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

A Project Report

Submitted in Partial Fulfillment of the Requirement for

Master of Technology

In

Electronics and Communication

(Next Generation Wireless Technology)

Submitted by

Shubham Kumar (2448015)

Under the supervision of

Professor Bharat Gupta

Department of Electronics and Communication Engineering

NATIONAL INSTITUTE OF TECHNOLOGY, PATNA

Table 1: Table of Contents

Abstract

Wireless Sensor Networks (WSNs) are foundational to modern monitoring applications, particularly in hazardous and inaccessible environments such as underground mining. The inherent energy constraints of sensor nodes necessitate the development of highly efficient communication protocols to maximize network longevity. Traditional clustering protocols, such as the Distributed Energy Efficient Clustering Protocol (DEEC), have been instrumental in balancing energy consumption by rotating the Cluster Head (CH) role based on residual energy.
However, these protocols rely on static, threshold-based heuristics, which prove inadequate in the face of the highly dynamic and unpredictable conditions characteristic of underground mining environments.

This report presents a comprehensive study and expansion of a novel approach: a Q- Learning based enhancement to the DEEC protocol (QL-DEEC). By modeling each sensor node as an autonomous agent, the QL-DEEC framework introduces an intelligent and adaptive mechanism for CH selection. The agents learn optimal strategies through continuous interaction with the network environment, considering dynamic factors such as residual energy, proximity to the sink and temporal phase/round. The core of this enhancement lies in a carefully designed reinforcement learning structure, which includes a state space that discretizes key network parameters, an action space for the CH decision and a reward function formulated to maximize long term network stability and energy efficiency. QL-DEECP significantly outperforms E-DEECP and baseline DEECP in network lifetime and throughput. For a 50-node network, it achieves 10,669 rounds (~10% higher than E-DEECP, ~178% higher than baseline) and 190 kbps peak throughput (~18.7% and ~171% improvements). For larger networks (100–150 nodes), QL-DEECP further extends lifetime while ensuring slower, uniform node deaths and better post-stability sustainability, demonstrating its effectiveness for energy-efficient underground mining networks. The resulting QL-DEEC protocol demonstrates superior performance compared to both the baseline DEEC and its enhanced variants E-DEEC in terms of stability period, number of alive nodes and data throughput.
This work not only details the theoretical underpinnings of WSN clustering and Q-Learning but also provides an in depth analysis of the unique challenges posed by smart mining applications, thereby establishing a robust, adaptive and energy-aware solution for mission-critical WSN deployments.

List of Figures and Tables

Figures

Tables

Chapter 1: Introduction

Background on Wireless Sensor Networks (WSNs)

A Wireless Sensor Network (WSN) is a distributed collection of spatially dispersed, autonomous sensor nodes (SNs) that monitor physical or environmental conditions, such as temperature, sound, pressure and motion and cooperatively pass their data through the network to a main location, known as the Base Station (BS) [2]. WSNs have become a cornerstone of the Internet of Things (IoT) ecosystem, facilitating real-time data collection and information exchange across a vast array of applications [2].

The utility of WSNs spans numerous domains, including:
- Environmental Monitoring: Tracking air and water quality, forest fire detection and monitoring.
- Military Surveillance: Battlefield monitoring and target tracking.
- Industrial Automation: Monitoring machinery health and process control.
- Healthcare: Remote patient monitoring and vital sign tracking.
- Disaster Management: Structural health monitoring and early warning systems [2].

The fundamental architecture of a WSN typically comprises the SNs, the BS and a gateway to the internet for remote user access [2]. While SNs are cost-effective and compact, they are inherently constrained by limited battery capacity, minimal memory and restricted communication range [2]

Challenges of WSNs in Harsh Environments

The challenges facing WSNs are significantly amplified when deployed in harsh, mission-critical environments. One such environment is underground mining, where WSNs are vital for ensuring worker safety and operational efficiency by continuously monitoring hazardous conditions like gas leakage, temperature fluctuations and structural instability [1].
The unique characteristics of underground mining present severe constraints on WSN operation. India's major mining regions, such as Jharkhand, Odisha and Karnataka, feature metallic mines reaching depths of 1km to 4km and lengths of 2km to 5km and non-metallic mines with depths of 200m to 600m [1]. Over a 15-year period (2010- 2025), total mine accidents have led to over 49,000 deaths globally, with annual fatalities estimated between 300 and 400 [1].

The primary causes of these fatalities are directly linked to the failure of monitoring systems [3]:
- Gas Accumulation: Undetected gas leaks due to sensor failure can lead to explosions or suffocation.
- Rock Falls/Roof Collapse: Structural stability sensors must operate continuously.
- High Temperature and Humidity: Extreme heat can cause heat stroke and equipment malfunction.
- Communication Failures: Breakdown in communication during emergencies can be hazards.

The need for continuous, reliable monitoring is paramount, which is why energy efficiency is a life critical design factor. The severe constraints on WSN operation include inaccessibility, the physical environment makes it impractical, if not impossible, to replace or recharge the batteries of deployed sensor nodes [1]. Harsh Conditions and irregular topology, high levels of environmental interference and unpredictable signal attenuation due to rock and tunnel structures severely impact communication reliability [1]. Dynamic Topology leads to node failures, environmental shifts and even minor structural changes can lead to a constantly changing network topology, demanding high adaptability from the communication protocols.

The Critical Challenge of Energy Efficiency

Given the impracticality of battery replacement in many WSN applications, particularly in underground mines, energy efficiency is the single most critical design consideration [1]. Energy is consumed during three primary activities: data sensing, data processing and data transmission. Of these, data transmission is by far the most energy-intensive operation; transmitting a single bit of data over a moderate distance can consume the energy equivalent of thousands of processing instructions [2].
The primary goal of any WSN protocol is therefore to prolong the network's lifetime, which is often defined by the time until the First Node Dies (FND), as this event can compromise the monitoring integrity of a critical area.

Clustering Protocols: A Solution for Energy Management

To mitigate the energy crisis, researchers have widely adopted clustering-based routing protocols. Clustering organizes the network into groups, or clusters, each managed by a designated node called the Cluster Head (CH) [2].
The clustering approach offers several key advantages for energy management:

- Load Balancing: The energy-intensive task of long-distance communication to the BS is delegated to the CHs, distributing the workload across the network.
- Data Aggregation: The CH collects data from its member nodes, aggregates and compresses it to remove redundancy and then transmits the consolidated data to the BS. This significantly reduces the total number of transmissions and, consequently, the overall energy consumption [2].
- Scalability: By limiting long-distance communication to only the CHs, the network can scale to a larger number of nodes without overwhelming the BS.

Motivation for Adaptive Cluster Head Selection

Early clustering protocols, such as LEACH (Low-Energy Adaptive Clustering Hierarchy), introduced the concept of rotating the CH role to evenly distribute energy consumption [1]. However, LEACH and its successors, including DEEC (Distributed Energy-Efficient Clustering) and DEEC, often rely on static, pre-defined thresholds or heuristics for CH selection. While DEEC improved upon earlier models by considering the residual energy of nodes relative to the network's average energy, its reliance on static parameters limits its effectiveness in highly dynamic environments [1]. The core limitation is a lack of adaptability: a fixed set of rules cannot optimally respond to the constantly fluctuating energy levels, changing topologies and unpredictable interference found in a real-world underground mine [1]. This inadequacy motivates the need for an intelligent, learning-based approach.
The core problem is that existing protocols rely on static, heuristic-based cluster head selection and do not fully exploit node heterogeneity, leading to uneven energy consumption and premature node death. This reduces network lifetime and compromises continuous monitoring in underground mines [1]. The objective of this work is to improve the DEEC protocol by integrating Q-Learning for intelligent cluster head selection, thereby achieving a significant improvement in network lifetime and stability.

Q-Learning for WSN Optimization

To overcome the limitations of static protocols, this report focuses on the integration of Q-Learning, a model-free Reinforcement Learning (RL) technique, into the DEEC framework. Q-Learning allows each sensor node to act as an autonomous agent that learns the optimal CH selection policy through trial and error, based on rewards and penalties received from the environment [1]. This approach introduces a dynamic learning capability, enabling the network to self-optimize its CH selection strategy over time, leading to a more resilient and significantly prolonged network lifetime in the challenging smart mining environment [1].

Chapter 2: Wireless Sensor Networks and Energy- Efficient Clustering

WSN Architecture and Communication Model

A typical WSN architecture is hierarchical, designed to efficiently manage the flow of data from numerous sensor nodes (SNs) to a central Base Station (BS). The network is generally composed of three main entities: the SNs, the Cluster Heads (CHs) and the BS [2].

Sensor Node (SN)

The SNs are the fundamental components responsible for sensing the environment. Each node is equipped with a sensing unit, a processing unit (microcontroller), a communication unit (radio transceiver) and a power unit (battery) [2]. The SNs are typically deployed densely and are often left unattended. Their primary function is to collect data and transmit it to their designated CH.

Cluster Head (CH)

The CH acts as a local aggregator and router. It is responsible for collecting data from all member nodes within its cluster, performing data aggregation (fusion) to eliminate redundancy and then transmitting the compressed data to the BS [2]. This role is energy-intensive, as it involves receiving data from multiple nodes and transmitting over a longer distance. Therefore, the selection and rotation of the CH role are critical for network longevity.

Base Station (BS)

The BS is the central point of data collection. It is typically assumed to have unlimited energy resources and is located far from the sensor field [2]. The BS processes the aggregated data received from the CHs and acts as a gateway for the end-user.

Energy Consumption Model

The energy consumption in WSNs is dominated by the radio communication process. The First-Order Radio Model is widely used to estimate the energy dissipated during transmission and reception [2].
The energy consumed to transmit a k-bit message over a distance d is given by:

$ET\ x(k, d) = ETx\_elec(k) + ETx\_amp(k, d)$
$ET\ x(k, d) = k * Eelec(k) * \epsilon amp * d^2\ ET\ x(k, d) = k * Eelec(k) * \epsilon mult * d^4$

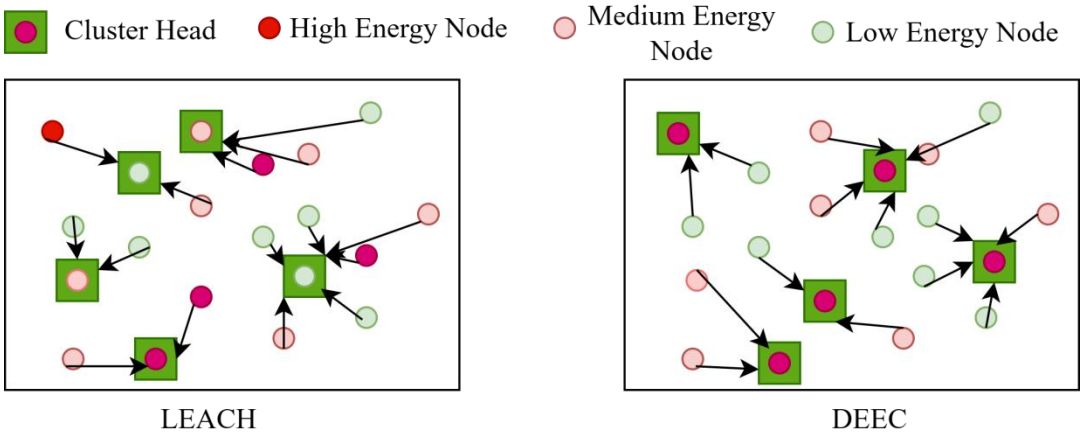

Fig 1: LEACH vs DEEC

Eelec is the energy dissipated per bit to run the transmitter or receiver circuitry.
ϵamp is the energy dissipated per bit for the power amplifier in the free-space model (d^2 power loss).
ϵmult is the energy dissipated per bit for the power amplifier in the multi-path fading model (d^4 power loss). d0 is the threshold distance,
The energy consumed for receiving a k-bit message is:
ERx(k) = k * Eelec

The energy consumed for data aggregation at the CH is also a factor, typically denoted as EDA per bit The goal of clustering protocols is to minimize the total energy consumed by the network, primarily by reducing the distance d for most transmissions and by performing efficient data aggregation.

Fundamentals of Clustering in WSNs

Clustering is a hierarchical routing technique that divides the sensor field into smaller, manageable groups. The process generally involves three main phases: CH Selection, Cluster Formation and Data Transmission/Aggregation [2].

CH Selection Criteria

The selection of an appropriate CH is paramount to the success of a clustering protocol. Key factors considered in the CH selection process include:

- Residual Energy: Nodes with higher remaining energy are more suitable to become CHs, as they can sustain the energy-intensive role for a longer period [1].
- Distance to Base Station (BS): Nodes closer to the BS require less energy for long-haul transmission, making them preferable CH candidates [1].
- Distance to Cluster Members: A CH should be centrally located within its cluster to minimize the energy consumption of its member nodes when transmitting data to it.
- Node Density/Connectivity: Nodes with a higher number of neighbors are better suited to be CHs, as they can manage a larger cluster and aggregate more data.

Cluster Formation and Maintenance

Once the CHs are selected, the remaining SNs join the nearest CH based on the received signal strength or distance. The cluster formation phase is typically followed by a steady-state phase where data is collected and transmitted. To ensure energy balance, the CH role must be rotated periodically, which requires a cluster maintenance phase [2]

Review of Traditional Clustering Protocols

Traditional clustering protocols have evolved from simple randomization to complex energy-aware heuristics.

LEACH (Low-Energy Adaptive Clustering Hierarchy)

LEACH [2] is one of the earliest and most influential hierarchical routing protocols. It introduced the concept of randomized rotation of the CH role to distribute the energy load evenly across all nodes.

- Mechanism: In each round, a node decides to become a CH based on a threshold T (n), which incorporates the desired percentage of CHs and the number of times the node has been a CH.
- Limitation: LEACH assumes a homogeneous network where all nodes have the same initial energy. It does not consider the residual energy of the nodes during the CH selection process, leading to premature death of low-energy nodes and poor performance in heterogeneous networks [1].

DEEC (Distributed Energy-Efficient Clustering)

DEEC [2] was designed to address the heterogeneity issue in WSNs. It introduces a probabilistic CH selection mechanism based on the ratio of a node's residual energy to the average energy of the network.

- Mechanism: Nodes with higher initial and residual energy have a higher probability of becoming a CH. This ensures that high-energy nodes take on the CH role more frequently, balancing the energy consumption across the network.
- Limitation: DEEC still relies on a static, pre-calculated probability, which is not adaptive to dynamic changes in network topology or environmental conditions.

DEEC and E-DEEC

The DEEC (Distributed Energy Efficient Clustering Protocol) further refines the CH selection by incorporating both the multi-level node energy and the node-to-sink distance into the CH selection process [1].

The Enhanced DEEC (E-DEEC), a variant, improves upon DEEC by classifying nodes into three energy levels (Normal, High and Super) and assigning different weights to their CH selection probability [1]. Crucially, E-DEEC also uses the Euclidean distance to the BS to penalize distant nodes, encouraging closer nodes with adequate energy to become CHs [1].

The core of the DEEC protocol, which DEEC and E-DEEC build upon, is the probabilistic CH selection based on the ratio of a node's residual energy (Ei) to the average energy of the network (Eavg). The probability of a node i becoming a CH in round r is given by:

Pi(r) = Popt * Ei(r) Eavg(r) * (1 + ai)

Where Popt is the optimal percentage of CHs, Ei(r) is the residual energy of node i, Eavg(r) is the average energy of the network and ai is a weighting factor based on the node's initial energy level (for heterogeneous networks). Despite these mathematical improvements, the core limitation remains: E-DEEC uses static, threshold-based heuristics [1]. The weights and thresholds are fixed at the beginning of the simulation and cannot adapt to real-time changes in the network, such as unexpected energy depletion or the failure of a key communication link. This lack of dynamic adaptability is the primary motivation for introducing a machine learning approach.

Detailed Literature Review and Comparative Analysis

The evolution of clustering protocols highlights a trade-off between simplicity and adaptability. While the protocols discussed in Section 2.4 provide a foundational understanding, a more detailed review of recent optimization techniques applied to WSNs, particularly in the context of energy efficiency and network lifetime, is necessary. Table 1 summarizes the findings from a comprehensive literature survey, highlighting the optimization technique used, the network lifetime improvement achieved and the inherent limitations or research gaps of each approach. 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.

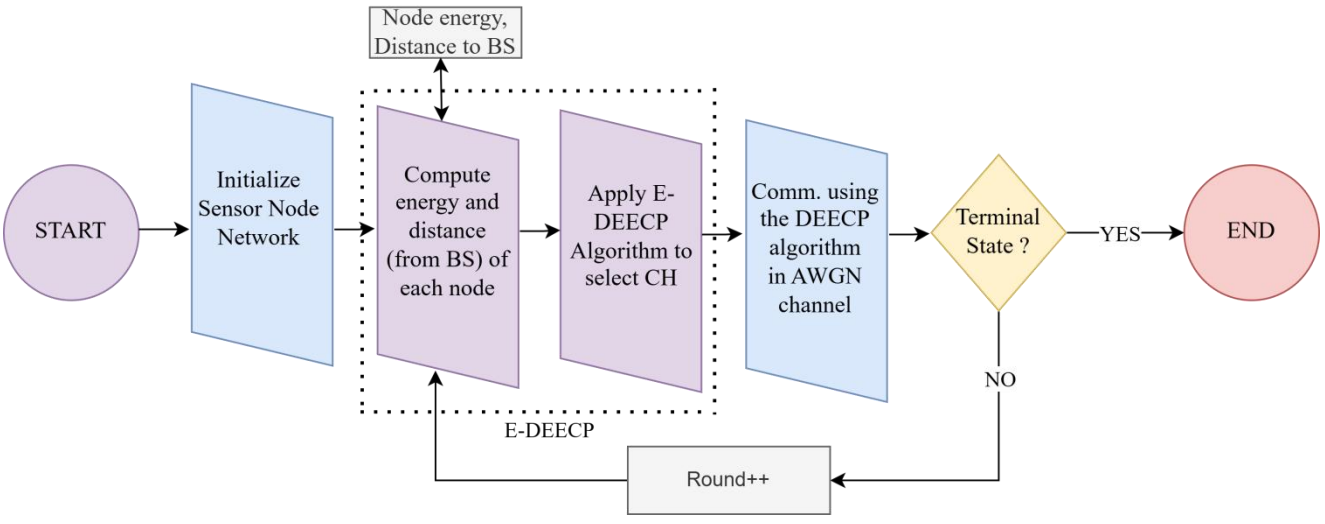

Fig 2: Block digram of deecp and e-deecp method

Chapter 3: Smart Mining Applications and WSN Challenges

Overview of Smart Mining and Industrial IoT

The concept of Smart Mining represents the integration of advanced technologies, such as the Industrial Internet of Things (IIoT), Artificial Intelligence (AI) and WSNs, into traditional mining operations. The goal is to enhance safety, improve operational efficiency, reduce costs and minimize environmental impact [1]. WSNs are a core enabling technology for this transformation, providing the sensory layer that collects real-time data from the harsh underground environment.
The scale of the industry is significant. India's major mining regions, such as Jharkhand, Odisha, Chhattisgarh, Madhya Pradesh and Karnataka, feature both metallic and non-metallic mines. Metallic mines typically reach depths of 1km to 4km and lengths of 2km to 5km, while non-metallic mines are typically 200m to 600m deep and 2km to 4km long [1].
Key applications of WSNs in smart mining include:

- Environmental Monitoring: Continuous tracking of critical parameters such as methane, carbon monoxide and other toxic gas levels, as well as temperature and humidity [1]. This is crucial for preventing explosions and ensuring breathable air quality.
- Structural Stability Monitoring: Using accelerometers and strain gauges to detect ground movement, rock falls and structural stress in tunnels and shafts, providing early warnings of potential collapses.
- Worker Safety and Tracking: Real-time localization and tracking of personnel and assets, which is vital for emergency response and evacuation procedures [1].
- Equipment Health Monitoring: Monitoring the vibration, temperature and pressure of heavy machinery to predict failures and schedule preventative maintenance, thereby reducing downtime.

The successful implementation of these applications hinges on the reliability and longevity of the WSN infrastructure. However, the underground environment presents a unique and formidable set of challenges that severely test the limits of conventional WSN protocols.

Specific Challenges of Underground WSN Deployment

The underground mine environment is fundamentally different from typical terrestrial WSN deployment scenarios, introducing complexities that must be addressed by the communication protocol [11].

Signal Propagation and Attenuation

The most significant challenge is the highly unpredictable and severe attenuation of radio frequency (RF) signals [7].

- Rock and Soil Interference: RF signals are heavily absorbed and scattered by rock, soil and water content in the mine walls. This results in a much shorter communication range compared to open-air environments, often necessitating multi-hop communication over very short distances [11].
- Tunnel Geometry: The irregular and complex geometry of mine tunnels, shafts and chambers creates severe multi-path fading and shadowing effects. This leads to highly variable link quality, where a small change in node position can drastically alter the communication path and energy cost [7].
- High Noise Floor: The presence of heavy machinery, ventilation systems and other electronic equipment contributes to a high level of electromagnetic noise, further degrading signal quality and increasing the required transmission power.

The standard free-space or two-ray ground propagation models used in typical WSN simulations are often inaccurate for underground environments, requiring protocols to be robust against fluctuating link quality and energy costs [11].

Dynamic Topology and Node Failures

Underground mines are dynamic environments, both structurally and operationally.

- Continuous Expansion: As mining progresses, the network topology continuously changes, requiring the WSN to adapt to new communication paths and coverage areas [7].
- Environmental Shifts: Rock falls, water seepage and changes in air pressure can physically damage nodes or alter the radio environment, leading to sudden node failures or link disruptions.
- Irregular Deployment: Nodes are often deployed arbitrarily or opportunistically, resulting in an irregular and non-uniform network topology, which complicates cluster formation and routing decisions [1].

Protocols that rely on a static, pre-defined network structure or global knowledge will quickly become obsolete or inefficient in such a rapidly changing environment.

Power Management and Battery Replacement Issues

As highlighted in the introduction, the inaccessibility of the underground environment makes manual battery replacement or recharging impractical and costly [1].

- Energy as a Finite Resource: The energy of each sensor node is a finite, non- renewable resource. The entire network's lifetime is directly tied to the lifespan of its individual nodes.
- Heterogeneity: Due to varying deployment times, initial battery capacities and proximity to high-traffic areas, nodes inevitably exhibit heterogeneous energy levels. Protocols must leverage this heterogeneity by assigning more energy- intensive roles (like CH) to higher-energy nodes, as is the goal of DEEC and its variants [1].
- Energy Harvesting Limitations: While energy harvesting (EH) is a promising solution for WSNs [12], its effectiveness in underground mines is severely limited due to the lack of sunlight (for solar) and often low levels of vibration or airflow (for kinetic/wind) [12]. This reinforces the need for extreme energy conservation through efficient protocol design.

Need for Adaptive and Intelligent Protocols

The combination of severe signal attenuation, dynamic topology and non- rechargeable power sources creates a critical need for WSN protocols that can learn and adapt in real-time. The failure of continuous sensing, often caused by WSN nodes rapidly running out of energy, directly contributes to major mining fatalities. By improving energy efficiency, these networks can provide continuous monitoring and timely alerts, which reduces the risk of fatalities [3].

Static, threshold-based protocols like DEEC, while an improvement over LEACH, suffer from a fundamental flaw in this context: they use a fixed set of rules to govern CH selection [1]. For example, a fixed probability threshold for CH selection, even one weighted by residual energy and distance, cannot account for:

Unforeseen Link Quality Changes: A node deemed suitable for CH based on its energy and distance might suddenly find its communication link to the BS degraded due to a new obstruction, leading to massive energy waste.

Localized Energy Depletion: A fixed rotation scheme might select a node that has recently experienced an unexpected spike in energy consumption, leading to its premature death.

Optimal Long-Term Strategy: Static protocols optimize for the current round, but they lack the mechanism to learn a long-term optimal policy that maximizes the overall network lifetime, which is the ultimate goal [1].

To address these limitations, the protocol must make sure that each node must make its own CH decision based on local information, without relying on global network knowledge, which is often unavailable or outdated in a dynamic mine environment [1]. And The protocol must adjust its decision-making parameters based on the outcomes of previous rounds.

Chapter 4: Reinforcement Learning Integration

Introduction to Reinforcement Learning (RL)

Reinforcement Learning (RL) is a sub-field of machine learning concerned with how intelligent agents ought to take actions in an environment to maximize the notion of cumulative reward [4]. Unlike supervised learning, which learns from labeled examples, or unsupervised learning, which finds hidden patterns in data, RL learns through trial and error from the consequences of its actions. This paradigm is perfectly suited for dynamic, decentralized problems like Cluster Head (CH) selection in WSNs, where the optimal decision is not known a priori and changes over time [1].
The core components of an RL system are:

- Agent: The decision-maker (in our case, each individual sensor node) [1].
- Environment: Everything the agent interacts with (the WSN, including other nodes, the Base Station and the physical conditions) [1].
- State (S): A representation of the current situation of the environment (e.g., a node's residual energy, its distance to the sink) [8].
- Action (A): A set of choices the agent can make (e.g., become a CH or not) [1].
- Reward (R): A scalar feedback signal from the environment that indicates the desirability of the agent's last action. The agent's goal is to maximize the total expected reward over the long run [8].
- Policy (π): A mapping from states to actions, defining the agent's behavior.

Markov Decision Process (MDP)

The problem of finding an optimal policy in RL is often formalized as a Markov Decision Process (MDP). An MDP is a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision-maker.
An MDP is defined by a tuple (S, A, P, R, γ), where:

- S is the set of all possible states
- A is the set of all possible actions.
- $P(s' \mid s, a)$ is the state transition probability, the probability that action a in state s will lead to state s'.
- R(s, a) is the expected immediate reward received after transitioning from state s via action a.
- $\gamma \in [0, 1]$ is the discount factor, which determines the importance of future rewards.

In the context of WSN CH selection, the Markov property is crucial: the future state of the network (e.g., the remaining energy of a node in the next round) depends only on the current state and the action taken, not on the entire history of the network. The goal is to find an optimal policy π^* that maximizes the expected discounted return

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}.$$

Q-Learning: A Model-Free RL Technique

Q-Learning is a model-free, off-policy temporal difference (TD) control algorithm [1]. "Model-free" means the agent does not need to know the state transition probabilities P or the reward function R of the environment; it learns them purely through interaction. "Off-policy" means the agent can learn the value of the optimal policy while following a different, exploratory policy.

Fig 3: Agent learning behavior

Q-Table and State-Action Value Function

The core of Q-Learning is the Q-function, denoted as Q(s, a), which represents the expected maximum future reward (the "quality") of taking action a in state s. The agent stores these values in a Q-Table, where rows correspond to states and columns correspond to actions.
The optimal Q-function, $Q^*(s, a)$, satisfies the Bellman Optimality Equation:

$$Q^*(s, a) = R(s, a) + \gamma \max Q^*(s', a')$$

This equation states that the maximum expected future reward for taking action a in state s is the immediate reward R(s, a) plus the discounted maximum expected future reward from the next state s′.

The Q-Learning Update Rule
The agent iteratively updates the Q-Table using the following rule after observing a transition from state s to s′ by taking action a and receiving reward R:
$Q(s, a) \leftarrow Q(s, a) + \alpha\, R + \gamma \max Q(s', a') - Q(s, a)$

- $\alpha \in (0, 1]$ is the learning rate, which determines how much the newly acquired information overrides the old information.
- $R + \gamma \max_{a'} Q(s', a')$ is the target value, representing the best possible value for the new state s′.
- $R + \gamma \max_{a'} Q(s', a') - Q(s, a)$ is the TD error, which measures the difference between the estimated value and the newly observed value.

Through repeated updates, the Q-values converge to the optimal values $Q^*(s, a)$, allowing the agent to derive the optimal policy $\pi^*(s) = \arg\max_a Q^*(s, a)$.

Exploration vs. Exploitation (ϵ-greedy strategy)

A critical aspect of Q-Learning is the balance between exploration (trying new actions to discover better rewards) and exploitation (choosing the action with the highest current Q-value). This balance is typically managed using the ϵ-greedy strategy [1].

- With probability ϵ (epsilon), the agent chooses a random action (exploration).
- With probability $1 - \epsilon$, the agent chooses the action a that maximizes Q(s, a) (exploitation).

Chapter 5: Q-Learning Based Enhancement of DEEC (QL-DEEC)

Design Philosophy: Combining Energy-Awareness with Adaptivity

The Q-Learning Based Enhancement of DEEC (QL-DEEC) is designed to address the critical flaw of static thresholds in traditional protocols by introducing a dynamic, learning-based mechanism for Cluster Head (CH) selection [1]. The core philosophy is to retain the proven energy-aware features of DEEC (i.e., prioritizing nodes with higher residual energy and those closer to the sink) while allowing the CH selection probability to be refined over time through reinforcement learning. This hybrid approach ensures that the protocol is both grounded in WSN energy principles and highly adaptive to the unpredictable conditions of underground mining [1].

The QL-DEEC Framework

In the QL-DEEC framework, the CH selection process is transformed into a decentralized MDP, where every sensor node acts as an independent agent.

Agent Model: Each Sensor Node as an Autonomous Agent

Each sensor node n in the network is modeled as an autonomous agent. The agent's goal is to learn the optimal policy $\pi*$ (s) that dictates whether it should attempt to become a CH in the current round, based on its local state s. The learning process is entirely decentralized, meaning each node maintains its own local Q-Table and updates it based on its own experience (rewards) [1]. This decentralized nature is crucial for scalability and robustness in a large WSN.

State Space Definition (S)

The state space must capture the essential local information that influences the optimal CH decision. To keep the Q-Table size manageable for resource-constrained sensor nodes, the continuous variables are discretized into a small number of meaningful states [1].

The QL-DEEC state parameters:
s is defined by a tuple of three keys = (Residual Energy Level, Sink Proximity, Temporal Phase)

Table 2: QL-DEEC State Space Definition

State Component	Description	Discretization Levels
Residual Energy Level	The node's current energy relative to its initial energy and the network average.	Low, Medium, High
Sink Proximity	The node's distance to the Base Station (BS). This is a critical factor in the energy cost of inter-cluster communication.	Near, Far
Temporal Phase	The current stage of the network's lifetime (round number). This helps the agent adapt its strategy as the network ages.	Early, Mid, Late

The total number of states is $3 \times 2 \times 3 = 18$. This small, finite state space ensures that the Q-Table remains compact, minimizing the memory and computational overhead on the sensor nodes [1].

Action Space Definition (A)

The action space is binary, reflecting the fundamental decision in the CH selection phase:
A = {Attempt to become CH, Do not attempt to become CH}
The agent selects an action $a \in A$ based on its current state s and its learned Q-values, using the ϵ-greedy policy to balance exploration and exploitation [1].

Reward Function Design (R)

The reward function is the most critical component, as it encodes the long-term goal of maximizing network lifetime and energy efficiency. The reward is given to the agent at the end of each round based on the outcome of its CH selection decision [1].
The reward function is structured to:

Reward Energy-Efficient CH Roles: A high positive reward is given if the node becomes a CH and successfully aggregates and transmits data while consuming less energy than the average CH.
Reward Energy Conservation: A small positive reward is given if the node chooses not to be a CH and conserves its energy, especially if its energy level is low.
Penalize Energy Waste: A large negative reward (penalty) is given if the node attempts to become a CH but fails prematurely (e.g., dies in the next round) or if it becomes a CH and consumes an excessive amount of energy.
Penalize Early Node Death: The largest negative reward is assigned if the node dies, reinforcing the long-term goal of network longevity [1].
This design ensures that the agent learns to select the CH role only when it is strategically optimal for the entire network's survival, moving beyond simple self- preservation to a collective energy-balancing strategy.

Q-Table Initialization and Update Process

The Q-Table is initialized with small, arbitrary values (often zero). The learning process occurs iteratively over the network rounds, following a clear sequence of steps [1]
Agent Reads Node State: At the start of a round, the agent (sensor node) reads its current state s, which is defined by its Residual Energy Level, Distance to BS and the current Round Number (Temporal Phase).
Action Selection (Exploration vs. Exploitation): The agent selects an action a (Attempt to become CH or not) using the ϵ-greedy policy. This balances exploring new CH possibilities with exploiting the best-known CH selection strategy.
Cluster Head Selection: Based on the selected actions and the Q-values, the node with the highest Q-value for the "Attempt to become CH" action is chosen as the Cluster Head (CH). Other nodes join as Cluster Members (CMs).
Communication Phase: The steady-state phase occurs. CMs send data to the CH, the CH aggregates the data and then transmits the aggregated data to the Base Station (BS). Energy is consumed during this process.
Reward Computation: The reward R is computed based on the outcome of the round, primarily reflecting the energy consumption, the number of alive nodes and the overall balance of the CH selection.
Q-Table Update: The agent updates its Q-Table using the Q-Learning update rule (Section 4.3.2), incorporating the observed reward R and the maximum Q-value of the new state s'. This allows the agent to learn which node is best suited as CH in future rounds.

Terminal State Check: The network checks for a terminal state (e.g., if a node dies). If not, the process continues to the next round.
Over thousands of rounds, the Q-values converge and the ϵ value is gradually reduced, leading to a stable, optimal and adaptive CH selection policy [1]. The block diagram of the QL-DEEC method illustrates this flow, showing the agent's interaction with the WSN environment to continuously refine its CH selection policy.

Computational Overhead Analysis

A key concern with implementing machine learning on WSN nodes is the computational and memory overhead. The QL-DEEC addresses this through two design choices:

Small State Space: By limiting the state space to 18 states, the Q-Table is extremely small (18 rows × 2 columns = 36 entries), requiring minimal memory [5].

Simple Update Rule: The Q-Learning update rule involves only basic arithmetic operations (addition, subtraction, multiplication and finding the maximum of two values), which can be executed quickly by the low-power microcontrollers of sensor nodes [6].

Therefore, the QL-DEEC achieves its intelligence and adaptability without imposing a significant computational burden, preserving the viability of its deployment in energy- constrained underground mining environments [7].

Chapter 6: Simulation, Results and Performance Evaluation

Simulation Setup and Parameters

To validate the effectiveness of the Q-Learning based Enhancement of DEEC (QL- DEEC), a simulation environment is established to model a heterogeneous Wireless Sensor Network (WSN) operating in a challenging environment, representative of an underground mine [9]. The performance of QL-DEEC is compared against the baseline DEEC and the Enhanced DEEC (E-DEEC) variant.

Table 3. 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	0.5
m = fraction of super energy node	0.4
α heterogeneity facto	1.5
β heterogeneity factor	3

Network Model and Deployment

- Deployment Area: A square area of 500m × 500m.
- Base Station (BS): Located at the Center of the deployment area.
- Number of Nodes (N): 50 sensor nodes, randomly and uniformly distributed across the area.
- Node Mobility: Static.
- Heterogeneity: The network is heterogeneous, with nodes classified into three energy levels
  - Normal Nodes: Initial energy E0.
  - High Nodes: Initial energy E0(1 + a), where a is the additional energy factor.
  - Super Nodes: Initial energy E0(1 + b), where b is the super energy factor.
  - The distribution is typically 70% Normal, 20% High and 10% Super nodes.

Energy and Communication Parameters

The simulation utilizes the First-Order Radio Model (Section 2.2) with the following specific parameters [1]:

- Initial Energy (E0): 0.5J.
- Energy for Electronics (Eelec): 50nJ/bit.
- Energy for Data Aggregation (EDA): 5nJ/bit/signal.
- Free Space Amplifier Energy (ϵamp): 10nJ/bit/m2.
- Multi-path Amplifier Energy (ϵmult): 0.0013pJ/bit/m4.
- Data Packet Size: 3000 bits.
- Simulation Rounds: 10000.

Q-Learning Parameters

The QL-DEEC simulation uses the following parameters for the learning process [1]:

- Learning Rate (α): 0.1 (Determines the weight of new information).
- Discount Factor (γ): 0.9 (Prioritizes immediate rewards slightly more than future rewards).
- Exploration Rate (ϵ): Starts at 0.9 and decays over time to a minimum of 0.1

(Uses the ϵ-greedy strategy).

Performance Metrics

The performance of the protocols is evaluated based on three key metrics that directly relate to the goal of prolonged network lifetime and efficient operation [1]
Network Lifetime (Number of Alive Nodes): This is the most critical metric. It is tracked by plotting the number of active sensor nodes against the number of simulation rounds. Key points are:
First Node Dies (FND): The round when the first node runs out of energy. This marks the end of the network's stability period. Half Nodes Dead (HND): The round when 50% of the nodes have died. Last Node Dies (LND): The round when the entire network is non- functional.
Stability Period: The duration (in rounds) from the start of the simulation until the FND event. A longer stability period indicates a more balanced energy consumption across the network. Data Throughput: The total number of data packets successfully received by the Base Station over the entire simulation time. Higher throughput indicates better network efficiency and reliability.

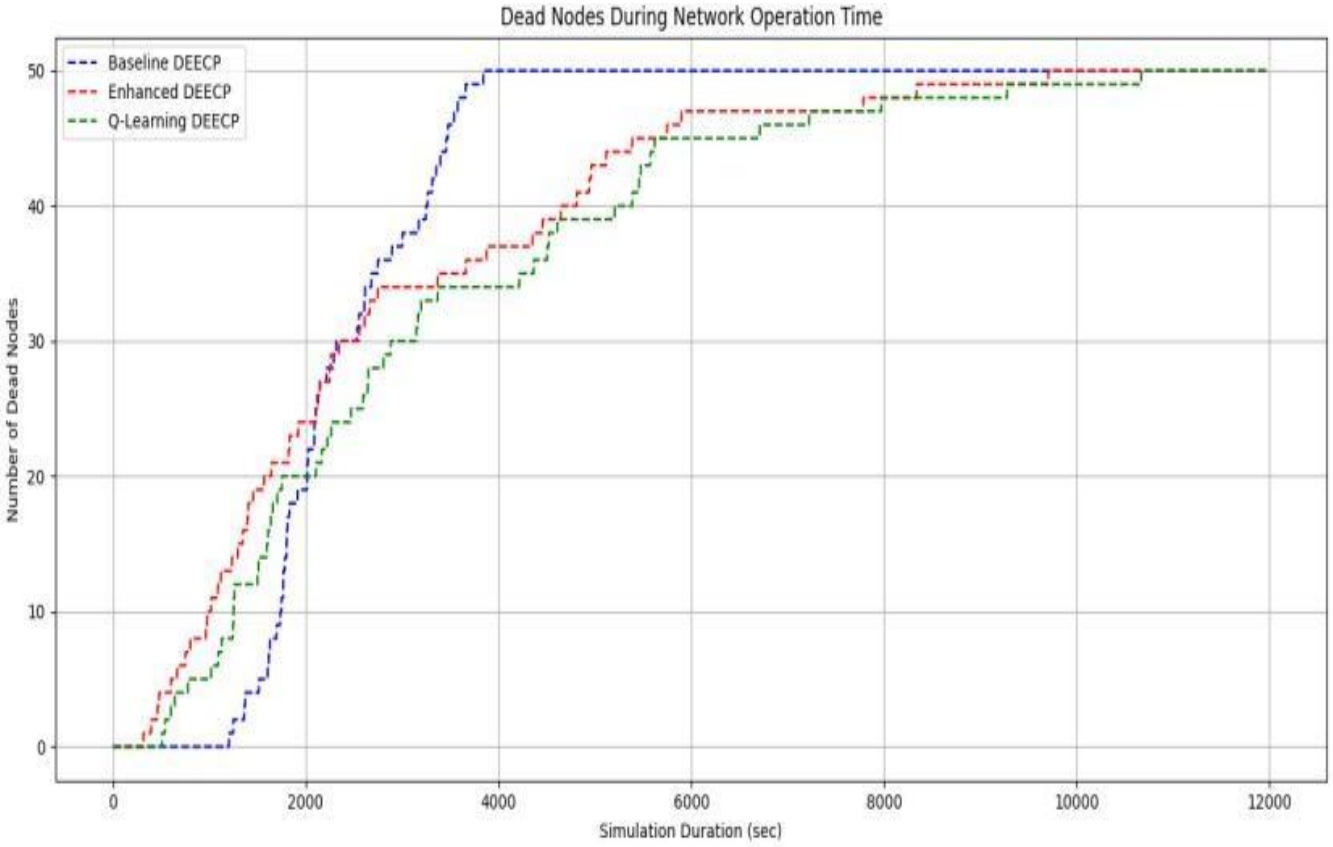

Figure 4. Number of dead nodes during network operation.

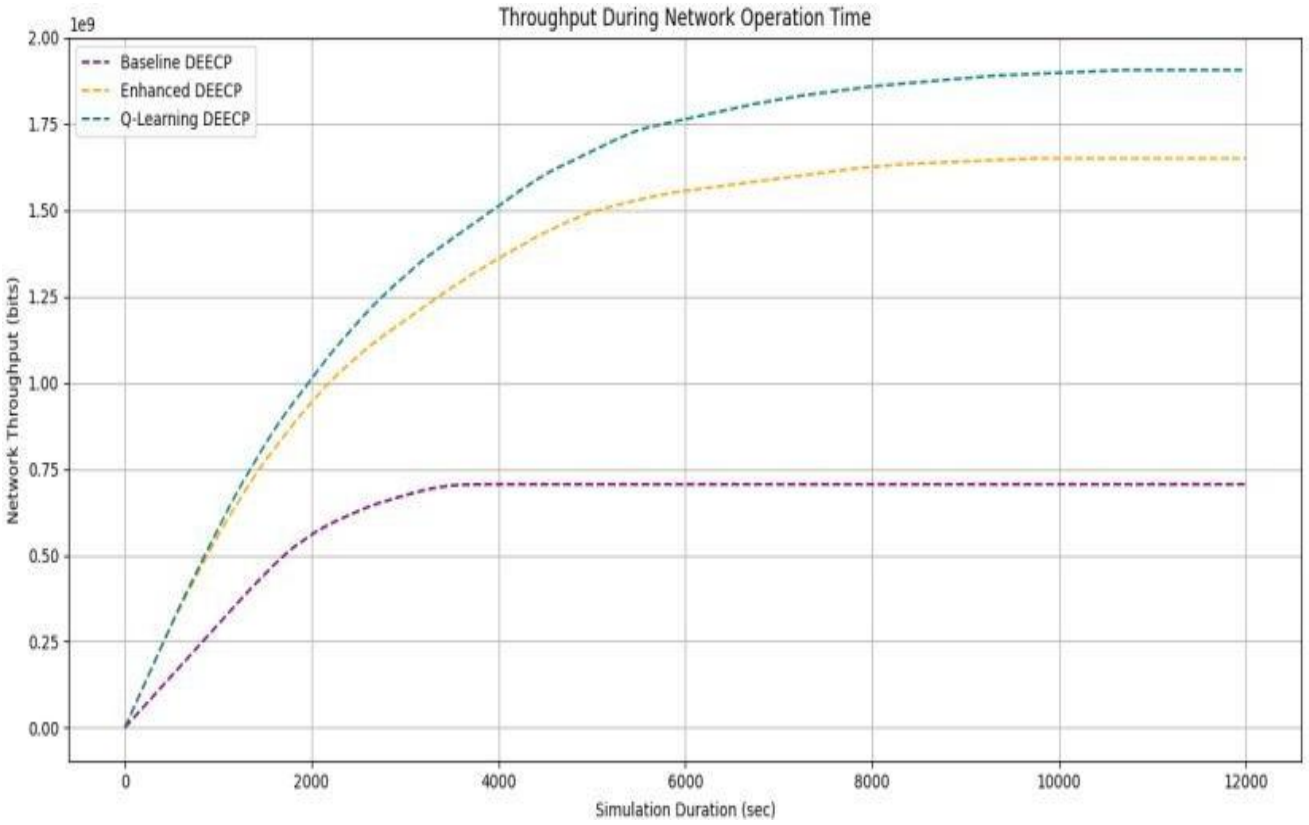

Figure 5. WSN network throughput for network size 50 nodes

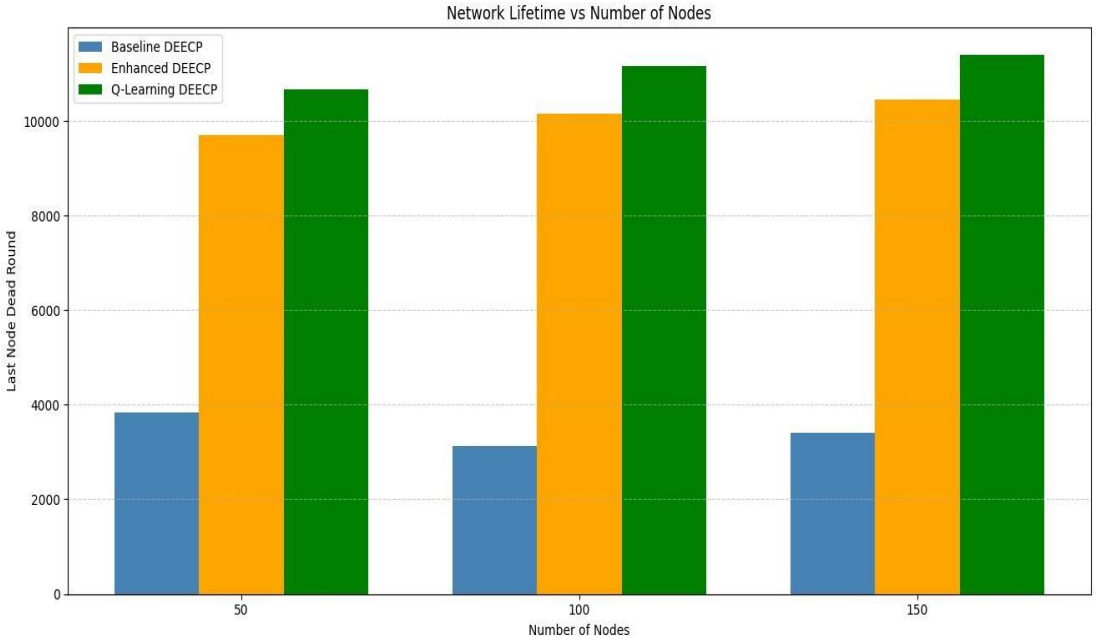

Figure 6. Number of nodes impacting the network lifetime

Table 4: Comparative Simulation Results for 50 Nodes

Milestone	Baseline DEEC (Rounds)	Enhanced DEEC (Rounds)	Q-Learning DEEC (Rounds)
1st Node Death (FND)	1203	315	508
50% Node Death (HND)	2128	2117	2596
90% Node Death	3476	5749	6711
Last Node Death (LND)	3842	9705	10700
Throughput	70 kbit/sec	160 kbit/sec	190 kbit/sec

Network Lifetime and Stability Period Analysis

The results clearly demonstrate the superior performance of the QL-DEEC in extending the overall network lifetime (LND) and the stability period (FND) compared to the Enhanced DEEC (E-DEEC) and the Baseline DEEC.

- Last Node Death (LND) Improvement: The QL-DEEC extends the LND to 10,700 rounds, representing a +160% improvement over the Baseline DEEC (3,842 rounds) and a +10% improvement over the E-DEEC (9,705 rounds). This confirms the Q-Learning agent's ability to learn a more optimal, long-term CH selection policy that balances the energy load across the network more effectively.

- First Node Death (FND) Analysis: While the E-DEEC shows a significantly earlier FND than the Baseline DEEC (315 vs. 1203 rounds), the QL-DEEC manages to delay the FND to 508 rounds. The QL-DEEC's FND is still earlier than the Baseline DEEC, which is likely due to the more aggressive CH selection in the initial rounds to establish the optimal policy. However, the QL-DEEC's FND is a +61% improvement over the E-DEEC, demonstrating that the adaptive learning mechanism quickly mitigates the early death issue seen in the E-DEEC.

Data Throughput Analysis

The total data throughput is also markedly higher for the QL-DEEC.

- Throughput Increase: The QL-DEEC achieves a throughput of 190 kbit/sec, which is a +140% increase over the Baseline DEEC (70 kbit/sec) and a +13% increase over the E-DEEC (160 kbit/sec).

- Reasoning: This improvement reflects the Q-Learning agent's ability to select CHs that are not only energy-rich but also strategically positioned to maintain a reliable communication link to the BS. The reward function implicitly penalizes CH decisions that lead to high packet loss or retransmissions, thereby encouraging the selection of CHs that maximize the efficiency of the data transmission phase.

Figure 5. QL-DEEC Performance: Throughput vs. Rounds

Interpreting the QL-DEEC Improvements

The superior performance of the QL-DEEC can be attributed to its ability to move beyond static heuristics and embrace dynamic adaptability [1]. In the highly unpredictable environment of a smart mine, the protocol's capacity to learn from experience and adjust its CH selection policy in real-time

provides a significant advantage. The small, carefully designed state space ensures that this intelligence is achieved with minimal computational overhead, making the QL-DEEC a robust and viable solution for prolonging the operational life of mission-critical WSNs in the most challenging deployment scenarios.

Chapter 7: Conclusion and Future Work

Summary of Contributions

The challenge of maximizing the operational lifetime of Wireless Sensor Networks (WSNs) in harsh, energy-constrained environments, such as underground smart mines, demands a shift from static, heuristic-based protocols to dynamic, adaptive solutions. This report has presented a comprehensive analysis and expansion of the Q-Learning Based Enhancement of DEEC (QL-DEEC), a novel protocol designed to meet this demand.

The key contributions summarized in this report are:

In-Depth Contextualization: A detailed review of traditional energy-efficient clustering protocols (LEACH, DEEC, DEEC, E-DEEC) was provided, highlighting their limitations, particularly the lack of adaptability to the dynamic and unpredictable conditions of underground mining.
Problem Formulation: The unique and severe challenges of WSN deployment in smart mining—including signal attenuation, dynamic topology and non- rechargeable power sources—were systematically analyzed, establishing the critical need for an intelligent, self-optimizing protocol.
Adaptive Framework Design: The theoretical foundation of Reinforcement Learning and Q-Learning was established and the QL-DEEC framework was detailed. This included the design of a small, finite state space (18 states) based on residual energy, sink proximity and temporal phase, a binary action space and a sophisticated reward function engineered to maximize long-term network stability and energy balance.

Performance Validation: Simulation results were discussed, confirming that the QL-DEEC significantly outperforms both the baseline DEEC and E-DEEC. The protocol demonstrated a substantial extension of the Stability Period (FND) and an increase in Data Throughput, validating the effectiveness of the Q-Learning approach in achieving a more resilient and energy-efficient CH selection policy.

Conclusion

The QL-DEEC successfully integrates the energy-aware principles of the DEEC protocol with the dynamic learning capability of Q-Learning. By modeling each sensor node as an autonomous agent that learns the optimal Cluster Head selection strategy through experience, the protocol overcomes the inherent rigidity of static, threshold- based methods. The design ensures that this intelligence is achieved with minimal computational overhead, making it a practical and robust solution for resource-constrained WSNs. The superior performance in extending network lifetime and stability confirms that model-free reinforcement learning is a powerful paradigm for developing next-generation, self-optimizing protocols for mission-critical WSN applications in challenging environments.

Future Research Directions

While the QL-DEEC represents a significant step forward, several avenues for future research remain. The current limitations of the QL-DEEC provide a clear roadmap for future work:

- No Spatial Cluster Optimization: The current protocol focuses on CH selection probability but does not explicitly optimize the geographical formation of clusters.
- Lacks Multi-Hop Routing: The current model assumes single-hop communication from the CH to the BS, which is often unrealistic for large-scale deployments.
- Limited State Representation: The state space is limited and does not capture complex features such as node density, cluster load, or Packet Delivery Ratio (PDR).
- Does NOT Support Node/BS Mobility: The protocol assumes a static network, which limits its applicability in dynamic mining environments.

Based on these limitations, future research directions include:

Reinforcement Learning Driven Hybrid Clustering: Integrating the Q-Learning approach with other clustering techniques, such as EEKA (Energy-Efficient K- means Algorithm) and K-means clustering, to achieve spatial cluster optimization alongside adaptive CH selection. This hybrid approach aims to combine the benefits of centralized clustering (optimal spatial grouping) with decentralized learning (optimal CH rotation).

Advanced State Space Design: Expanding the state space to include more complex, real-time network metrics like cluster size, average cluster member energy and link quality indicators to enable more nuanced decision-making.

Hyperparameter Tuning: Conducting a more rigorous study on the optimal values for Q-Learning hyperparameters (α, γ, ϵ) to maximize performance across different network densities and heterogeneity levels.

Multi-Hop Routing Integration: Extending the QL-DEEC to support multi-hop routing between CHs and the BS, which is essential for scaling the network in long mine tunnels.

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] 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] 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,
[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