Q-Learning Based Enhancement of DEECP for Prolonged Lifetime in Smart Mining Wireless Sensor Networks

Shubham Kumar
Department of Electronics and Communication
National Institute of Technology
Patna, India
shubhamk.pg24.ec@nitp.ac.in

Bharat Gupta
Department of Electronics and Communication
National Institute of Technology
Patna, India
bharat@nitp.ac.in

***Abstract*—The Wireless Sensor Networks (WSNs) are vital for underground mining applications, enabling continuous environmental monitoring and early hazard detection. Existing protocols, such as Distributed Energy Efficient Clustering Protocol (DEECP), have extended network lifetime by incorporating the multi-level node energy and node to sink distance into the Cluster Head (CH) selection process. However, these methods rely on static, threshold based parameters, limiting their adaptability in the highly dynamic and unpredictable underground environment. This paper presents a Q-Learning based enhancement to the DEECP protocol, introducing an intelligent and adaptive approach to CH selection in heterogeneous WSNs. Each sensor node is modeled as an autonomous agent capable of learning optimal CH election strategies through interaction with the environment, considering factors such as residual energy, sink proximity and temporal phase. The proposed reinforcement learning framework dynamically refines CH decisions over time, achieving extended network lifetime and improved energy distribution without imposing significant computational overhead on energy constrained sensor nodes. Simulation results confirm that the proposed method significantly outperforms both the original and the enhanced DEECP variant in terms of number of nodes alive and data throughput, demonstrating its effectiveness in energy constrained underground mining environments.**

Keywords—Q-learning, WSN DEECP, Energy Efficiency, Cluster Head Selection, Reinforcement Learning.

I. INTRODUCTION

Wireless sensor networks (WSN) serve as reliable technique to collect information from a wide range of distributed sensors, it acts as a crucial enabling technology for underground mining environments where continuous monitoring of hazardous conditions such as gas leakage, temperature rise and structural instability is essential for worker safety and operational efficiency. These networks typically consist of distributed sensor nodes with limited energy, computing power and communication capabilities which are often deployed in remote or hard to access areas to monitor environmental parameters and transmit the collected data to a central base station. Due to the harsh and inaccessible nature of underground environments, replacing or recharging sensor node batteries is often impractical, making energy efficiency a critical design consideration [1].

Traditional WSN protocols such as Low-Energy Adaptive Clustering Hierarchy (LEACH) [12] have introduced clustering mechanisms to reduce energy consumption. By rotating the role of the CH, which aggregates and forwards data to the base station, LEACH achieves some energy balancing. However, its random CH election and lack of residual energy consideration make it unsuitable for resource constrained and mission critical environments such as mines. To address the shortcomings of random CH selection, more sophisticated protocol, DEECP was proposed. DEECP considers the residual energy of nodes relative to the network average energy when electing CHs, thus enabling more informed decision-making and extending network lifetime. Despite its improvements over LEACH, DEECP in its original form suffers from limitations in heterogeneous networks where nodes have varying initial energy levels [12]. It also neglects the geographical distance between nodes and the base station, a critical factor in underground networks where communication range is limited and energy cost increases with distance [2].

To address this challenge, Enhanced Distributed Energy Efficient Clustering Protocol (E-DEECP) algorithm was developed and widely adopted [15]. The E-DEECP has shown notable improvements in prolonging network lifetime by considering factors such as residual energy and distance to the sink when selecting CHs. In this enhanced model, nodes are grouped into three energy levels Normal, High and Super energy nodes. These categories represent nodes with increasing energy capacities and are assigned different weights or probabilities during CH election. This classification ensures that high energy nodes has greater chances of become CHs, thereby balancing the energy load across the network more effectively [3]. Additionally, Euclidean distance formula is used to figure out node to sink (base station) proximity. This consideration is crucial because nodes located away from the sink loss compartively more energy during data transmission. The enhanced protocol penalizes such distant nodes when calculating CH suitability, thereby encouraging closer nodes with adequate energy to take on the CH role. Furthermore, an improved threshold equation is used for CH selection that integrates the proportion of residual energy, node type (on basis of energy level) and distance factor into a unified probabilistic model. This formulation significantly improved network stability period, throughput and lifetime when compared to baseline DEECP [4].

However, these improvements largely rely on static threshold-based heuristics, which lack the flexibility to adapt to the dynamic and unpredictable conditions typical of underground mining environments. Factors such as irregular

topology, fluctuating energy consumption patterns and environmental interference can significantly impact network behavior, making static thresholds insufficient for sustained performance and adaptability. As a result, there is a growing need for adaptive and learning-based methods that can respond intelligently to environmental changes and network dynamics over time [5].

To address these limitations, this paper proposes a novel Q-learning based enhancement to the DEECP protocol. Q-learning is a model-free Reinforcement Learning (RL) technique, allows nodes to learn from their environment over time and make adaptive CH selection decisions based on ongoing network conditions. In the proposed approach, each sensor node is modeled as an autonomous agent that observes local features such as distance to the sink, residual energy (low, medium, high) and temporal phase (early, mid, or late) to select actions (i.e., whether to become CH or not) that maximize long-term energy efficiency and network lifetime. The learning process is designed to be lightweight and decentralized, ensuring that computational and memory overheads remain minimal, thereby preserving the viability of deployment in energy-constrained underground environments. The reinforcement learning framework allows nodes to refine their CH selection strategy over time based on received rewards, which are formulated to reflect the quality of the CH decision in terms of energy preservation and contribution to network longevity.

Figure 1. Generic RL-based Decision-Making Framework

This dynamic learning capability introduces a level of adaptability and intelligence absent in conventional threshold based approaches, making the proposed method highly resilient and suitable for real-world underground mining scenarios. Extensive simulation results demonstrate that the Q-learning based DEECP significantly outperforms both the baseline DEECP and the Enhanced DEECP (E-DEECP) in terms of stability period, number of alive nodes and data throughput, marking a step forward in smart and energy-aware WSN protocol design [6].

Our proposed method leverages Q-learning, a model-free reinforcement learning (RL) technique that empowers individual nodes to learn optimal CH selection strategies through continuous interaction with their environment. Unlike supervised learning methods, Q-learning does not require prior knowledge or labeled datasets. Instead, each sensor node acts as an autonomous agent that learns by receiving rewards or penalties based on its actions [7]. Through repeated exploration and exploitation, each node updates its Q-values, which represent the expected long-term reward of taking a specific action in a given state. Over time, this learning process leads to more informed and energy-efficient CH selections that adapt to the network's evolving conditions. The key advantage of incorporating Q-learning is its ability to adapt in real-time, allowing the network to respond dynamically to changes such as node energy depletion, environmental interference, or topological changes caused by node failures. This level of adaptability is particularly critical in underground mining scenarios, where sensor nodes must operate autonomously in unpredictable and harsh conditions, without centralized control or global knowledge [8].

Our contributions can be summarized as follows:

i. We propose a Q-Learning based enhancement to the DEECP protocol for CH selection in heterogeneous WSNs deployed in underground mining environments.

ii. We design an agent based framework where each node autonomously learns optimal CH selection behavior using a combination of local energy, distance and temporal phase(rounds) information.

iii. We develop an efficient reward structure that encourages energy-efficient CH roles while penalizing actions that lead to early node deaths.

iv. We validate the proposed method through extensive simulations and compare it against both the baseline DEECP and E-DEECP variant. The results show significant improvements in key performance metrics such as the number of alive nodes, stability period and throughput.

The remainder of this paper is organized as follows:

Section II reviews related work on clustering protocols and reinforcement learning in WSNs. Section III provides the technical background. Section IV details the proposed Q-Learning-based cluster head (CH) selection mechanism. Section V presents and discusses the simulation results. Section VI summarizes the paper's conclusions and outlines potential future research directions.

II. Related Works

Efficient data communication and prolonged network lifetime have been long-standing challenges in Wireless Sensor Networks (WSNs), especially in harsh and dynamic underground environments. Numerous energy-aware clustering protocols have been proposed over the years to address the limited battery capacities and harsh deployment conditions of underground WSNs.

Traditional protocols such as LEACH introduced the idea of randomized rotation of CHs to evenly distribute energy consumption across nodes [12]. However, LEACH assumes homogeneous networks and often performs poorly in heterogeneous or underground conditions due to unpredictable signal attenuation and varying energy levels among nodes. To overcome these limitations, protocols like DEEC and its variants were developed. DEEC utilizes the residual energy of nodes and the average network energy to probabilistically select CHs. Despite improvements in energy efficiency, DEEC often relies on static parameters and fails to dynamically adapt to varying underground scenarios [8]. The E-DEECP protocol, improves upon DEECP by incorporating heterogeneity in energy levels and adjusting CH selection based on energy and distance to sink awareness. This variant improves CH probability using energy and distance based weighting. While these enhancements lead to noticeable gains in network lifetime and stability, the protocol still utilizes static heuristics for CH selection.

Recently, there has been growing interest in machine learning and reinforcement learning techniques for intelligent decision making in WSNs. Several studies have applied Q-learning to optimize routing and cluster formation, allowing nodes to learn optimal policies based on dynamic network conditions. Based on recent literature, these works can be broadly categorized into optimization based enhancements, fuzzy logic driven approaches, mobility-aware protocols and machine learning/reinforcement learning enabled schemes.

A. Optimization-Based Enhancements to LEACH

Several studies have applied evolutionary or swarm based optimization to improve LEACH's cluster head selection and energy management.

In [9], the authors introduced an "Optimized Energy Efficient Path Planning" scheme for agricultural IoT monitoring using the Stable Election Algorithm (SEA), which selects CHs based on residual energy, neighbor count and a heuristic rotation index. This was integrated with multi-objective evolutionary algorithms such as "Ant Colony Optimization (ACO), Genetic Algorithm (GA) & Simulated Annealing (SA)" [9]. Their approach achieved up to 66% improvement in energy efficiency as well as network lifetime compared to standard LEACH. However, the design was tested only in simulations with network sizes up to 100 nodes, leaving scalability in large-scale deployments unaddressed.

In [10], the authors optimized LEACH parameters via Particle Swarm Optimization (PSO), using residual energy thresholds as the key CH selection criterion. This achieved 25% energy efficiency improvement and 30% lifetime extension over LEACH in static WSNs. However, the scheme assumes static nodes and lacks mechanisms for mobility handling, limiting its applicability in dynamic scenarios.

B. Fuzzy Logic–Based Hybrid Approach

In [11], the authors combined fuzzy inference with PSO to enhance CH selection. Their fuzzy rules considered residual energy, node centrality and distance to the base station, improving adaptability in heterogeneous deployments. The hybrid fuzzy LEACH + PSO improved energy efficiency by 22% and extended lifetime by 18% compared to LEACH. Nevertheless, the fuzzy inference stage increased control overhead, which could negatively affect scalability in dense deployments.

C. Mobility-Aware Protocols

In [12], the authors addressed CH selection in mobile heterogeneous WSNs, incorporating residual energy and a mobility factor into a modified multi-hop LEACH protocol. This provided ~15% and ~20% improvements in energy efficiency and network lifetime, respectively. Despite these gains, the model assumes uniform mobility, which is rarely the case in real-world deployments and does not address large-scale topology changes.

D. Machine Learning / Reinforcement Learning

In [13], the authors leveraged Deep Reinforcement Learning (DRL) for UAV trajectory planning to minimize combined UAV flight energy and WSN communication energy. The CHs were selected to reduce overall system energy, with path optimization performed using a Pointer Network and A* search. While innovative, the approach assumes a pre-clustered WSN and does not explicitly quantify percentage improvements, making comparisons difficult. Additionally, latency and reliability aspects were not considered.

In [14], the authors proposed a hybrid DEEC + EEKA + Q-learning model for smart city and industrial IoT WSNs. The CH selection was based on DEEC's residual energy ratio and EEKA constraints, with K-means clustering and Q-learning enhancing adaptability. This yielded 17.5% energy efficiency and 14% lifetime improvements compared to LEACH. The main limitation is the computational overhead introduced by reinforcement learning and the lack of energy harvesting (EH) integration for sustainable operation.

In [15], the authors introduced the E-DEECP model for underground mining applications, in which sensor nodes are classified into three categories based on their energy levels and equipped with distance awareness capabilities. This approach achieved a significant 24% improvement in network lifetime. However, the model lacks adaptability to the highly dynamic environmental conditions typically found in underground mines.

E. Comparative Trends and Limitations

From the above works, several trends emerge:

- Optimization-based approaches (Al-Kaseem, Salman) achieve high performance gains but face scalability or mobility limitations.
- Fuzzy logic hybrids (Gamal) improve decision accuracy at the cost of increased control overhead.
- Mobility-aware methods (Mohapatra) address dynamic topologies but rely on unrealistic mobility assumptions.
- Environment-specific designs (Raed) excel in niche domains but lack generalization.
- ML/RL-based methods (Zhu, Aleem) promise adaptability and self-learning capabilities, but often suffer from computational complexity and incomplete performance evaluations.

Despite notable advancements, there is no existing approach that simultaneously:

- Adapts CH selection dynamically via self-learning mechanisms,
- Maintains computational efficiency suitable for large-scale deployments,
- Supports heterogeneous energy levels and mobility
- Is validated across both harsh and general-purpose WSN environments.

Due to high computational overhead, dependency on precise environmental sensing, lack of adaptability to dynamic network conditions and restricted real-world validation persist. Moreover, most existing ML-based methods employ static training without online adaptation, making them less effective in fluctuating topologies or heterogeneous environments. These limitations motivate our proposed approach, which leverages Q-learning-enhanced cluster head selection to balance energy efficiency, adaptability and real-time feasibility by addressing both the adaptability gap and the deployment practicality by integrating adaptive learning

into energy-efficient CH selection, while controlling complexity and preserving applicability in diverse WSN scenarios which is often missing in prior work. Moreover, most previous studies did not take into account fading effects or network failure scenarios.

TABLE 1. Summary of energy efficiency studies in WSN

PAPER & YEAR	CH SELECTION CRITERIA	OPTIMIZATION TECHNIQUE USED	ENERGY EFFICIENCY IMPROVEMENT (%)	NETWORK LIFETIME IMPROVEMENT (%)	LIMITATIONS (RESEARCH GAP)
"Al-Kaseem et al. (2021) Optimized Energy-Efficient Path Planning with Multiple Mobile Sinks" [9]	Stable Election Algorithm (SEA): Residual energy, neighbor count, heuristic rotation index	Multi-objective Evolutionary Algorithms (ACO, GA, SA)	Up to 66% vs. standard LEACH	Up to 66% vs. standard LEACH	Scalability issues beyond 100 nodes Simulation-only
"Zhu et al. (2021) UAV Trajectory Planning for Energy Minimization by Deep RL" [13]	CHs selected to minimize total UAV+WSN energy	Ptr-A* (Pointer Network + A* search)	Not explicit	Not explicit	Ignores latency/reliability Assumes pre-clustered WSN
"Mohapatra et al. (2022) Mobility Induced Multi-Hop LEACH Protocol in Heterogeneous Mobile Network" [12]	Residual energy, node mobility factor	Modified multi-hop LEACH	~15% vs. standard LEACH	~20% vs. standard LEACH	Assumes uniform mobility Limited scalability
"Salman et al. (2022) Optimization of LEACH Protocol for WSNs in Terms of Energy Efficiency and Network Lifetime" [10]	Residual energy threshold	Particle Swarm Optimization (PSO) tuned LEACH	25% vs. standard LEACH	30% vs. standard LEACH	Simulation only No mobility support
"Gamal et al. (2022) Enhancing Lifetime of WSNs Using Fuzzy Logic LEACH + PSO" [11]	Fuzzy rules: residual energy, node centrality, distance to BS	Fuzzy LEACH + PSO	22% vs. standard LEACH	18% vs. standard LEACH	Increased control overhead from fuzzy inference
"Raed et al. (2022) Efficient Energy Mechanism for Underground Mining Monitoring" [15]	Residual energy, link stability	Enhanced LEACH	20% vs standard LEACH	25% vs standard LEACH	Focused on harsh underground environment No mobility support
"Aleem & Thumma (2025) Hybrid DEEC + EEKA + RL in IoT-WSNs" [14]	Residual energy ratio (DEEC) + EEKA constraints	Hybrid: DEEC + EEKA + K-means + Q-learning	17.5% vs standard LEACH	14% vs standard LEACH	Computational overhead from RL No integration of energy harvesting or adaptive learning rate tuning.

III. TECHNICAL BACKGROUND

A. Wireless Sensor Networks and Energy Constraints

Wireless Sensor Networks (WSNs) consist of spatially distributed sensor nodes deployed to monitor physical or environmental parameters such as temperature, pressure, vibration or humidity [1]. Each node typically comprises sensing, data processingand communication modules, all powered by a limited-energy battery. Since replacing or recharging batteries is often impractical, energy efficiency becomes a primary design goal. Communication-related tasks, particularly data transmission to the base station (BS), are the dominant source of energy consumption in WSNs. Therefore, routing and clustering strategies play a critical role in prolonging network lifetime [15].

B. Clustering and Cluster Head Selection

Clustering divides sensor nodes into groups, each with a designated Cluster Head (CH) responsible for aggregating data from member nodes and forwarding it to the BS. This reduces redundant transmissions and saves energy. However, the CH role is energy-intensiveand improper CH selection can lead to premature energy depletion of certain nodes, reducing network stability and coverage. Effective CH selection algorithms aim to balance energy load and maximize the First Node Death (FND), Half Node Death (HND)and Last Node Death (LND) metrics [15].

C. Reinforcement Learning (RL)

Reinforcement Learning (RL) is a subfield of machine learning where an agent learns to make a sequence of decisions by interacting with its environment [7]. Unlike supervised learning, which depends on labeled input-output examples, RL operates on a trial-and-error basis, guided by feedback in the form of rewards or penalties [17]. RL has found effective applications in wireless sensor networks (WSNs) for tasks such as energy-efficient routing, cluster head (CH) selection, adaptive transmission power controland dynamic scheduling, due to the stochastic, dynamicand partially observable nature of these networks.

The RL framework is commonly modeled as a Markov Decision Process (MDP), characterized by the tuple <S, A, P, R, γ>. Here, S denotes the set of possible states describing the environment, A represents the set of actions available to the agent, P defines the state transition probabilities P(s′ | s, a), R is the reward function R(s, a) that provides scalar feedbackand γ (with values between 0 and 1) is the discount factor that balances the importance of immediate versus future rewards. The learning cycle in RL involves the agent observing the current state st, selecting an action at according to a policy π(s), experiencing a transition to the next state st+1 and receiving a reward rt. Based on this feedback, the agent updates its policy with the aim of discovering an optimal policy π^* that maximizes the expected cumulative discounted reward [16].

In the context of WSNs, RL plays a critical role in making adaptive decisions under resource limitations. For example, the problem of selecting a cluster head can be formulated as a decision-making process where the state captures factors like node energy levels, positions and network density. The action involves designating a node as the cluster headand the reward function is often tied to metrics such as energy consumption efficiency, network longevityand the success rate of packet delivery.

D. Q-Learning

Q-Learning is a model-free RL algorithm that does not require prior knowledge of the environment's dynamics P(s′ | s,a). Instead, it learns an action-value function Q(s,a), which estimates the expected return (cumulative discounted reward) when taking action 'a' in state 's' and thereafter following the optimal policy [18]. The Q-value update rule is given by

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

where s represents previous state, s ′ represents new state 'a' represents chosen action, 'α' represents learning rate, 'γ' represents discount factor and Q(s, a) is a two-dimensional lookup table, row contains all possible states and column contains all possible actions.

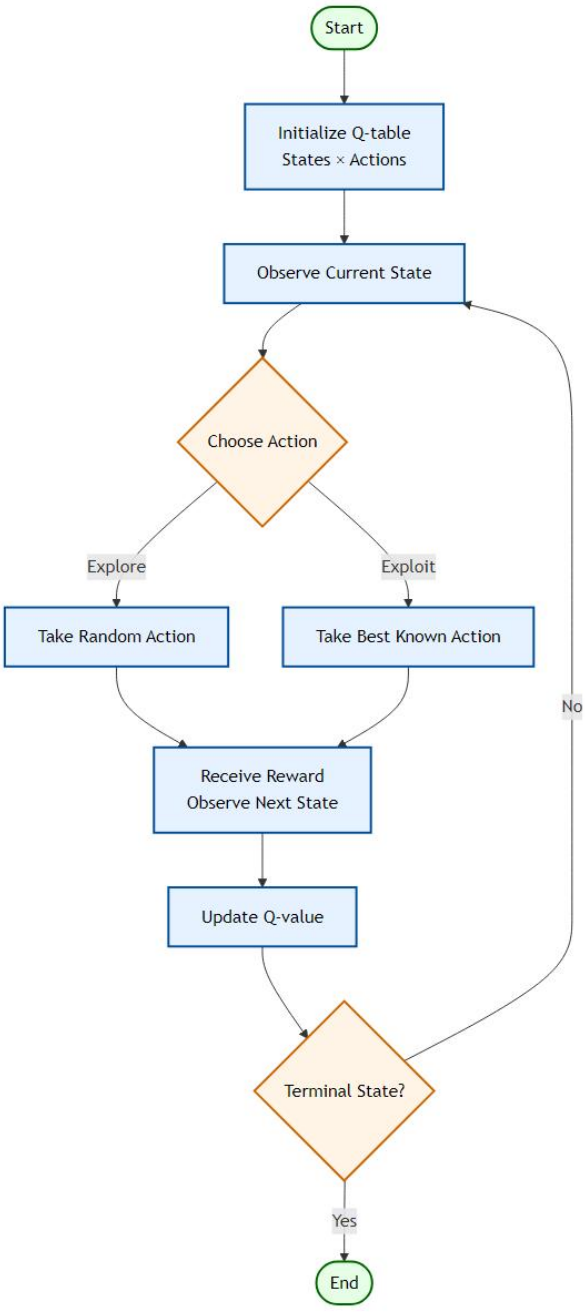

Figure 2. Workflow of Q-learning (RL) algorithm

Q-Learning uses an ε-greedy strategy to maintain a balance between exploration experimenting with less frequently chosen actions to find better results and exploitation actions that are already known to provide high rewards based on the agent's existing knowledge [18]. The ε parameter controls randomness in action selection, often decaying over time to gradually move from exploration to exploitation. This approach is has advantages for WSNs, in the context of WSN cluster head selection the 'state' can include residual energy, distance to base stationand current round status. The 'actions' to become CH or remain a normal node. The 'reward' is designed to favor balanced energy consumption, survival of CH and extended network lifetime. Q-Learning's model-free nature allows it to adapt to dynamic network conditions (e.g., varying node energy levels, mobilityand environmental interference) without predefined complex mathematical models of the network [18].

E. Clustering and Energy Models for Heterogeneous WSNs

LEACH Protocol is one of the oldest and most popular hierarchical routing protocols designed to reduce energy consumption in wireless sensor networks. Here nodes are organized into clusters, each with a CH and CHs are chosen randomly to balance energy load. Also LEACH uses TDMA to avoid intra-cluster collisions and CDMA for inter-cluster communication. Each node becomes a CH using a threshold.

$$T(n) = \begin{cases} \frac{p}{1-p\cdot\left(r \bmod \left(\frac{1}{p}\right)\right)}, & \text{if } n \in G \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where T(n) is threshold for node n to become CH, p is desired percentage of CHs per round, r is current round number, G is set of nodes that have not been CHs in the last 1/p rounds [1].

Whereas DEECP enhances the lifetime and stability of WSNs by making CH selection energy-aware, by introducing residual energy and average energy of the network as part of the CH selection logic. It also take in account that netwok can be heterogeneous with different energy level such as a normal nodes (basic energy) and advanced nodes (more energy). Each node becomes a CH using a threshold which is modified version of LEACH [15].

$$T(i) = \begin{cases} \frac{P_i}{1-P_i\cdot(r \bmod \left(\frac{1}{P_i}\right))}, & \text{if } i \in G \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where the T(n) is threshold for node n to become CH, Pi is probability of node i becoming a CH, r is current simulation round number and G is set of nodes that have not been CHs in the last 1/Pi rounds [15].

$$P_i = prop \times \left(\frac{E_i}{\bar{E}(r)} \right) \quad (3)$$

here the Pi is proportinal to the Ei which represents residual energy of node i and E(r) which represent average residual energy of the network at round r, 'prop' is optimal CH probability.

$$\bar{E}(r) = \frac{E_{\text{total}} \cdot \left(1 - \frac{r}{R}\right)}{N} \quad (4)$$

$$E_{\text{total}} = N \cdot (1-h) \cdot E_0 + N \cdot h \cdot E_0 \cdot (1+\alpha) \quad (5)$$

where E(r) represents average residual energy of the network at round r, N represents total numbers of nodes, R represents total numbers of rounds, Etotal represents total energy of the network, h represents faction of advance node and α represents heterogeneity factor that defines how much more energy an advanced node has compared to a normal node which is having initial energy Eo.

Whereas Enhanced DEECP (E-DEECP) introduces multi-level node heterogeneity that is, super, advanced and normal nodes with three different energy levels, also adds distance awareness to further improve CH selection efficiency. In E-DEECP CH selection algorithm uses weighted probability function considering, residual energy of the node, average residual energy of the network, weight factor based on node type and distance between node and base station [15].

$$P_i = \frac{\text{prop} \cdot E_i}{(1 + h(\alpha + m \cdot \beta)) \cdot E(r)} \quad (6)$$

Probability of node i becoming a CH, computed based on its residual energy relative to the network average energy, adjusted for heterogeneity factors.

$$\begin{aligned} E_{\text{total}} = N \cdot E_0 \cdot (1 - m - h) \cdot W_0 \\ + m \cdot W_1 + h \cdot W_2 \end{aligned} \quad (7)$$

$$T(i) = \begin{cases} \frac{P_i}{d\cdot\left(1-P_i\cdot\left(r \bmod \left(\frac{1}{P_i}\right)\right)\right)}, & \text{if } i \in G \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where the equation 8, represents threshold for node i to become CH, Pi is probability of node i becoming a CH and r is current round number. Here the threshold computation also incorporates the distance parameter, enabling distance-aware CH selection that favors nodes optimally positioned relative to the base station.

$$d = \frac{1}{|S|} \sum_{i \in S} \sqrt{(x_i - x_{BS})^2 + (y_i - y_{BS})^2} \quad (9)$$

$$W_i = \begin{cases} W_o = 1, & \text{node } i \text{ is a Normal node} \\ W_1 = 1 + \alpha, & \text{node } i \text{ is an Advanced node} \\ W_2 = 1 + \beta, & \text{node } i \text{ is a Super node} \end{cases} \quad (10)$$

where β represents heterogeneity factor that defines how much more energy an super node has compared to a normal node, m represents faction of advance node, d represents average distance from node 'i' to the base station, calculated by Euclidean distance formula [15].

The lifetime of a network is directly influenced by how efficiently each node consumes energy during

communication, computationand sensing operations. Among these, data transmission (Tx) and data reception (Rx) are the most energy-intensive activities, especially in clustered architectures where CH must forward aggregated data to the base station (BS). Therefore, it is essential to model and quantify the energy required for both transmitting and receiving data packets.

$$E_{Tx}(k,d) = \begin{cases} k \cdot E_{elec} + k \cdot \varepsilon_{fs} \cdot d^2, & \text{if } d < d_0 \\ k \cdot E_{elec} + k \cdot \varepsilon_{mp} \cdot d^4, & \text{if } d \geq d_0 \end{cases} \quad (11)$$

The equation 11, computes the energy needed to transmit a k-bit packet over a distance d, Longer distances (≥ d0) incur significantly higher energy costs due to multi path fading, where k is size of the data packet in bits, E_elec is energy required per bit for transmission/reception circuitry, ε_fs is amplifier energy for free-space model, ε_mp is amplifier energy for multi-path model, d is distance between sender and receiver, d_0 is threshold distance to switch between free-space and multi path,

$$E_{Rx}(k) = k \cdot E_{elec} \quad (12)$$

The equation 12, calculates the energy required to receive k bits, receiving costs are independent of distance, as no amplification is needed.

$$E_{DA}(k) = k \cdot E_{DA} \quad (13)$$

Here in the equation 13, E_DA is energy required for data aggregation for k bits before transmission.

$$E_{CH} = \left(\frac{N}{k} - 1\right) kE_{elec} + N \cdot E_{DA} + kE_{elec} + k\varepsilon_{fs} d_{BS}^2 \quad (14)$$

$$E_{nonCH} = k \cdot E_{elec} + k \cdot \varepsilon_{fs} \cdot d_{CH}^2 \quad (15)$$

The equation 14 and 15 calculates the total enegy consume by the CHs and CMs, when considering no multi path fading.

These energy estimations are essential for accurately modeling the energy dynamics of heterogeneous WSNs. By quantifying the per-round consumption for both Cluster Heads (CHs) and Cluster Members (CMs), it becomes possible to predict network lifetime, identify potential energy bottlenecksand design optimal CH rotation strategies.

In clustering-based protocols, CH nodes consume significantly more energy than CMs due to additional tasks such as data aggregation, long-distance transmission to the base stationand intra-cluster communication management. Therefore, CH selection algorithms often incorporate node distance to the sink, residual energy and temporal phase to balance the load across the network and delay the first node death (FND).

IV. THE METHODOLOGY

The proposed Q-Learning enhanced DEECP (QL-DEECP) integrates reinforcement learning into the CH selection process to enable adaptive, energy-aware decisions in a heterogeneous WSN. This design replaces static threshold-based heuristics with a decentralized, node-level learning mechanism, making it highly suited for the dynamic and unpredictable conditions of underground mining

The state space will discretize energy, distance and round parameters into total of 27 unique combinations, while the action space will consist of two choices that is becoming a CH or not. The reward function will be designed to promote the selection of high energy, sink proximal CHs that survive the round, penalize premature CH deathsand modestly reward surviving non-CH nodes to encourage balanced energy consumption [19]. An ε-greedy policy will balance exploration and exploitation, enabling each node to refine its CH selection strategy over time. The proposed approach will be evaluated in the same simulation environment as Baseline DEECP and E-DEECP, under multiple network sizes, ensuring fair comparison through identical clustering and communication models. Performance will be measured in terms of stability period, network lifetimeand throughput, with the expectation that QL-DEECP will demonstrate superior adaptability and energy efficiency in dynamic underground mining scenarios. In the proposed QL-DEECP framework, the learning process of each node is governed by the standard Q-learning update rule, expressed in the equation 16.

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \quad (16)$$

In this context Q(s,a) denotes the action-value function, which predicts the expected total reward an agent can achieve by taking action a in state s and then following the optimal policy afterward. The learning rate α, which ranges between 0 and 1, controls how much the newly gained information influences the existing value estimates, balancing the trade-off between learning speed and stability. The immediate reward received after performing action ‘a’ is represented by r, while the discount factor γ(gamma), also between 0 and 1, determines the relative importance of future rewards as compared to immediate ones [7]. The term Q(s′,a′) represents the maximum estimated value of the next state s′ over all possible actions a′, which helps the agent make decisions that maximize long-term returns. The difference between the target value [r+γmaxQ(s′,a′)] and the current estimate Q(s,a) is referred to as the Temporal Difference (TD) error [7] which quantifies the discrepancy between predicted and actual outcomes. By iteratively applying this update after every round of operation, each node refines its action-value estimates, allowing it to improve CH selection decisions over time.

Over multiple rounds, the Q-values converge toward an optimal policy that balances the trade-off between becoming a CH and preserving energy for future participation. This adaptive decision-making ensures that high-energy nodes close to the sink are utilized more effectively, while low-energy or distant nodes are preserved for later stages of the network lifetime. As a result, QL-DEECP is expected to prolong the stability period, delay the last node death, increase overall throughputand maintain better coverage compared to both DEECP variants.

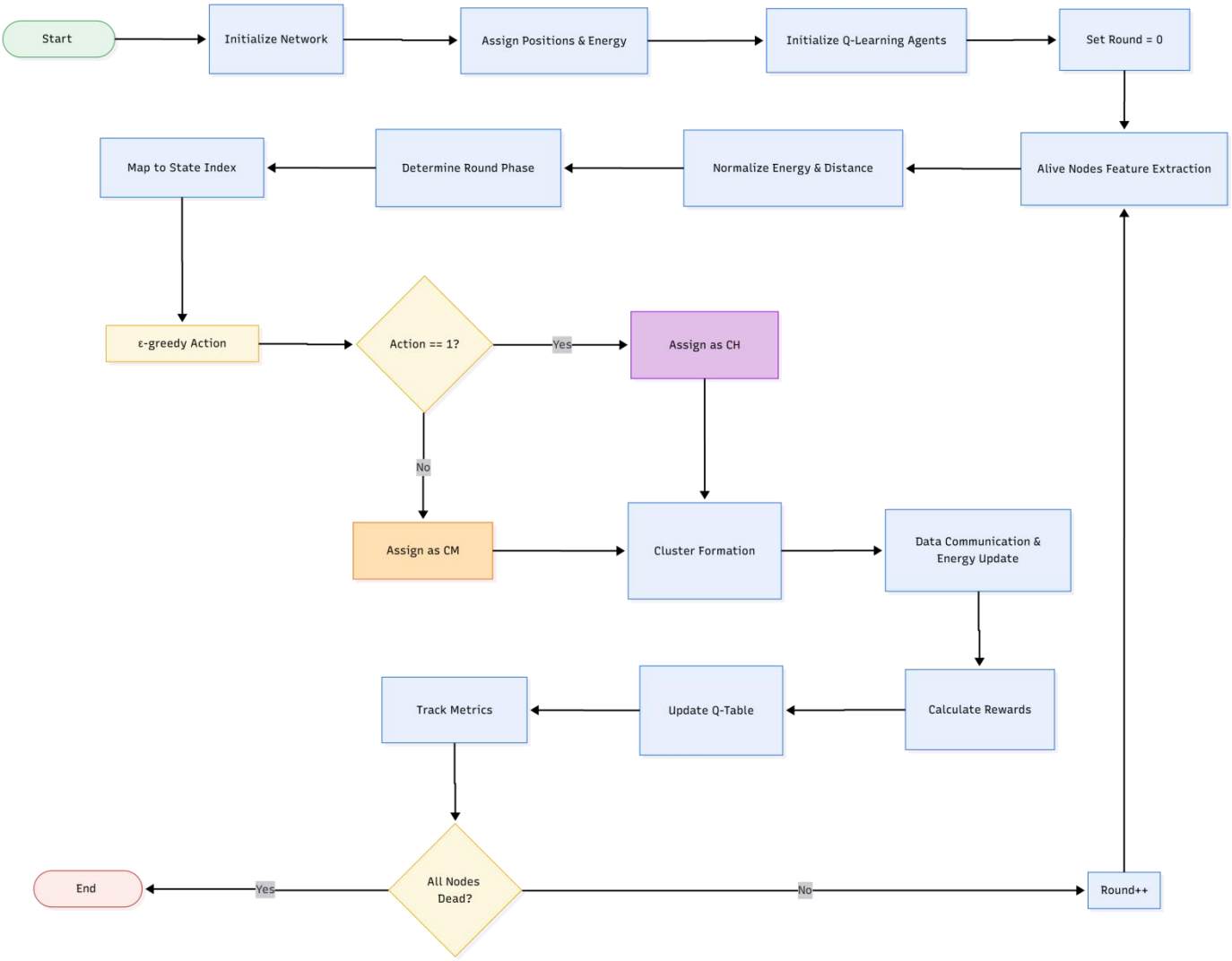

Figure 3. Flow diagram of the proposed QL-DEECP algorithm

The operational flow of the proposed QL-DEECP protocol, beginning with network initialization and node parameter assignment. Each node extracts features such as residual energy and distance to the sink, which are normalized and mapped to a discrete state index. Using an ε-greedy policy, nodes decide whether to act as a CH or a Cluster Member (CM). CHs form clusters and coordinate data communication, after which rewards are calculated and Q-tables are updated based on the Q-learning formula. The process repeats for subsequent rounds until all nodes deplete their energy [18] [19].

A. Optimization Targets

Q-DEECP specifically aims to reduce premature node death by preventing low-energy nodes from becoming CHs, balance energy utilization across all nodes to delay the first node death (FND) and last node death (LND) [11], enhance throughput by maintaining stable CH coverage over more rounds, increase adaptability in dynamic underground conditions where communication links and node energy vary unpredictably and decentralize decision-making, eliminating reliance on a global controller for CH assignment [20].

B. Working of QL-DEECP

Step 1 – Network Initialization: At deployment, all nodes are randomly positioned within the monitoring area. Each node is assigned an initial energy level depending on its type (Normal, High-energy, or Super node). Q-Learning agents are instantiated for each node, with Q-tables initialized to zero.

Step 2 – State Representation: Each node's state is determined by three key features that is Normalized Residual Energy (Enorm) that is current energy as a fraction of the maximum possible energy for that node type. Normalized Distance to Sink (Dnorm) that is Euclidean distance to the sink normalized by the network's diagonal length and temporal phase that is categorical representation of the simulation phase, early, mid or late rounds. These features are discretized and mapped to a finite state index for the Q-Learning agent.

Step 3 – Action Space: Each node can take one of two actions in every round. Action 0 act as a Cluster Member (CM) and Action 1 act as a Cluster Head (CH).

Step 4 – Action Selection (ε-Greedy Policy): Where nodes follow an ε-greedy policy, with probability ϵ, select a random action (exploration) and with probability 1- ϵ, select the action with the highest Q-value (exploitation) [19]. This allows nodes to explore potential new strategies while gradually converging toward optimal CH selection behavior in dynamically changing mine environment without the need of heavy computation .

Step 5 – Cluster Formation: Where all nodes that chose the CH action broadcast their status. CM nodes join the nearest CH based on Euclidean distance, forming non-overlapping clusters.

Step 6 – Data Transmission and Energy Update: CM nodes send their sensed data to the CH. CHs perform data aggregation and transmit the aggregated data to the sink node. Energy consumption is computed using the radio energy dissipation model for both transmission and reception which is guided by the equation 11, 12 and 13.

Step 7 – Reward Calculation: Each node receives a reward based on its performance during the round. High positive reward if a CH survives the round with significant residual energy and is close to the sink. Smaller positive reward for CMs that survive. Strong negative reward for CHs that die during the round (energy depletion).

Step 8 – Q-Table Update: Using the observed reward and the next state, the node updates its Q-table according to the equation 16. Then these process repeats for each round until all nodes are dead. Throughout the simulation, metrics such as First Node Death (FND), Last Node Death (LND)and Cumulative Throughput are recorded for performance evaluation.

Algorithm 1: QL-DEECP – Q-Learning–Based Distributed Energy-Efficient Clustering Protocol

Initialize

Network parameters, Node positionsand Energy levels.

Initialize Q-learning agents with zeroed Q-tables.

Set number of round (r) to 0.

While (any node is alive):

Identify alive nodes.

For (each alive node):

Extract features: normalized residual energy, normalized distance to sinkand temporal phase.

Map features to a discrete state index.

Select action using ϵ-greedy policy:

Generate a random number rnum between 0 and 1

If rnum < ε: *action = random choice between 0 or 1 # explore*

Else: *action = action with max Q value for current state # exploit*

If action = 1: *Assign node as Cluster Head (CH).*

Else: *Assign node as Cluster Member (CM).*

Form clusters by assigning CMs to nearest CHs.

Communication phase: CMs send data to CHs; CHs aggregate and forward to sink.

Update node energies based on transmission/reception costs.

End For

For(each alive node):

Calculate reward:

If CH and survived*: 0.8+ 0.2 * Enorm + 0.1 · (1−Dnorm) + 0.2 * AliveRatio*

If CH and dead: *−1.0*

If CM and survived: *0.1 * Enorm + 0.1 * AliveRatio*

Else: *0.0*

Update Q-table: As per equation 16

End For

Increment r by 1.

End while.

QL-DEECP protocol assumes that the base station is located at or near the center of the sensor network. This central placement is crucial because it helps minimize the average communication distance between nodes and the base station, which directly impacts the energy consumption of the sensor nodes.

C. *Evaluation Parameters*

The wireless sensor network was simulated in a PYTHON environment over a 200 × 200 underground mine area. The simulation parameters are listed in Table 2. Three network scales were considered, with 50, 100and 150 nodes randomly deployed within the defined area. The base station is positioned at the center of the mine. The communication and clustering among member nodes are conducted over limited ranges, specifically 2 to 4 meters for miner movement and 6 to 11 meters for roof fall detection.The energy model and transmission and reception are considered under low-power wireless conditions, following the IEEE 802.15.4 that is Zigbee standard.

Performance was evaluated using three key metrics network lifetime, number of alive nodesand network throughput. Network lifetime was defined as the total operational duration before complete energy depletion of all nodes [15]. The number of alive nodes was tracked throughout the simulation to assess energy efficiency and fault tolerance. Network throughput was measured as the total number of packets successfully transmitted to the base station. This

Table 2. WSN parameters for underground WSNs [15].

Network Parameters	Value
WSN area size	200 x 200
Number of nodes	50, 100, 150
Initial energy	0.5 J
Packet size	3000 bits
Data aggregation energy	5n J/bit/signal
Transmit energy	50n J/bit
Receive energy	50n J/bit
Free space loss	10n J/bit
Multipath loss	0.0013p J/bit
Simulation time	12000 sec
h = fractions of high energy node (eq. 7)	0.5
m = fraction of super energy node (eq. 7)	0.4
α heterogeneity factor (equation 10)	1.5
β heterogeneity factor equation 10)	3

The simulation will evaluate WSN performance by tracking the number of dead and active nodes, measuring power consumption, monitoring data transmission to and from the base station and assessing the overall network lifetime.

V. RESULTS AND DISCUSSION

This section presents and discusses the performance evaluation results of the DEECP variants and the proposed mechanism. The analysis focuses on the network stability period, the duration during which power consumption remains steady and energy distribution across the network is balanced, which is a primary objective for the evaluated protocols. The simulations assume that all nodes are either stationary or exhibit only micro-mobilityand energy losses due to dynamic random channel conditions and fading effects are neglected.

Figure 4 illustrates the number of operational nodes (total of 50 nodes) versus simulation rounds for all three DEECP protocol. Baseline DEECP exhibited a rapid decline in the number of alive nodes shortly after the stability period ended, caused by uneven energy consumption among nodes. E-DEECP improved stability by incorporating energy and distance awareness into the CH selection process, delaying the first node death and maintaining higher node availability for a longer duration. The proposed QL-DEECP further extended the stability period and sustained a higher count of alive nodes throughout the simulation. This improvement stems from its adaptive CH selection, which learns to balance energy usage dynamically and avoid overburdening individual nodes [21].

For 50 nodes, first node death (FND) occurred at round 508 for QL-DEECP, compared to 315 for E-DEECP. While the baseline recorded a higher FND of 1203 in this case, it suffered from much faster node death afterwards, quickly leading to network collapse. In contrast, Q-Learning maintained gradual, uniform node death over time.

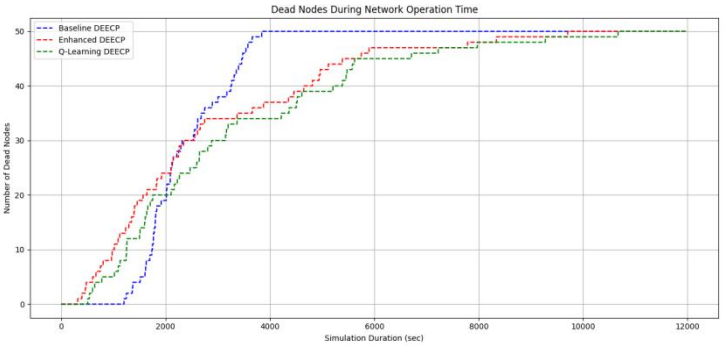

Figure 4. Number of dead nodes during network operation.

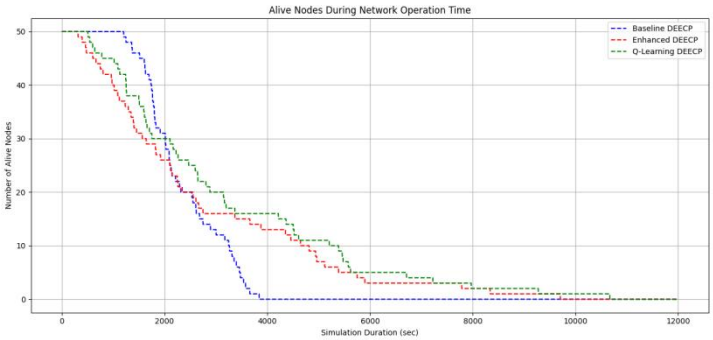

Figure 5. Number of alive nodes during network operation

Figure 5, which reflects the Last Node Death (LND) or total network lifetime, QL-DEECP significantly outperforms the other two methods by surviving until round 10,669, compared to 9,705 for E-DEECP and only 3,842 for the baseline. This represents an improvement of ~10% over E-DEECP and ~178% over the baseline. The extended lifetime results directly from QL-DEECP's adaptive CH selection mechanism, where each node learns from its past role assignments and the network's current energy state.

Q-Learning's strength lies in post-FND longevity, keeping far more nodes alive in the mid-to-late stages, where baseline protocols deteriorate rapidly.

Figure 6, clearly show that QL-DEECP achieves the highest cumulative throughput across all simulation rounds. All protocols show an upward trend in network performance during the initial phase of operation. DEECP maintains stability until around 4000 roundsand E-DEECP sustains this stable performance until approximately 9000 rounds, whereas QL-DEECP perform better by getting stability at 10000 rounds. In terms of peak performance, DEECP reaches a maximum throughput of 70 kbps and E-DEECP

more than doubles this by achieving 160 kbps, while in other hand QL-DEECP achieve 190 kbps, this clearly demonstrates the effectiveness of the QL-DEECP approach and validates the underlying design strategy. The results indicate that, under the QL-DEECP approach, sensor nodes stay active for longer periods when they are linked to a CH and possess sufficient energy for communication. Furthermore, the method ensures that CHs are selected based on having adequate residual energy, enabling them to remain operational and reliably transmit aggregated data to the base station.

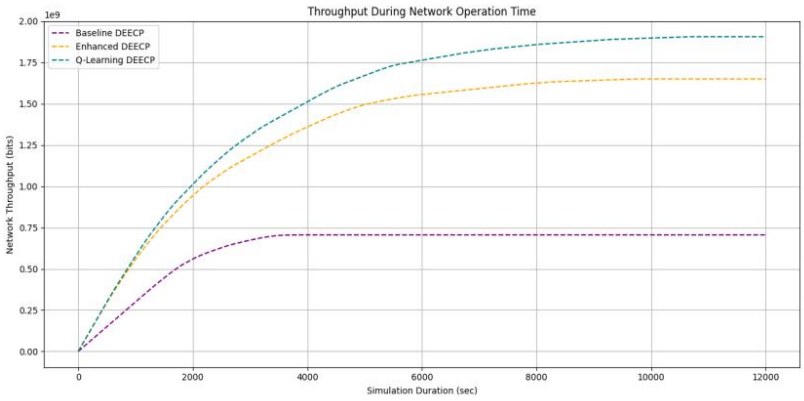

Figure 6. WSN network throughput for network size 50 nodes

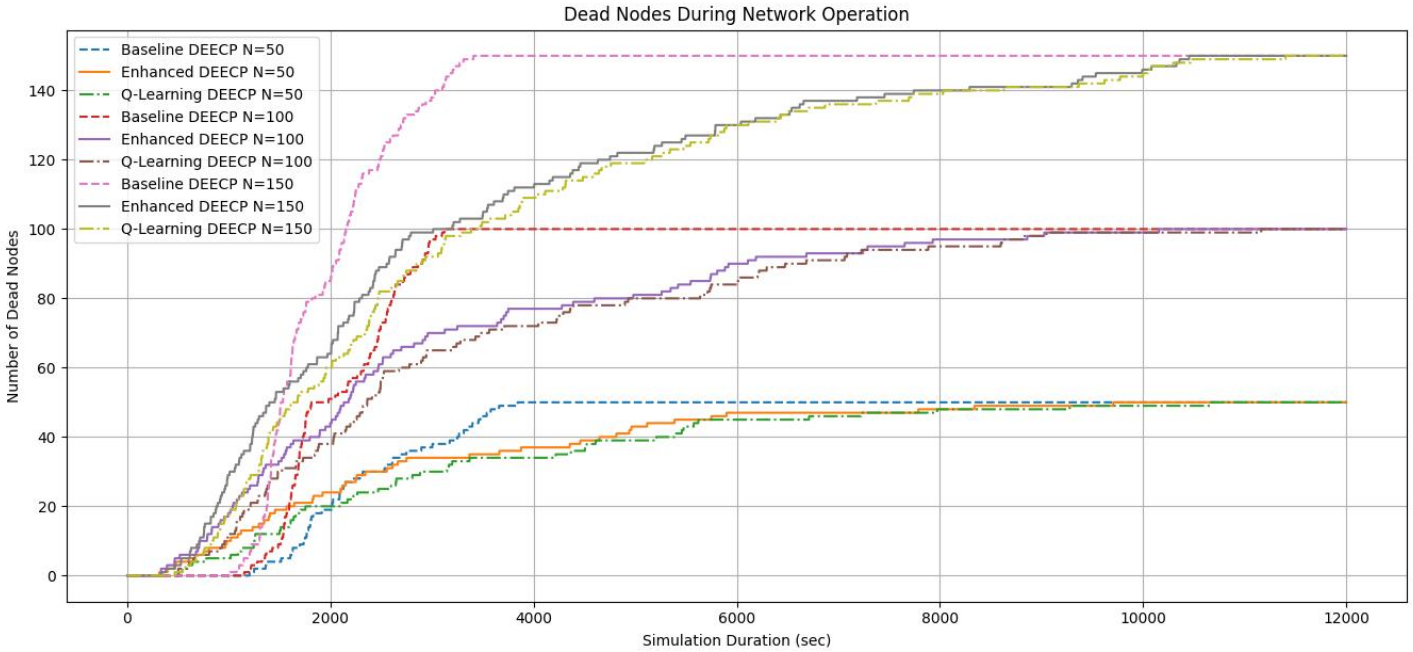

Figure 7. Number of dead nodes during network operation for different network size

Figures 7 and 8 show that, as the number of nodes increases during the stable operating phase, all algorithms benefit in terms of performance, with the Q-Learning approach consistently delivering the highest results. This improvement is linked to the higher proportion of nodes being selected as optimal CHs, particularly under the proposed method. The increase in node count leads to more balanced energy consumption across the network, which in turn extends the stability period. Using the same parameters as in Table 3, the simulations were conducted with 50, 100 and 150 nodes to study the effect of network size on stability.

As seen in Figure 8, increasing the number of nodes allows the proposed method to keep more nodes operational for longer periods. High-energy nodes can serve as CHs over extended duration, supporting network operations beyond 10,000 simulation rounds and ensuring prolonged WSN lifetime. Although adding more nodes enhances performance and lengthens operational time, it also raises deployment costs and requires widespread sensor placement in the underground mine. Therefore, a trade-off must be considered between the desired operational period and the acceptable deployment cost to achieve optimal performance in such environments. The algorithm intelligently balances energy usage across the network, resulting in a more gradual and uniform depletion of node energy, thereby extending the stability period.

Additionally, a larger number of nodes under the proposed algorithm enables greater data collection from the mining area while reducing the per-node data load. Even though higher data volumes can increase energy consumption for selected CHs, the algorithm's adaptive selection mechanism maintains network stability for the longest possible time by prioritizing the most suitable nodes for CH roles.

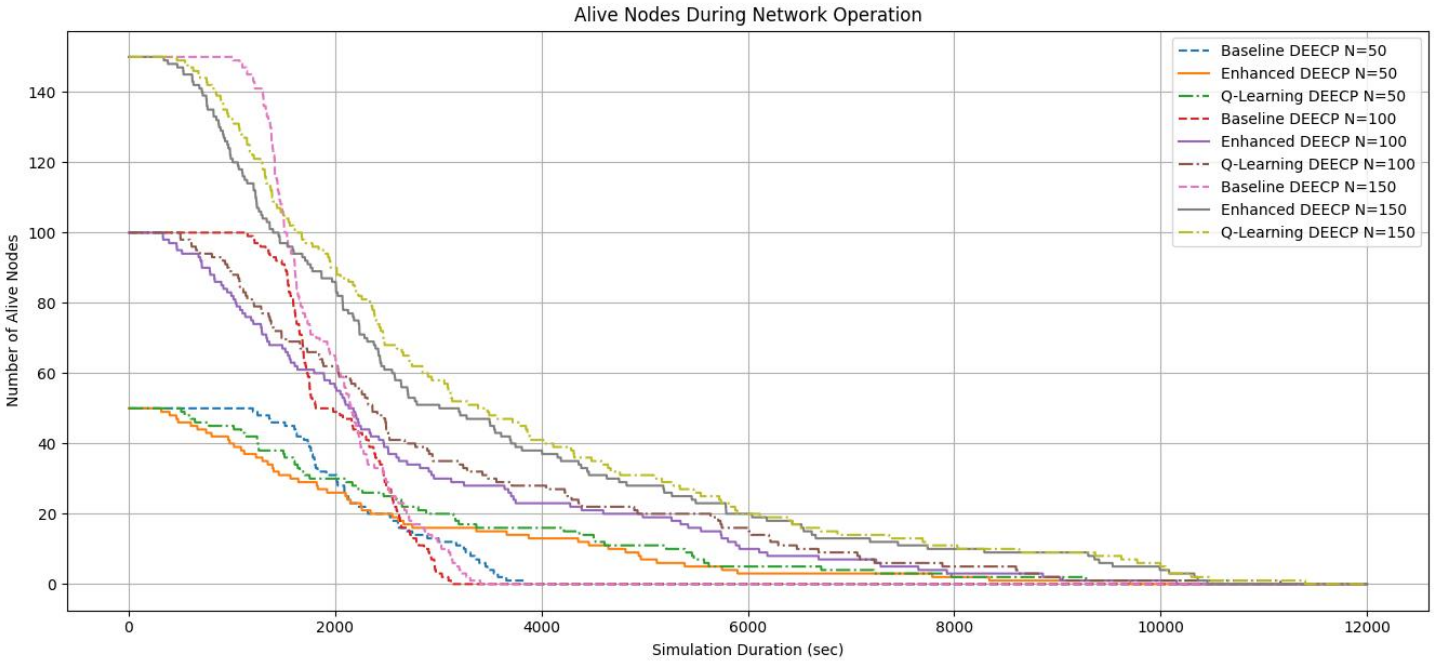

Figure 8. Number of alive nodes during network operation for different network size

Moreover, a larger node population enables QL-DEECP to distribute data transmission loads more evenly. This not only reduces the strain on individual CHs but also mitigates the risk of premature CH energy depletion a common limitation in traditional DEECP approaches. Even when total data volume grows with more nodes, the Q-learning driven selection process ensures that the network remains stable for the maximum possible duration .

VI. CONSLUSION

In this work, we introduced QL-DEECP, a Q-learning enhanced DEECP protocol tailored for heterogeneous WSN deployments in underground mining applications. While the traditional DEECP protocol incorporates residual and average energy to guide CH selection, its two-level energy model struggles to fully exploit energy heterogeneity. The enhanced DEECP (E-DEECP) addressed this by introducing three energy levels and distance-aware CH selection, but it still relied on static thresholds, limiting adaptability in dynamic mining environments.

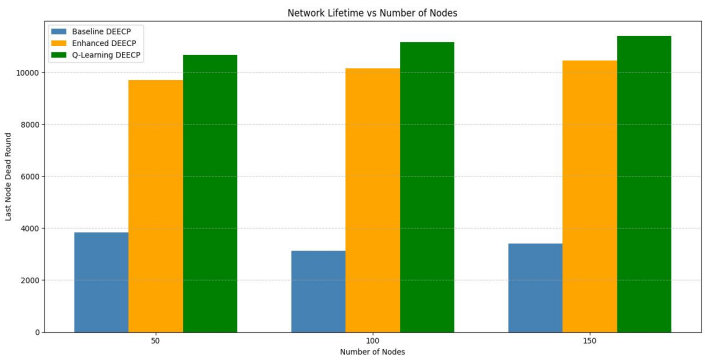

Figure 9. Number of nodes impacting the network lifetime

From figure 9, we can conclude that proposed algorithm perform better than Baseline DEECP and E-DEECP across different network sizes. The results clearly show that Baseline DEECP suffers from the shortest lifetime, with LND values of 3,842, 3,123and 3,407 rounds for 50, 100and 150 nodes, respectively.

While E-DEECP significantly improves lifetime through its three-level energy model and distance-aware CH selection achieving 9,705, 10,156and 10,452 rounds respectively whereas QL-DEECP consistently delivers the best performance. Specifically, QL-DEECP extends lifetime to 10,669, 11,162and 11,401 rounds for the same network sizes.

```
Simulation Started

Running for NUM_NODES=50
Milestone                Baseline DEECP      Enhanced DEECP      Q-Learning DEECP
1st Node Death                     1203                 315                  508
50% Nodes Dead                     2128                2117                 2596
90% Nodes Dead                     3476                5749                 6711
Last Node Death                    3842                9705                10669

Running for NUM_NODES=100
Milestone                Baseline DEECP      Enhanced DEECP      Q-Learning DEECP
1st Node Death                     1151                 331                  500
50% Nodes Dead                     2046                2179                 2363
90% Nodes Dead                     2905                6109                 6684
Last Node Death                    3123               10156                11162

Running for NUM_NODES=150
Milestone                Baseline DEECP      Enhanced DEECP      Q-Learning DEECP
1st Node Death                     1011                 340                  466
50% Nodes Dead                     1761                2234                 2413
90% Nodes Dead                     2903                6619                 6868
Last Node Death                    3407               10452                11401

Simulation Completed
```

Figure 10. Number of dead node as round progress

By analyzing figure 10, we can confirm that the integration of Q-learning improves CH election fairness, balances energy consumption more effectivelyand significantly prolongs network stability. However, challenges remain in further extending lifetime under strict underground energy constraints and in optimizing the trade-off between deployment cost and performance for dense networks. Future work will focus on combining Q-learning with swarm intelligence based optimization to refine CH selection accuracy. Additional research will also address dynamic data flow management, considering sensor type and

data volume, to further enhance stability and performance in underground mine, we can integrate intra-cluster multi-hop communication along with, varying packet size for each sensor node to better reflect real-world network traffic and optimize performance.

REFERENCES

[1] S. M. Chowdhury and A. Hossain, ‘‘Different energy saving schemes in wireless sensor networks: A survey,’’ Wireless Pers. Commun., vol. 114, no. 3, pp. 2043–2062, Oct. 2020.

[2] F. M. Salman, A. A. Mohammedand A. F. Mutar, “Optimization of LEACH protocol for WSNs in terms of energy efficient and network lifetime,” Journal of Cyber Security and Mobility, vol. 12, no. 3, pp. 275–296, May 2023, doi: 10.13052/jcsm2245-1439.1232.

[3] G. M. T. Tamilselvan and K. Gandhimathi, ‘‘Network coding based energy efficient LEACH protocol for WSN,’’ J. Appl. Res. Technol., vol. 17, no. 1, pp. 251–267, Jun. 2019

[4] W. Heinzelman, A. Chandrakasanand H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” in *Proc. 33rd Hawaii Int. Conf. System Sciences*, 2000, pp. 1–10, doi: 10.1109/HICSS.2000.926982.

[5] A. Chehri, R. Saadane, N. Hakemand H. Chaibi, ‘‘Enhancing energy efficiency of wireless sensor network for mining industry applications,’’ Proc. Comput. Sci., vol. 176, pp. 261–270, Jan. 2020.

[6] P. Kathiroli and K. Selvadurai, ‘‘Energy efficient cluster head selection using improved sparrow search algorithm in wireless sensor networks,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 2021, pp. 1–12, Sep. 2021.

[7] R. S. Sutton and A. G. Barto, “Reinforcement learning: An introduction,” MIT Press, 2nd ed., 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[8] H. Zhang, G. Liuand T. Jiang, “Parameter configuration scheme for optimal energy efficiency in LoRa-based wireless underground sensor networks,” IEEE Transactions on Vehicular Technology, vol. 74, no. 6, pp. 9961–9973, Jun. 2025, doi: 10.1109/TVT.2025.3351526.

[9] B. R. Al-Kaseem, Z. K. Taha, S. W. Abdulmajeedand H. S. Al-Raweshidy, “Optimized energy–efficient path planning strategy in WSN with multiple mobile sinks,” IEEE Access, vol. 9, pp. 79994–80007, Jun. 2021, doi: 10.1109/ACCESS.2021.3087086.

[10] F. M. Salman, A. A. Mohammedand A. F. Mutar, “Optimization of LEACH protocol for WSNs in terms of energy efficiency and network lifetime,” Journal of Cyber Security and Mobility, vol. 12, no. 3, pp. 275–296, May 2023, doi: 10.13052/jcsm2245-1439.1232.

[11] M. Gamal, N. E. Mekky, H. H. Solimanand N. A. Hikal, “Enhancing the lifetime of wireless sensor networks using fuzzy logic LEACH technique-based particle swarm optimization,” *IEEE Access*, vol. 10, pp. 36935–36948, Mar. 2022, doi: 10.1109/ACCESS.2022.3163254.

[12] S. Mohapatra, P. K. Behera, P. K. Sahoo, S. K. Bisoy, K. L. Huiand M. Sain, “Mobility induced multi-hop LEACH protocol in heterogeneous mobile network,” *IEEE Access*, vol. 10, pp. 132895–132907, Dec. 2022, doi: 10.1109/ACCESS.2022.3228576.

[13] [13] Q. Zhu, X. Wu, J. Wangand Z. Xu, “Deep reinforcement learning-based UAV trajectory planning for energy-efficient wireless sensor networks,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15646–15655, Dec. 2020, doi: 10.1109/TVT.2020.3035872.

[14] [14] A. Aleem and R. Thumma, “Hybrid energy-efficient clustering with reinforcement learning for IoT-WSNs using knapsack and K-means,” *IEEE Sensors Journal*, vol. 25, no. 15, pp. 30047–30058, Aug. 2025, doi: 10.1109/JSEN.2025..

[15] R. Alsaqour, E. S. Ali, R. A. Mokhtar, R. A. Saeed, H. Alhumyaniand M. Abdelhaq, “Efficient energy mechanism in heterogeneous WSNs for underground mining monitoring applications,” *IEEE Access,* vol. 10, pp. 100123–100138, Jul. 2022, doi: 10.1109/ACCESS.2022.3188654.

[16] Y. Zhao, S. Yangand Q. Wu, “An energy-efficient clustering routing method for wireless sensor networks based on an improved LEACH algorithm,” *Sensors,* vol. 19, no. 21, p. 4654, Nov. 2019, doi: 10.3390/s19214654.

[17] Y. Duan, X. Chen, R. Houthooft, J. Schulmanand P. Abbeel, “Benchmarking deep reinforcement learning for continuous control,” in *Proc. 33rd International Conference on Machine Learning (ICML)*, 2016, pp. 1329–1338.

[18] Y. Wang., “Reinforcement learning for reasoning in large language models with one training example,” *arXiv preprint arXiv:2504.20571*, Apr. 2025. [Online]. Available: https://arxiv.org/abs/2504.20571

[19] S. Sharma, S. Chaurasiaand P. K. Jana, “Q-learning based energy efficient and load balanced clustering algorithm for wireless sensor networks,” *Pervasive and Mobile Computing,* vol. 64, p. 101142, Dec. 2020, doi: 10.1016/j.pmcj.2020.101142.

[20] H. Mohammadi Rouzbahani, H. Karimipour and L. Lei, "Optimizing Resource Swap Functionality in IoE-Based Grids Using Approximate Reasoning Reward-Based Adjustable Deep Double Q-Learning," in *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 522-532, Aug. 2023, doi: 10.1109/TCE.2023.3279138

[21] H. Saberi, C. Zhang and Z. Y. Dong, "A Multi-Agent Deep Constrained Q-Learning Method for Smart Building Energy Management Under Uncertainties," in *IEEE Transactions on Smart Grid,* vol. 15, no. 5, pp. 4649-4661, Sept. 2024, doi: 10.1109/TSG.2024.3386896.