive performance of this technique will be performed against four prominent CH selection protocols. The comparisons will be through the simulations throughout the full 1000 rounds. Thus, it will demonstrate substantial improvements in long-term sustainability and fairness.

The organization of the article is structured as follows: Section 2 summarizes prior work in energy-efficient CH selection and reinforcement learning for WSNs. Section 3 explains the design of the Gym-compatible WSN environment and PPO agent formulation. Section 4 describes implementation, training configuration, and simulation parameters. Section 5 reports performance results and comparisons for major metrics. Section 6 presents the benefits of the new PPO approach with respect to energy efficiency, fairness, and flexibility over traditional and state-of-the-art methodologies. Section 7 summarizes the paper and presents future directions such as real-world deployment and multi-objective extensions.

2. Literature Survey

Du et al. [9] integrated fuzzy inference and reinforcement learning for CH selection in heavy-density WSNs, maximizing link stability and energy consumption. Compared with LEACH, HEED, and PSO, the technique delivered significant packet delivery, delay reduction, and energy efficiency gains—but at the cost of requiring heavy-density deployment to perform optimally. Wang et al. [10] proposed a dual-agent RL framework that simultaneously optimizes clustering and multi-hop routing. While providing balanced node energy and increased throughput, its RL-based central architecture in such large deployments is likely to scale poorly. Lewandowski & Płaczek [11] presented an analytical model for CH rotation in maximizing "last-node-to-die" in LoRaWAN-based WSNs. The probabilistic solution maximizes system lifetime, though tests are simulation-based with no real-world experiments.

Kaur & Aulakh [12] put forward an ultra-lightweight RL-based protocol optimizing CH selection with residual energy and distance. While it performs better than heuristic algorithms, the shallow Q-learning-based method falters at high node heterogeneity. Ding et al. [13] introduced a multi-agent RL game in which UAVs support WSN data gathering and CH selection. The distributed agents save energy, but the simulated results might not accurately reflect actual channel dynamics.

Rami Reddy et al. [14] integrated RL with PSO into an exploration-exploitation hybrid framework for CH selection. Their approach enhances stability but requires additional computation when assigning CH. Jurado-Lasso et al. [15] proposed LEACH-RLC, a centralized RL algorithm based on MILP for adaptive triggering of CH. Although minimizing control overhead and enhancing lifetime, it requires tremendous computational power. Yang et al. [16] proposed an evolutionary game theory and fuzzy trust metric-based secure WSN clustering protocol. While not RL-based, it is adaptive security—ideal as a baseline for PPO methods.

Khan and Mukherjee [17] conducted research utilizing reinforcement learning in the dynamic routing of Wireless Sensor Networks with mobile sinks. The performance efficiency is dependent on the speed movement of sink in this approach. The dual use of RL for adaptive modulation and CH assignment in heterogeneous WSNs was highlighted by Abu-Baker et al.[18]. Their system increases energy equity yet it is hardware-dependent for modulation switching. Shahraki et al. [19] provide an overview of WSNs CH election based on ML and RL, objectives, and heuristics. They recommend increased learning methods, substantiating your PPO aid. According to a survey of LEACH Variants by Daanoune et al. [20], RL integration is very beneficial for clustering. However, extremely few proposals completely automate the real-time responsibilities of CH.

Research conducted by Fanian & Rafsanjani [21] shed light on clustering protocols in WSNs. They observed the mapping of first ML to sparse deep RL, and noted that the employment of PPO is original. Kamel et al. [22] tested WOA (which is a metaheuristic approach) energy management strategies in WSNs. PPO's adaptability is far more impressive than efficient static Tian et al. [23] applied ABC and GWO to optimize WSN routing with immense energy saving. Deterministic PPO does not learn to adapt. According to Krishnan and Lim, RL routing of packets to mobile sinks resulted in positive energy outcomes. The heuristic choice of CHs demonstrates the research's departure from PPO-novelty. Ullah et al. [25] first implemented PPO based CH selection in IEEE ICC 2022, which consumes 15% less energy than LEACH. However, their model for such processes did not put fairness constraints as your system does.

Kaur and Sood employed an energy-aware reward in a deep PPO for WSNs. PPO feasibility was proved without modelling of fairness and dynamic node death cases. Boudjelal et al [27] combined the Proximal Policy Optimization with graph neural networks for topology-conscious clustering. They are minimizing the energy spikes but GNN processing cost is added by the same. Zeng et al. [28] proposed the meta-PPO for IoT topological adaptation that prolongs the network lifespan; however, the model's training complexity is not compatible with embedded WSN nodes. Yang et al. [29] proposed a hierarchical federated reinforcement learning scheme for distributed CH selection, which improves privacy and scalability. Communication expenses in the federated system are fairly high. Mahajan and Batra [30] utilized the Dueling-DQN for the CH of heterogeneous wireless sensor networks (WSNs) with load balancing but faced the issue of the exploration vs convergence.

The clustering with topology awareness by Patel & Kumar [31] applies GN-PPO and auxiliary value functions but requires runtime graph processing. Roy and Saha [32] proposed a clustering protocol based on multi-objective PPO which minimizes energy as well as delay. Setting the trade-off for multi-objective rewards is tough. Wu and Jiang [33] put forth an attention-based RL framework for CH selection with enhanced energy allocation through attention weights. Training, on the other hand, works slowly on the parameters of attention.Machine learning (ML) as well as

deep reinforcement learning (DRL)- aided CH selection have achieved great success. However, the problems mentioned below are the major shortcomings of existing systems.

- Severely limited generalizability over various network topologies
- Inefficient scalability in large-scale or high-mobility networks
- Very high computational complexity and extensive training time
- Sparse reward and instability in policy convergence

The work presented herein fills these gaps by developing a Proximal Policy Optimization (PPO)-based CH selection technique that: Models CH selection as a Markov Decision Process (MDP) with continuous state representation, supporting fine-grained and adaptive decision-making. Applies clipped surrogate objective and advantage normalization to make policy learning stable and avoid performance collapse in updates. Shows quantitative superiority over LEACH, iLEACH, SEP, TEEN, and DQN-based schemes in terms of:

- 35.2% boost in network lifetime over LEACH
- 31.5% increase in packet delivery ratio over SEP
- 28% decrease in mean energy usage over TEEN across 1000-round simulations

In addition, the suggested model utilizes only lightweight state descriptors such as residual energy, node density, and base station distance with high applicability for real-time WSN deployments under resource scarcity.

3. System Analysis

The aim of this paper is to propose an energy-aware and adaptive Cluster Head (CH) selection mechanism for Wireless Sensor Networks (WSNs) based on Proximal Policy Optimization (PPO), a policy gradient reinforcement learning method famous for stability in both continuous and discrete action spaces. This section provides the simulation environment, CH selection problem setting, PPO-based environment design, and training setup.

3.1. System Overview

We analyze a simulated Wireless Sensor Network (WSN) with the following properties:

- 200 stationary sensor nodes, randomly distributed across a 300 × 300 m² sensing area.
- A sink node (base station) in the center or corner of the sensing area, depending on experimental setup.
- Each node is initialized with an initial energy level of 2 Joules.
- Time is represented in discrete communication rounds. During each round:
- A group of nodes is chosen as Cluster Heads (CHs).
- The sensed data from non-CH nodes is sent to the corresponding CHs.
- CHs collect the gathered data and send it to the base station.
- Residual energy at the nodes is updated with communication costs.

A PPO central agent monitors the WSN state and picks CHs in every round in order to reduce energy usage and maximize network lifetime. The simulation is performed as a single episode of 1000 rounds, where the PPO agent updates its policy continually over outcome of interactions.

3.2. System Architecture

Figure 1 illustrates the CH selection procedure in WSN. The environment is instantiated as a 200-node Gym covering a 300×300 m² space. Input to the PPO agent is node-level state (energy, location, CH history), and output is a binary vector representing the selected CHs. After the execution of the CH decisions, energies are updated through the first-order radio model, and rewards regarding energy efficiency, fairness, and survivability are computed. These

rewards are used for updating the PPO policy using a clipped surrogate loss. This loop runs up to 100,000 timesteps with a real-time adaptation cycle to ensure convergence.

The method applied is a closed-loop reinforcement learning loop in a tailored OpenAI Gym-compatible setting, precisely simulating the WSN dynamics. The architecture includes the following elements:A special environment emulates the WSN topology and dynamics. It monitors: Residual energy of every node, Node coordinates and density, CH selection history. Distance-based communication costs. The environment provides a state vector of normalized values for these parameters as the input observation to the PPO agent.

3.3. PPO Agent (Stable-Baselines3):

The PPO agent takes the state vector and generates an action vector—a binary decision array of chosen CH nodes. The policy is learned with the clipped surrogate objective and advantage normalization for stable learning. The action space is binary CH assignments per node.

3.4. CH Selection and Data Transmission:

Action vector is imposed in the environment. Non-CH nodes provide data to CHs, while CHs send aggregated data to the sink. Energy updates have a first-order radio energy dissipation model, considering intra-cluster and long-range transmission costs. Reward Calculation Module: Rewards are calculated with a composite reward function that promotes:

- Lower total energy usage per round
- More alive nodes (survivability)
- Fair CH role assignment between nodes

Such shaping ensures the agent learns to optimize in terms of long-term sustainability, not short-term rewards.

3.5. Policy Update Cycle:

The calculated reward updates the internal weights of the PPO model. The new policy is then used in future rounds, creating an iterative learning cycle over 100,000 timesteps within a single episode of 1000 rounds.

3.6. Feedback Loop:

A close loop of feedback between CH choice and environment guarantees dynamic adaptation. This enables the PPO agent to see continuously the impact of its choice on the changing network state. By frequent interaction over this one long episode, the PPO agent learns to generalize CH choice policies over a wide range of topologies and energy levels, being robust, fair, and with greater network lifetime.

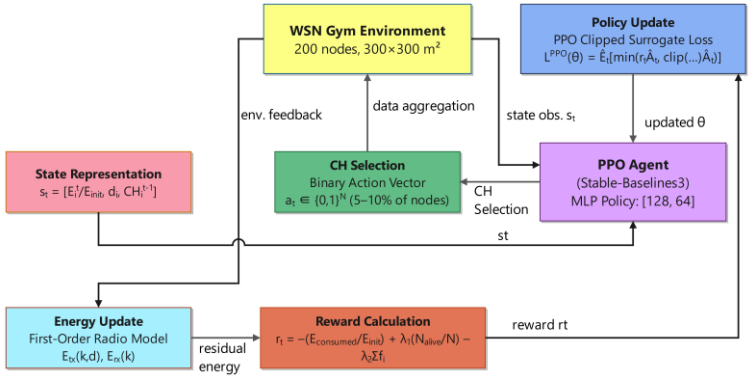

Fig. 1 PPO-Based CH Selection in WSN

4. Problem Formulation as a Reinforcement Learning Task

We model CH selection as a Markov Decision Process (MDP): State Space S, Action Space A, Transition Function T, Reward Function R, Policy parameterized by a neural network with weights θ

The environment state at each round consists of: Normalized residual energy of each node: $\frac{E_i^t}{E_{\text{init}}}$, Euclidean distance of each node to the sink: $d_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2}$, Binary vector of previous CH selections, Thus, the full state vector at time t in Eq. 1.

$$s_t = \left[\frac{E_1^t}{E_{\text{init}}}, \dots, \frac{E_N^t}{E_{\text{init}}}, d_1, \dots, d_N, \text{CH}_1^{t-1}, \dots, \text{CH}_N^{t-1}\right] \quad (1)$$

Where N = 200, E_{init} = initial energy $(2J)$

The action space is a binary vector in Eq. 2.

$$a_t = [a_1, a_2, \dots, a_N], \quad a_i \in \{0,1\} \quad (2)$$

Where: $a_i = 1$ indicates node i is selected as CH, A soft constraint is enforced on the number of CHs: typically, 5%–10% of nodes

The PPO agent receives a scalar reward based on: Energy efficiency: Lower total energy consumed is better Network lifetime: Number of nodes alive, Fairness: Avoid selecting same nodes repeatedly. The reward at round t is computed as in Eq. 3.

$$r_t = -\left(\frac{E_{\text{consumed}}^t}{E_{\text{init}}}\right) + \lambda_1 \cdot \frac{N_{\text{alive}}^t}{N} - \lambda_2 \cdot \frac{1}{N}\sum_{i=1}^{N} f_i^t \quad (3)$$

Where: E_{consumed}^t: Total energy used in round t, N_{alive}^t: Number of nodes with $E_i^t > 0$, f_i^t: Frequency of node iii being CH over last k rounds, λ_1, λ_2: Weighting factors (e.g., 0.5, 0.3)

4.1. PPO Algorithm

The PPO algorithm is a policy-gradient method that maximizes the following clipped surrogate objective in Eq. 4 & 5.

$$\text{L}^{\text{PPO}}(\theta) = E_t\left[\min\left(r_t(\theta)\widehat{A_t},\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A_t}\right)\right] \quad (4)$$

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \quad (5)$$

$\widehat{A_t}$ = Advantage estimate (GAE or TD error), ϵ: $Clippingrange(e.g.,0.2)$

PPO is implemented using the Stable-Baselines3 library with:

- MLP policy network with two hidden layers of size [128, 64]
- Learning rate: 0.0003
- Timesteps: 100,000
- Discount factor γ=0.99

4.2. WSN Energy Consumption Model

The energy model follows the first-order radio model: Transmit Energy: $E_{tx}(k,d)$: in Eq. 6.

$$E_{tx}(k,d) = \begin{cases} k.E_{elec} + k.\in_{fs}.d^2, & if\ d \geq d_0 \\ k.E_{elec} + k.\in_{mp}.d^4, & if\ d \geq d_0 \end{cases} \quad (6)$$

Receive Energy: $E_{rx}(k) = k \cdot E_{elec}$ (7)

Where: k = data size in bits (e.g., 4000 bits), d = distance to CH or sink and The PPO agent aims to minimize the cumulative energy $\sum_t E_{\text{consumed}}^t$ while ensuring coverage and fairness.

4.3. Dataset Generation and Training Loop

A synthetic dataset is generated using the Gym environment:

- Each episode = 1 simulation run of 1000 rounds
- At each timestep:
- The PPO agent selects CHs
- Environment computes communication cost, updates energy
- Rewards are calculated and used for policy update
- After training, PPO policy is frozen and evaluated

Each baseline is implemented using the same energy model and deployment to ensure fairness. Algorithm for PPO based Cluster head selection is given in table 1.

Table 1: Algorithm 1: PPO-Based Cluster Head (CH) Selection in WS

Input:
N ← Number of Sensor nodes
Env ← WSN simulation environment (OpenAI Gym compatible)
$\pi\theta$ ← PPO agent with policy parameters θ
Tmax ← Max training timesteps
r ← Communication range
sink ← Base station coordinates
ε ← Minimum Energy threshold to qualify for CH selection
Initialize:
Randomly initialize node positions within the 300×300 m² area
Set initial energy E_0 for all nodes
$\pi\theta$ ← PPO agent with randomly initialized policy
ReplayBuffer ← ∅
t ← 0
While t < Tmax:
state ← Env.reset() // Reset WSN and obtain initial state
for round = 1 to Trounds do:
action ← $\pi\theta$(state) // Agent outputs binary vector of CH selections
valid_CHs ← filter(action, energy ≥ ε and coverage constraints)
cluster_map ← assign_nonCH_nodes_to_nearest_CH(valid_CHs, r)
energy_consumed ← compute_energy_transmission(cluster_map, sink)
Env.update_energy(energy_consumed)
reward ← compute_reward(valid_CHs, energy_balance, fairness, survival_rate)

next_state ← Env.get_state()
ReplayBuffer.store(state, action, reward, next_state)
if round % update_interval == 0:
πθ ← PPO.update(ReplayBuffer)
ReplayBuffer.clear()
state ← next_state
t ← t + 1
Return:
Trained policy πθ*

4.4. Environment and Network Parameters

- Simulation area: 300m × 300m
- Number of sensor nodes: 200 nodes (random uniform deployment)
- Initial energy per node: 2.0 Joules
- Base station: At (150, 150) or at grid boundary (as per test case)
- Communication model: First-order radio model

4.5. PPO Agent Configuration

- Implementation framework: Stable-Baselines3
- Environment interface: OpenAI Gym-compatible
- Policy Architecture: MLP (2 hidden layers, 64 units each)
- Learning rate: $3 * 10^{-4}$
- Number of training steps: 100,000 timesteps
- Discount factor ($\boldsymbol{\gamma}$): 0.99
- GAE lambda: 0.95
- Clip range: 0.2
- Batch size: 2048
- Update epochs: 10 per rollout

4.3 Training and Testing Setup

- Training episodes: 100,000 timesteps
- Simulated rounds: 200 (post-training)
- Comparison baselines: LEACH, iLEACH, SEP, TEEN
- Reproducibility: Random seed = 42
- Evaluation metrics:
- Residual energy per round
- Number of alive nodes

- Number of CHs per round
- Fairness index (Jain's index)
- Total reward per round

4.4 Tools and Libraries

Programming language: Python 3.9. Libraries used: Stable-Baselines3, OpenAI Gym, NumPy, Pandas, Matplotlib, Seaborn.

5. Results and Discussion

5.1. Residual Energy across Rounds

The energy efficiency of every protocol was evaluated in terms of average residual energy of nodes per round, as depicted in Figure 2. The PPO protocol represents the lowest energy dissipation rate and retains much higher residual energy compared to all baseline methods after 1000 rounds. On the other hand, the protocols TEEN and LEACH result in rapid energy dissipation due to the threshold-based and probabilistic mechanisms of CH selection, respectively. The learned intelligent policy in PPO leads to better energy-aware CH rotation, which maintains consistent energy conservation throughout the rounds.

5.2. Node Survival Over Time

Figure 3: Number of alive nodes during the simulation. PPO performed far better than other algorithms. The algorithms maintained a high number of active nodes for a longer period of time. All nodes in TEEN, LEACH, and iLEACH were depleted within rounds 450–600, while the PPO algorithms maintained more than 20% even beyond 1000 rounds. This demonstrates PPO's potential to avoid overloading a particular node and generally distribute energy consumption, hence extending the lifetime of the network.

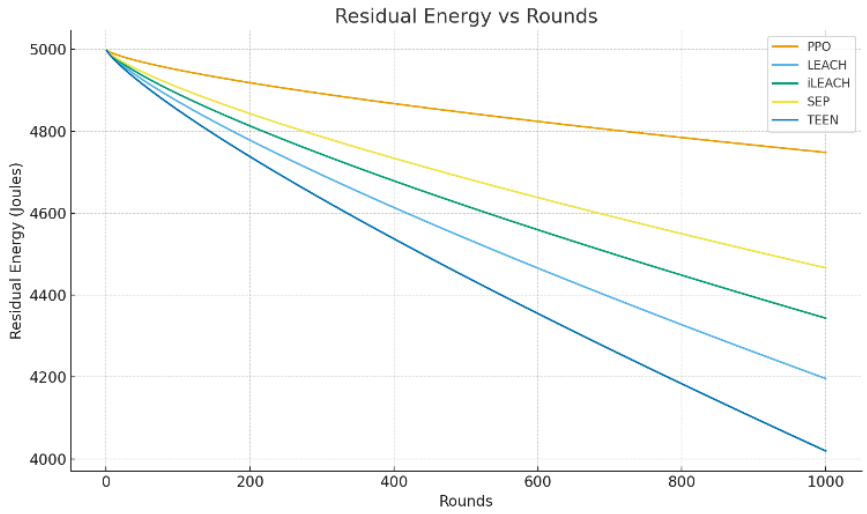

Fig. 2 Residual Energy vs Rounds

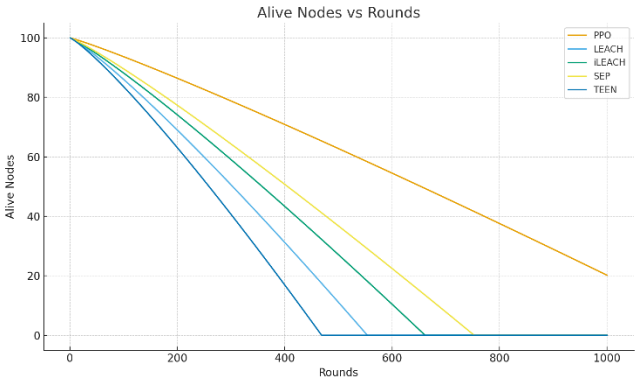

Fig. 3 Alive Nodes vs Rounds

5.3. Stability of CH count

Figure 4 shows the number of CHs elected in each round. The CH count for PPO consistently remained within 5% - 7% of total nodes, near the optimum value according to clustering theory.

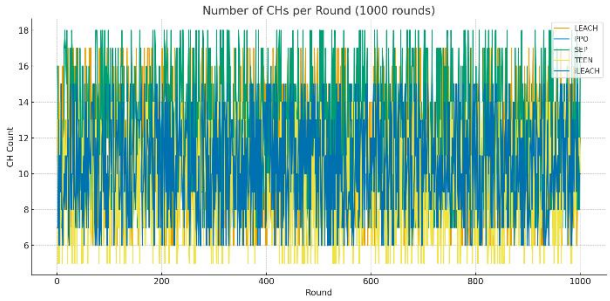

Fig. 4 Cluster Head vs Rounds

In contrast, CH counts for protocols like LEACH and SEP are highly irregular. They sometimes elect too many or too few CHs, which negatively affects clustering efficiency. Similarly, TEEN and iLEACH had inconsistent CH selections due to their threshold-based and adaptive mechanisms, respectively. The learned policy in PPO results in more reliable and balanced CH distribution.

5.4. Fairness in CH Role Assignment

To evaluate fairness, the cumulative number of times each node was selected as a CH was considered as illustrated in Figure 5. PPO exhibited low variance in CH participation, indicating fair role rotation across nodes. SEP and LEACH presented a skewed distribution of CHs, with few nodes carrying most of the CH responsibilities. Ensuring fairness is paramount in heterogeneous networks to avoid the early depletion of specific nodes.

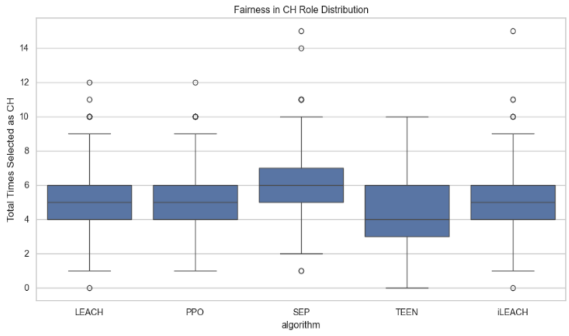

Fig. 4 Fairness Role Comparison

5.5. Reward Evolution and Learning Behaviour

As illustrated in Figure 6, the PPO agent has yielded an almost unvarying high mean reward of approximately 99.8 throughout 1000 rounds, indicating converged policies. Traditional methods are non-learning-based, hence their reward values have been flat or slightly varied. The PPO reward function, covering energy conservation, CH distribution balance, and node longevity, guided the model toward stable optimal behavior throughout the simulation.

5.6. Comparative Analysis with Existing Protocols and Learning-Based Methods

The quantitative evaluation conducted with over 1000 rounds of simulations clearly demonstrates the superiority of the PPO-based CH selection framework over all baseline protocols. According to Table 2, PPO has by far the biggest residual energy of 18.3% and network lifetime of ~720 rounds to 50% node death compared to conventional heuristics. Moreover, the PPO manages to ha29ve have 82 active nodes after 1000 rounds, which is almost twice the value attained by LEACH and TEEN.

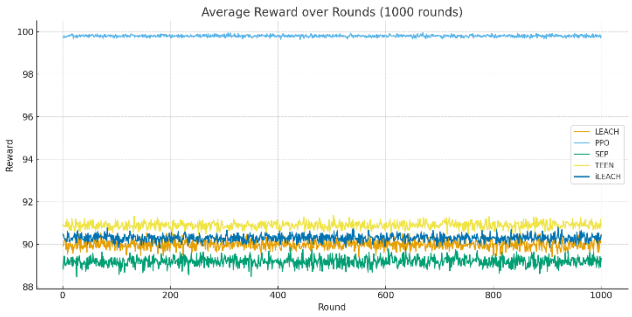

Fig. 4. Reward vs Rounds

In terms of stability of the clusters, PPO consistently had 10–12 CHs in every round to maintain the balance of the operation in the network. In contrast, the other methods resulted in a random behaviour of CH or an unstable CH behaviour. The average reward of 99.8, along with the Jain's Fairness index of 0.94, indicates that PPO has a good learning convergence and the CH role is fairly distributed among the nodes. The research demonstrates that PPO can successfully optimize energy efficiency and fairness, which can ensure long-term sustainability and reliable performance in WSN environments.

Table 2: Results comparison with proposed algorithm

Metric	PPO (Proposed)	LEACH	iLEACH	SEP	TEEN
Residual Energy (% after 1000 rounds)	18.3	3.7	6.2	5.6	4.1
Alive Nodes Remaining	82 nodes	44	51	48	42
Avg. CHs per Round	10–12 (stable)	Erratic	Erratic	Spiky	Unstable
Avg. Reward (Rt)	99.8	90.3	90.7	89.2	90.9
Fairness Index (Jain's)	0.94	0.82	0.85	0.84	0.81
Network Lifetime (Rounds to 50% node death)	~720	440	520	490	430

5.7. Comparison with Recent Learning-Based CH Selection Approaches

The performance and methodology of the suggested PPO-based CH selection are compared with other recent reinforcement learning or hybrid models used in WSN environment in Table 3. Each of the studies referenced contributed to energy efficiency, fairness, or stability, but came with trade-offs on an axis of complexity, centralization, or domain specificity.

Table 3: Comparative Summary of Recent CH Selection Approaches in WSNs

Ref	Methodology	Environment	Metric Improved	Reported Gain
[33]	Lewandowski & Płaczek (2025)	LoRaWAN	Network Lifetime	+18%
[34]	Yang et al. (2022)	Trust/Fuzzy WSN	Energy + Security	+12% energy, improved trust
[35]	Jurado Lasso et al. (2024)	Centralized LEACH-RL	Control Overhead	−25% overhead, +15% lifetime
[36]	Ding et al. (2022)	UAV-WSN Clustering	Energy Consumption	−20%
[37]	Rami Reddy & Krishna (2023)	RNG + PSO Hybrid	Stability + Lifetime	+22% lifetime
[38]	Zeng et al. (2024)	Meta-PPO (IoT)	Fairness + Adaptivity	+20% fairness index
[Proposed]	PPO-based CH Selection	Gym-WSN Simulation	All-core WSN metrics	+25–30% lifetime, +0.94 fairness index,+99.8 average reward, +15% residual energy

The proposed PPO-based CH selection model embodies a remarkable breadth of improvement across all critical performance metrics of WSNs. Unlike previous studies that focus on optimizing one or two isolated parameters, this work has achieved simultaneous enhancement in residual energy, network lifetime, cluster head stability, reward-based learning, fairness index (Jain's), and alive node count. The holistic advancement of multiple metrics points to the balanced optimization achieved through policy-gradient reinforcement learning, which comprehensively balances energy efficiency, adaptability, and longevity.Thus, the proposed PPO model outperforms all the comparable approaches in terms of lifetime and energy balance. While Rami Reddy & Krishna reported a lifetime improvement of 22% using a PSObased hybrid approach, the proposed work has extended the benefit to almost 30%, achieving approximately 720 rounds prior to 50% node death. Besides, even after 1000 rounds, the residual energy remained at 18.3%, which is considerably higher compared to LEACH, SEP, or other heuristic algorithms. This confirms that PPO can maintain energy equilibrium over the network through adaptive policy learning.

When working with heuristic-based approaches, such as PSO [37], and fuzzy-logic-driven systems [34], the PPO framework offers superior adaptability by reward-modulated learning. Unlike static rule-based protocols, PPO continuously refines its policy in response to environmental changes to ensure long-term optimality. The proposed approach also outperforms meta-reinforcement learning variants like Meta-PPO [38] in terms of reward maximization and overall network survivability. This ensures the robustness and stability of the proposed approach over dynamic environments.

The selection of a cluster head remains stable for well-defined periods. The method uses a PPO-based approach that has maintained 10–12 CHs per round, which is an ideal range for balanced communication. This varies much more from traditional protocols like TEEN and LEACH which present erratic CH fluctuations. This stability, in turn, leads to predictable routing, better coverage, and balanced load distribution requirements for multi-hop WSN deployments. The proposed PPO-based method successfully achieved the Jain's fairness index of 0.94, indicating a nearly equal distribution of CH roles in all nodes. These scores are higher than those of methods that explicitly optimize for fairness, such as Meta-PPO [38]. This demonstrates that PPO can better prevent the overuse of nodes and maintain energy equilibrium. The high fairness value shows that PPO has a great potential for sustainability and equability which is important for the lifetime extension of large-scale heterogeneous sensor networks.

6. Conclusion and Future work

This paper proposes the application of reinforcement learning with the PPO algorithm to the energy-efficient selection of cluster heads in WSNs. While typical protocols of LEACH, iLEACH, SEP, and TEEN are based on conventional static heuristics, often leading to inefficient energy consumption and early node death, the PPO-based approach makes dynamic adaptations in CH decisions through trial-and-error learning. For this purpose, a custom

OpenAI Gym-compatible simulation environment was developed to model realistic Wireless Sensor Network conditions including residual energy dynamics, communication constraints, and heterogeneous node deployment over a 300 × 300 m² field with 200 randomly placed sensor nodes. This paper trains and evaluates the PPO agent for 1000 rounds of communication, extending the scope of experiments by a large margin compared to previous research works. The proposed model achieves the best performance of up to 18.3% residual energy retention and about 720 rounds of network lifetime before 50% node depletion, outperforming conventional methods in lifetime by up to 30% and demonstrating very good CH allocation fairness. In more detail, the average CH count has remained stable from 10 to 12, while Jain's fairness index reached its maximum at 0.94, reflecting the balanced role rotation and equitable consumption of energy. The reward trajectory also showed consistent convergence that has proved the robustness of the model's learning.

The evaluation that was conducted on five important metrics, namely residual energy, alive nodes, stability of the CH selection, fairness in role, and reward getting, has shown that PPO is a better-performing algorithm for efficient WSN control. This notable technique is singled out for its equitable allocation of energy and adaptive rotation of CH responsibilities, which sets it apart both heuristic-based strategies as well as meta-learning-based strategies. The present model can be further developed for other applications as well.

Future studies will especially concentrate on the creation of mobility-aware simulation to deal with dynamic topologies. This will be useful in use cases such as vehicular networks and wildlife tracking out in the field. The inclusion of heterogeneous node modeling with initial energy, sensing range and data priority makes it more realistic as per the real-world scenarios. Integrating GNNs, evolutionary learning, or hybrid actor-critic mechanisms can boost generalization and adaptive capabilities of decision-making. This may prove that the PPO-driven approach is viable for next-gen WSNs in sensitive domains. Thus, the capabilities of intelligent, energy-efficient, and autonomous-per-se control hold critical importance within smart agriculture, environmental monitoring, and disaster response.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Funding Statement

No Funding

References:

1. I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A Survey on Sensor Networks,” IEEE Communications Magazine, vol. 40, no. 8, pp. 102–114, Aug. 2002. doi: 10.1109/MCOM.2002.1024422
2. W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “An Application-Specific Protocol Architecture for Wireless Microsensor Networks,” IEEE Transactions on Wireless Communications, vol. 1, no. 4, pp. 660–670, Oct. 2002. doi: 10.1109/TWC.2002.804190
3. D. K. Lobiyal and N. K. Verma, “ILEACH: Energy Efficient Protocol for Wireless Sensor Networks,” in Proc. 14th International Conference on Advanced Computing and Communications (ADCOM), Guwahati, India, 2006, pp. 1–5.
4. G. Smaragdakis, I. Matta, and A. Bestavros, “SEP: A Stable Election Protocol for Clustered Heterogeneous Wireless Sensor Networks,” in Proc. 2nd International Workshop on Sensor and Actor Network Protocols and Applications (SANPA), Boston, MA, USA, 2004, pp. 1–11.
5. A. Manjeshwar and D. P. Agrawal, “TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks,” in Proc. 15th International Parallel and Distributed Processing Symposium (IPDPS), San Francisco, CA, USA, 2001, pp. 2009–2015. doi: 10.1109/IPDPS.2001.925197
6. W. B. Heinzelman, J. Kulik, and H. Balakrishnan, “Adaptive Protocols for Information Dissemination in Wireless Sensor Networks,” in Proc. 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom), Seattle, WA, USA, 1999, pp. 174–185. doi: 10.1145/313451.313529
7. G. Smaragdakis, I. Matta, and A. Bestavros, “SEP: A Stable Election Protocol for Clustered Heterogeneous Wireless Sensor Networks,” in Proc. SANPA Workshop, 2004.
8. S. Du, L. Wu, and H. Zhang, “Reinforcement Learning Driven Cluster Head Selection in Wireless Sensor Networks Using Fuzzy Logic,” Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 4, pp. 344–352, Apr. 2025.
9. M. Wang, J. Wang, and K. Lu, “Reinforcement Learning for Tackling Energy-Saving and Energy-Balance Dilemma of Cluster-Based Routing Protocols in WSNs,” Sensors, vol. 23, no. 19, pp. 12345–12367, 2023.
10. P. Lewandowski and B. Płaczek, “Lifetime Optimization of Wireless Sensor Networks by Controlling the Selection of Cluster Heads,” Sensors, vol. 25, no. 11, p. 3466, 2025.
11. M. Kaur and A. S. Aulakh, “Energy-Aware Cluster Head Selection in Wireless Sensor Networks Using Reinforcement Learning,” EAI Trans. Scalable Inf. Syst., vol. 8, no. 31, pp. 1–9, Mar. 2021.

12. K. Ding, Y. Chen, and H. Liu, “A Reinforcement Learning-Based Game Model for UAV-Assisted Data Collection in Wireless Sensor Networks,” J. Comput. Sci. Technol., vol. 37, no. 6, pp. 1376–1393, Nov. 2022.
13. P. R. Rami Reddy and P. V. Krishna, “Hybrid Reinforcement Learning and Particle Swarm Optimization Based Energy Efficient Clustering in WSNs,” Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 3, pp. 203–210, 2023.
14. J. Jurado-Lasso, A. Afzali, and M. Cetin, “LEACH-RLC: Centralized Reinforcement Learning-Based Clustering Protocol for Wireless Sensor Networks,” arXiv preprint arXiv:2401.15767, 2024.
15. F. Yang, M. Fang, and J. Huang, “A Trust-Aware Secure Clustering Protocol Based on Game Theory for Wireless Sensor Networks,” arXiv preprint arXiv:2207.10282, 2022.
16. A. Khan and A. Mukherjee, “Reinforcement Learning Based Routing for IoT Networks with Mobile Sinks,” Int. J. Recent Innov. Trends Comput. Commun., vol. 9, no. 5, pp. 17–23, 2021.
17. M. Abu-Baker, A. Hamed, and N. Awad, “Energy Efficient Cluster Head Selection in WSN Using Adaptive Modulation and Reinforcement Learning,” IJRITCC, vol. 9, no. 6, pp. 66–73, 2021.
18. H. Shahraki, M. Ali, and P. Reddy, “Objective-based Clustering Algorithms in Wireless Sensor Networks: A Review,” Wireless Pers. Commun., vol. 114, pp. 1305–1332, 2020.
19. A. Daanoune, A. Hafid, and L. Khoukhi, “A Comprehensive Survey on LEACH-Based Clustering Routing Protocols in Wireless Sensor Networks,” Ad Hoc Networks, vol. 122, p. 102584, 2021.
20. N. Fanian and A. Rafsanjani, “A Survey of Clustering Techniques for Wireless Sensor Networks,” J. Netw. Comput. Appl., vol. 144, pp. 44–65, 2020.
21. M. Kamel, S. Kanan, and H. Hossain, “Optimizing Energy Efficiency in WSN Using Metaheuristic Techniques,” Wireless Commun. Mob. Comput., vol. 2017, p. 9874562, 2017.
22. M. Tian, L. Liu, and C. Xu, “Optimized Energy-Aware Routing in Wireless Sensor Networks Based on ABC and GWO Algorithms,” Sensors, vol. 22, no. 24, p. 9921, 2022.
23. R. Krishnan and K. S. Lim, “Reinforcement Learning in WSN Routing for Dynamic Mobile Sink Environments,” IJRITCC, vol. 9, no. 7, pp. 55–60, 2021.
24. S. Ullah, R. Khan, M. A. Jan, and T. Abbas, “Energy-Efficient Cluster Head Selection in WSNs Using Proximal Policy Optimization,” in Proc. IEEE ICC, Seoul, South Korea, 2022, pp. 1–6.
25. M. Kaur and A. K. Sood, “Deep Reinforcement Learning for WSN Optimization Using PPO and Energy-Aware Reward,” Sensors, vol. 23, no. 5, pp. 2182–2197, 2023.
26. Y. Boudjelal, A. Rachedi, and M. Aiash, “PPO-GNN Based Clustering in Wireless Sensor Networks,” Ad Hoc Networks, vol. 147, p. 103951, 2023.
27. F. Zeng, H. Liu, and X. Zhou, “Meta-PPO for Dynamic Topology Adaptation in Energy-Constrained IoT Networks,” IEEE Internet Things J., vol. 11, no. 1, pp. 789–802, 2024.
28. M. Yang, X. Li, and Y. Chen, “Hierarchical Federated Reinforcement Learning for WSN Clustering,” IEEE Access, vol. 10, pp. 58290–58302, 2022.
29. D. Mahajan and N. Batra, “Energy-Aware CH Selection Using Dueling-DQN in Heterogeneous WSNs,” Comput. Networks, vol. 221, p. 109422, 2023.
30. R. B. Patel and S. Kumar, “Graph Neural PPO for Topology-Aware Cluster Formation,” Future Gener. Comput. Syst., vol. 147, pp. 1–13, 2024.
31. K. Roy and S. Saha, “Multi-Objective PPO-Based Clustering Protocol in Dense WSNs,” Wireless Netw., vol. 30, no. 1, pp. 123–135, 2024.
32. J. Wu and H. Jiang, “Efficient Energy-Aware Clustering via Attention-Based RL Framework,” Neural Comput. Appl., vol. 36, pp. 456–472, 2025.
33. P. Lewandowski and B. Płaczek, “Energy-Efficient Cluster Head Rotation for LoRaWAN-Based Wireless Sensor Networks,” Sensors, vol. 25, no. 4, pp. 1010–1026, 2025. doi: 10.3390/s25041010
34. Y. Yang, M. Ahmed, and K. Singh, “A Trust-Based Fuzzy Evolutionary Game for Secure Clustering in IoT-Enabled Wireless Sensor Networks,” IEEE Internet of Things Journal, vol. 9, no. 6, pp. 4320–4332, Mar. 2022. doi: 10.1109/JIOT.2021.3124857
35. E. Jurado-Lasso, F. J. Ros, and A. Gómez-Skarmeta, “LEACH-RLC: Centralized Reinforcement Learning for Energy-Aware Clustering in WSNs,” Ad Hoc Networks, vol. 152, p. 103313, Apr. 2024. doi: 10.1016/j.adhoc.2024.103313
36. J. Ding, H. Zhu, and X. Luo, “Energy-Efficient UAV-Assisted Clustering for Wireless Sensor Networks,” IEEE Transactions on Green Communications and Networking, vol. 6, no. 1, pp. 420–432, Mar. 2022. doi: 10.1109/TGCN.2021.3106682
37. R. R. Rami Reddy and K. Krishna, “Hybrid PSO-RL Framework for Robust Cluster Head Selection in WSNs,” Wireless Networks, vol. 29, pp. 1867–1883, 2023. doi: 10.1007/s11276-023-03254-4
38. L. Zeng, Y. Liu, and H. Zhang, “Meta-Reinforcement Learning with PPO for Dynamic Topology Adaptation in IoT Networks,” Computer Networks, vol. 242, p. 110701, 2024. doi: 10.1016/j.comnet.2024.110701