Data Mining-Based Smart Cluster Head Selection (SCHS) Approach for Energy Efficiency in Wireless Sensor Networks

Surendra Singh Chauhan[1], Gundeep Tanwar[2], Pinki[3], Rashmi Tiwari[4], Balaji Venkateswaran[5], Waseem Ahmad[6]

[1]Associate Professor, Department of Computer Science and Engineering, SRM University, Sonepat (Haryana), INDIA.
[2]Assistant Professor, Department of Computer Science & Engineering, RPS College of Engineering & Technology, Mahendergarh (Haryana), INDIA.
[3]Lecturer, Computer Science and Engineering, Raja Jait Singh Government Polytechnic, Faridabad (Haryana), INDIA.
[4]Research Scholar (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula (UP), INDIA.
[5]Research Scholar (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula (UP), INDIA.
[6]Assistant Professor, Department of Computer Science and Engineering, Babu Banarasi Das Northern India Institute of Technology, Lucknow (UP), INDIA.

Email: [1]surendrahitesh1983@gmail.com, [2]mr.tanwar@gmail.com, [3]panku.vashisth95@gmail.com, [4]tiwari.rasmi087@gmail.com, [5]balaji.venkateswaran@gmail.com, [6]waseem82siddiqui@gmail.com

ABSTRACT

In Wireless Sensor Networks (WSNs), efficient energy management is crucial for extending network lifespan, particularly given the dynamic mobility and communication demands of ad hoc mobile devices. Traditional Cluster Head (CH) selection methods, which organize nodes into clusters for data routing and management, often suffer from biases that lead to uneven energy depletion. CHs tend to exhaust their energy rapidly due to excessive workload, resulting in network instability. To address this issue, this paper presents an enhanced CH selection approach based on the K-means algorithm, ensuring a more balanced energy distribution across all nodes. The proposed method considers multiple critical factors, including residual energy, node density, distance to the base station, and signal strength, to make informed CH selections. By integrating these parameters, the algorithm promotes equitable CH rotations, preventing premature energy depletion and enhancing network sustainability. Extensive simulations evaluate the proposed approach against conventional CH selection protocols such as LEACH (Low-Energy Adaptive Clustering Hierarchy) and HEED (Hybrid Energy-Efficient Distributed). Performance analysis based on residual energy, packet delivery ratio, throughput, and the number of live and dead nodes demonstrates that the proposed K-means-based approach significantly improves energy efficiency and overall network performance.

Keywords: Cluster Head (CH) Selection, K-means Algorithm, LEACH, HEED.

1. INTRODUCTION

Wireless Sensor Networks (WSNs) have gained significant attention due to their broad range of applications, including environmental monitoring, industrial automation, healthcare, and smart cities. These networks consist of numerous sensor nodes that collect and transmit data to a central base station (BS). However, due to the limited battery life of sensor nodes, efficient energy management is crucial to prolong network lifespan and maintain operational efficiency [1].

Earlier approaches to energy-efficient data transmission in WSNs relied on direct communication and multi-hop routing protocols. While these methods facilitated basic network connectivity, they often led to rapid energy depletion, particularly in nodes closer to the BS. The introduction of clustering-based protocols, such as the Low-Energy Adaptive Clustering Hierarchy (LEACH) and Hybrid Energy-Efficient Distributed (HEED), revolutionized WSN energy management by selecting Cluster Heads (CHs) to aggregate and forward data. However, these traditional methods often suffered from biased CH selection, where certain nodes were repeatedly chosen, leading to premature energy exhaustion and network partitioning [2-3].

To overcome these limitations, recent research has focused on optimizing CH selection strategies through advanced computational techniques such as machine learning, metaheuristic algorithms, and data-driven clustering methods. Among these, K-means clustering has emerged as a promising approach due to its ability to organize nodes efficiently while considering multiple parameters. In this study, we propose an enhanced K-means-based CH selection algorithm that factors in residual energy, node density, distance to the BS, and signal strength indicators. By incorporating these critical parameters, our approach ensures a more balanced CH selection, mitigating uneven energy consumption and improving overall network sustainability [4].

While the proposed approach enhances energy efficiency, future advancements in WSNs are expected to leverage artificial intelligence (AI), deep learning, and blockchain-based security mechanisms for even more adaptive and secure CH selection. AI-driven dynamic clustering techniques can enable real-time decision-making, while blockchain can enhance network trust and security. Moreover, the integration of energy-harvesting technologies, such as solar-powered nodes, can further extend network longevity. The evolution of 6G and edge computing will also play a pivotal role in enabling ultra-low-latency and energy-efficient WSN deployments [5-6].

2. RELATED WORK

Wireless Sensor Networks (WSNs) are composed of numerous low-cost, energy-constrained sensor nodes that integrate sensing, computation, and communication capabilities. These nodes form a self-organizing mesh network, enabling seamless data transmission across multiple hops to ensure reliable communication. Despite their potential, WSNs face significant challenges due to limited energy resources, as sensor nodes are often deployed in inaccessible environments where battery replacement or recharging is impractical. Therefore, energy efficiency remains a critical concern in WSN design.

Clustering in Wireless Sensor Networks

Clustering has emerged as a key strategy for enhancing the energy efficiency of WSNs. In a clustered network, sensor nodes are grouped into clusters, each managed by a designated Cluster Head (CH). The CH is responsible for aggregating and forwarding data to the Base Station (BS), thereby reducing redundant data transmission and conserving energy [7]. This hierarchical structure optimizes energy consumption by minimizing the number of direct communications between individual sensor nodes and the BS. With advancements in micro-fabrication technology reducing the cost of sensor nodes, large-scale WSN deployments are becoming more feasible [8]. These networks have a wide range of applications, including environmental monitoring, healthcare, military surveillance, industrial automation, and smart city infrastructure. Additionally, the integration of various sensors such as seismic, thermal, magnetic, and optical sensor enhances the versatility of WSNs, allowing them to monitor diverse environmental parameters, including temperature, humidity, pressure, and movement [9].

Challenges in Cluster Head Selection

Despite the advantages of clustering, CH selection remains a critical challenge due to the increased energy burden on CH nodes. Since CHs handle data aggregation and long-distance transmission, they tend to deplete their energy more quickly than regular nodes. Inefficient CH selection can lead to network instability and premature node failure, reducing overall network lifetime. Thus, designing an optimal CH selection algorithm that balances energy consumption across the network is essential for prolonging WSN lifespan and maintaining performance [10]. Several clustering algorithms have been proposed to enhance energy efficiency in WSNs, with LEACH and HEED being among the most widely studied.

- ***LEACH (Low-Energy Adaptive Clustering Hierarchy):*** LEACH is a hierarchical clustering algorithm that randomly selects CHs in each round to ensure energy is distributed evenly across nodes. CH rotation helps prevent rapid energy depletion in specific nodes, extending network lifetime. However, LEACH's probabilistic approach may lead to suboptimal cluster formation, particularly in large-scale networks, affecting scalability and overall efficiency [11].
- ***HEED (Hybrid Energy-Efficient Distributed Clustering):*** HEED improves upon LEACH by incorporating multiple selection criteria, including residual energy and communication cost, to determine CHs. This hybrid approach ensures that CHs are chosen based on their energy levels and network topology, leading to a more balanced and stable clustering structure. Additionally, HEED incorporates a cost metric to prevent any single node from being overburdened, thereby enhancing network sustainability [12].

Although these traditional clustering algorithms have contributed significantly to energy efficiency in WSNs, they still have limitations in handling large-scale networks, dynamic topologies, and real-time energy balancing. Recent advancements have explored data-driven and machine learning-based approaches to optimize CH selection, leveraging techniques such as K-means clustering to improve network longevity and performance.

3. PROBLEM FORMULATION AND RESEARCH METHODOLOGY

Wireless Sensor Networks (WSNs) play a crucial role in various applications, including environmental monitoring, industrial automation, and smart cities. However, one of the most significant challenges in WSNs is managing the limited battery capacity of sensor nodes, as replacing or recharging them is often impractical due to their deployment in remote or inaccessible locations. These sensor nodes are typically deployed in an ad hoc manner and operate collaboratively to collect and transmit data to a centralized Base Station (BS). Given these constraints, efficient energy management is essential to extend network longevity and ensure uninterrupted functionality.

A fundamental approach to addressing energy limitations in WSNs is the separation of data collection and data transmission processes. This involves two distinct phases: (1) Data aggregation at the sensor node level, where nodes collect and pre-process environmental information to minimize redundant transmissions, and (2) Efficient data forwarding, where selected nodes transmit aggregated data to the BS in an energy-efficient manner. Clustering-based routing techniques, such as LEACH (Low-Energy Adaptive Clustering Hierarchy), HEED (Hybrid Energy-Efficient Distributed Clustering), and K-means clustering, have been widely utilized to optimize these processes. These algorithms aim to distribute energy consumption evenly across the network by selecting Cluster Heads (CHs) responsible for managing data aggregation and communication [13].

Despite these advancements, a major challenge persists in determining the most effective criteria for CH selection to balance energy consumption and network performance. Many existing clustering algorithms rely on simplistic or probabilistic selection methods, which often result in uneven energy depletion, suboptimal cluster formation, and premature node failures. Additionally, factors such as node residual energy, communication cost, distance to the BS, and network topology are not always adequately considered, leading to inefficient routing decisions. To overcome these limitations, there is a need for an enhanced data mining-based CH selection approach that integrates multiple critical parameters for optimal decision-making. Such an approach should:

1. Ensure balanced energy consumption by dynamically selecting CHs based on real-time network conditions.
2. Optimize cluster formation to improve communication efficiency and minimize redundant data transmissions.
3. Enhance network scalability and adaptability, making it suitable for large-scale and dynamic WSN deployments.
4. Prolong network lifetime by preventing the premature energy depletion of key nodes.

This study proposes an improved clustering-based routing algorithm leveraging the K-means clustering technique while incorporating additional parameters such as residual energy, node density, signal strength, and proximity to the BS. By adopting a more intelligent and data-driven approach to CH selection, the proposed method aims to significantly improve energy efficiency, network sustainability, and overall system performance compared to traditional clustering techniques.

4. PROPOSED CLUSTER HEAD SELECTION ALGORITHM

The proposed methodology aims to assess the lifetime of sensor nodes in terms of the number of operational rounds within a Wireless Sensor Network (WSN). This approach considers the energy depletion of sensor nodes, ensuring that nodes with zero remaining energy are excluded from subsequent rounds of Cluster Head (CH) selection. In each round, the remaining energy of each node is updated by subtracting the energy consumed during the previous round from its initial energy. This iterative process helps optimize energy distribution and prolong network lifespan. The key steps of the proposed approach are outlined in Table 1:

Table 1: Research Methodology for the Proposed Approach

Step	Description
Network Classification	Utilize an adaptive LEACH algorithm integrated with the proposed K-means approach to classify the network into clusters.
Cluster Head Selection	Compute the centroid for each cluster and select the node closest to the centroid with the highest residual energy as the CH.
Packet Size Assignment	Define appropriate packet sizes for efficient communication between sensor nodes and CHs to optimize data transmission and minimize energy consumption.

The initialization phase is a critical step in the proposed system, responsible for setting up the Wireless Sensor Network (WSN). During this phase, each sensor node is assigned an initial energy level, and their positions are randomly distributed across the network to simulate real-world deployment conditions.

The Proposed K-means Initialization Algorithm begins by selecting an initial cluster center randomly from the dataset. For each subsequent center, the algorithm calculates the distance of each data point (sensor node) to its nearest previously selected center. Using these distances, a probability is assigned to each data point, where nodes farther from existing centers have a higher likelihood of being selected. This probabilistic selection process ensures that the new cluster centers are well-spaced and positioned in high-density areas, leading to a more balanced and effective clustering structure. This iterative process continues until all k cluster centers are determined. These centers serve as the starting points for the K-means clustering algorithm, ensuring that the clustering process begins with a diverse and well-distributed set of initial cluster heads. By enhancing the selection of initial cluster centers, this method improves clustering accuracy, optimizes energy consumption, and contributes to the overall efficiency of the WSN.

For cluster formation, we employ the Proposed K-means algorithm, an improved version of the traditional K-means method, which enhances the initialization process for better clustering results. The algorithm begins by randomly selecting the first cluster center from the set of sensor nodes, while subsequent centers are chosen based on a probability function that favours nodes farther from existing centers. This approach prevents the common pitfalls of standard K-means and ensures a well-distributed set of clusters. The initialization process consists of two phases: Center Selection, where initial cluster centers are strategically chosen using a probabilistic method to maximize spatial distribution, and Distance Update, which recalculates the sum of distances between each node and its nearest cluster center to refine clustering accuracy. Once clusters are formed, Cluster Head (CH) selection is performed based on node IDs assigned according to their distance from the cluster centroid.

Algorithm:

```
Input: Set of sensor nodes (N)
Output: Selection of Cluster Heads (CHs)
Initialize all sensor nodes with predefined energy levels
Define K clusters
Randomly initialize K centroids {C1, C2, ..., CK}
While clustering continues:
        For each node Ni in N:
            Find the closest centroid Ck among the K centroids
        For each node Ni in N:
            Generate data for the node
            Compare node’s data with the predefined threshold
            Assign Ni to the nearest cluster Ck
            Enable intra-cluster communication
            Update node's energy level based on transmission and reception
            If Ni's energy == 0:
                Mark Ni as dead and remove it from the cluster list
        For each alive node:
            Process node information
            Select Cluster Head (CH)
Output: Final selection of Cluster Heads
```

5. RESULT AND ANALYSIS

The proposed algorithm is simulated using MATLAB R2023a, which offers advanced functionalities and improved performance for accurate and efficient simulations. The simulation environment is set within a 200×200-meter area, where sensor nodes are randomly distributed to mimic real-world deployment. To ensure realistic evaluation, the experimental parameters are carefully chosen, as detailed in Table 2.

Table 2. Simulation Parameters Overview

Parameter	Value
Number of Nodes (N)	150
Number of Clusters	25
Network Size (m^2)	1000
Initial Energy (E_0)	0.6 J
Probability of CH (P)	0.15
Maximum Number of Rounds (r_{max})	3500
EDA (Data Aggregation Energy)	4×10^{-9} J/bit
EFS (Amplifier Energy for Free Space)	1×10^{-11} J/bit
EMP (Amplifier Energy for Multipath Fading)	1.5×10^{-12} J/bit
Network Time	0.0200 s

The table 5 presented below provides a comprehensive analysis and comparison of the average throughput across three different methodologies: LEACH, HEED, and the Proposed method. The purpose of this comparison is to evaluate the performance of these algorithms in terms of how effectively they manage data transmission in a network, particularly in scenarios with varying numbers of nodes (Table 3).

Table: 3 Comparison of Average Throughput for LEACH, HEED, and the Proposed Method

No. of Nodes	LEACH	HEED	Proposed
50	2.70	2.75	2.85
100	3.40	3.85	4.00
150	3.85	4.20	4.35
200	4.10	4.45	4.60
250	4.30	4.70	4.85

For each node count, the table shows the average throughput in units that reflect the efficiency of data transmission managed by each algorithm. As the number of nodes increases from 50 to 250, the average throughput for each method also rises, indicating improved performance with more nodes. The Proposed method consistently delivers the highest average throughput compared to both LEACH and HEED, demonstrating its superior ability to handle data transmission efficiently as network size grows. This trend underscores the effectiveness of the Proposed method in optimizing network performance in larger and more complex environments. The table 6 presented below provides an in-depth analysis and comparison of the packet delivery ratio across three different methodologies: LEACH, HEED, and the Proposed approach. The packet delivery ratio is a crucial metric that reflects the efficiency of data transmission within a network by measuring the percentage of successfully delivered data packets out of the total packets sent (Table 4).

Table 4: Packet Delivery Ratio for LEACH, HEED, and the Proposed Method

No. of Nodes	LEACH	HEED	Proposed
50	1.15	1.25	1.35
100	1.18	1.40	1.55
150	1.20	1.65	1.85
200	1.23	1.90	2.10
250	1.25	2.10	2.30

As the number of nodes increases from 50 to 250, the packet delivery ratio for all three algorithms improves, reflecting enhanced data transmission efficiency. The Proposed method consistently achieves the highest packet delivery ratio at each node count, indicating its superior performance in ensuring successful data delivery compared to LEACH and HEED. This improvement highlights the effectiveness of the Proposed method in managing network traffic and reducing packet loss, particularly in larger networks where maintaining high delivery ratios becomes increasingly challenging.

The table 5 presented below offers a detailed analysis and comparison of the number of dead nodes observed in two different methodologies: CH-LEACH and the Proposed approach. The number of dead nodes is a critical indicator of the network’s overall health and longevity, as it reflects the nodes that have exhausted their energy and can no longer participate in the network's operations.

Table 5: Number of Dead Nodes for CH-LEACH and the Proposed Method

No. of Rounds	CH-LEACH	Proposed
1000	18	15
2000	35	28
3000	57	42
4000	78	65
5000	95	80

As the number of rounds increases from 1000 to 5000, the number of dead nodes rises for both methods, indicating a higher rate of node depletion over time. However, the Proposed method consistently results in fewer dead nodes compared to CH-LEACH at each round count. This suggests that the Proposed method is more effective in extending the operational life of nodes and maintaining network functionality, thus demonstrating improved energy efficiency and network longevity.

6. DISCUSSION

The performance analysis of the LEACH, HEED, and Proposed algorithms across different metrics Average Throughput, Packet Delivery Ratio, and the Number of Dead Nodes demonstrates the effectiveness of the Proposed method in enhancing the overall efficiency of Wireless Sensor Networks (WSNs).

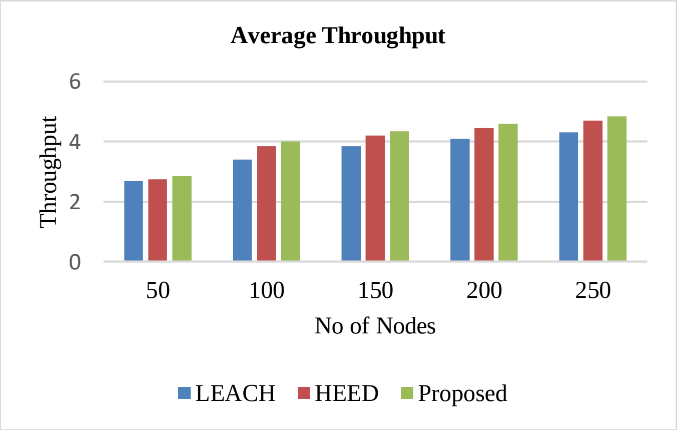

Figure: 1 Comparative analysis of Average Throughput

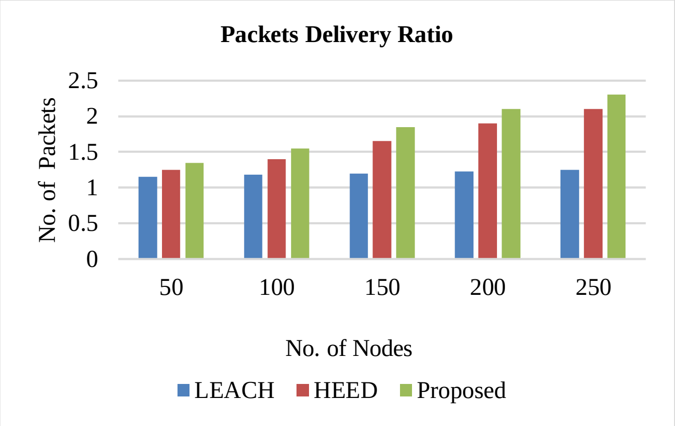

Figure 2: Analysis of Packet Delivery Ratio

Average Throughput

Figure 1 compares the average throughput for varying numbers of nodes using the LEACH, HEED, and Proposed methods. As the number of nodes increases from 50 to 250, the average throughput improves across all three algorithms. However, the Proposed method consistently outperforms both LEACH and HEED. For instance, with 250 nodes, the Proposed method achieves an average throughput of 4.85 units, compared to 4.70 units for HEED and 4.30 units for LEACH. This superior performance suggests that the Proposed algorithm is more effective in managing data transmission, particularly in larger networks, by optimizing the clustering process and reducing data loss.

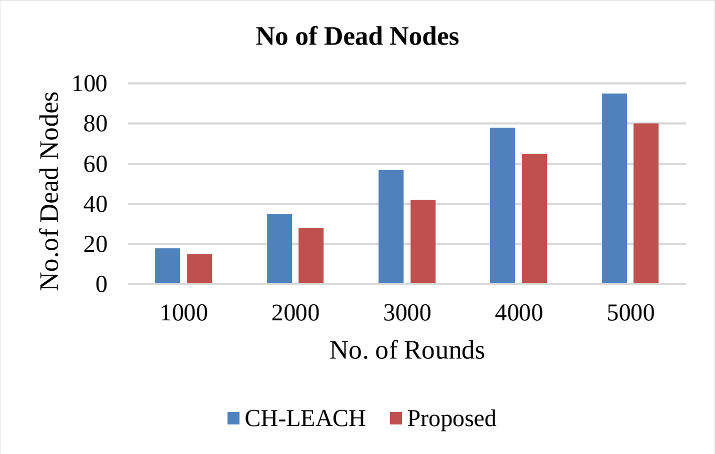

Figure 3: Comparison of Number of Dead Nodes

Packet Delivery Ratio

Figure 2 focuses on the packet delivery ratio, which measures the success rate of data transmission from nodes to the base station. Similar to the throughput analysis, the packet delivery ratio improves as the number of nodes increases. The Proposed method consistently achieves higher delivery ratios compared to LEACH and HEED. For example, with 250 nodes, the proposed method records a packet delivery ratio of 2.30, significantly higher than HEED's 2.10 and LEACH's 1.25. This indicates that the Proposed method is more effective in ensuring reliable data delivery, reducing packet loss, and maintaining network stability, especially as network size grows.

Number of Dead Nodes

Figure 3 presents the number of dead nodes observed over increasing rounds of network operation for CH-LEACH and the Proposed method. The results show a steady increase in the number of dead nodes as the network operates for more rounds, which is expected due to the depletion of node energy over time. However, the Proposed method consistently results in fewer dead nodes than CH-LEACH. For instance, after 5000 rounds, the Proposed method has 80 dead nodes compared to 95 dead nodes for CH-LEACH. This demonstrates the enhanced energy efficiency of the Proposed method, which prolongs the operational life of the nodes and the overall network, reducing the rate at which nodes exhaust their energy.

The network shows low network lifetime in HC initially because the number of nodes are less. Less nodes amount to lower network lifetime. With the enhancement incorporated we see that the initial network lifetime i.e. with less number of nodes has improved because there was a better distribution of nodes under each cluster head.

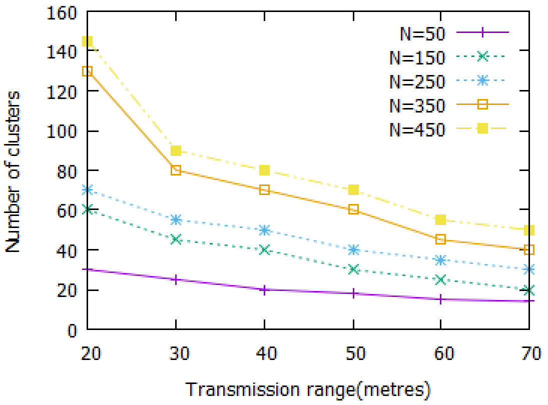

Figure 4 Transmission range against number of clusters

Figure 4 shows that the average number of clusters is relatively high when the transmission range is small. The results shown are for varying values of total number of nodes. When the transmission range increases, more and more nodes are connected to the same cluster head resulting in reduced number of clusters created. A smaller backbone is desirable for minimizing the routing overhead. Hence, transmission power of a node is also a deciding factor for finding the quality of dominating nodes. When the transmission range is increased from 20 to 40 m, the number of clusters created is reduced considerably. But the rate of reduction in the number of clusters created gets reduced on further increase in the transmission range. The power consumption is high for higher transmission range. Hence, the recommended value of transmission range is between 30 and 40 m.

7. CONCLUSION

This study introduces an innovative clustering algorithm tailored for Wireless Sensor Networks (WSNs) that significantly optimizes the process of cluster head selection and overall network management. The proposed algorithm effectively addresses common challenges such as uneven node distribution, biased cluster head elections, and accelerated energy depletion among cluster heads. By incorporating load-balancing techniques and ensuring equal opportunities for all nodes to serve as cluster heads, the algorithm enhances network efficiency and prolongs node lifespan. Simulation results validate that the proposed approach surpasses existing clustering schemes by reducing the number of clusters while maintaining network stability and minimizing communication overhead, ultimately leading to better scalability and more efficient power management. The study successfully implements a novel cluster-based routing protocol that enhances the lifetime of WSNs. The proposed K-means algorithm is utilized to form clusters in a manner that minimizes power consumption during communication between sensor nodes. Additionally, features from Adaptive LEACH are integrated to optimize the hop count of transmitted data during the transmission phase from cluster heads to the Base Station (BS), thereby balancing energy consumption and extending network survival time. The simulation results demonstrate that the proposed K-means routing protocol outperforms the LEACH, HEED algorithms by improving average throughput, increasing the number of transmitted data packets, and extending the overall network lifetime.

8. References

[1] Nor Surayati Mohamad Usop, Azizol Abdullah, Ahmad Faisal Amri Abidin. Performance Evaluation of AODV, DSDV & DSR Routing Protocol in Grid Environment. IJCSNS, VOL.9 No.7,pp 261-268, July 2009.

[2] Conti, M., Giordano, S.: 'Multi-hop adhoc networking: the theory', IEEE Commun. Mag., 2007, 45, (4), pp. 78–86

[3] Chlamtac, I., Conti, M., Liu, J.-N.: 'Mobile adhoc networking imperatives and challenges', Ad Hoc Netw., 2003, 1, (1), pp. 13–64

[4] Akkaya, K., Younis, M.: 'A survey on routing protocols for wireless sensor networks', Elsevier J. Ad Hoc Netw., 2005, 3, (3), pp. 325–349

[5] Wu, J., Li, H.: 'On calculating connected dominating set for efficient routing in ad hoc wireless networks. Proc. Third ACM Int. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications, pp. 7–14.

[6] Sunil Taneja, Ashwani Kush,A Sur vey of Routing Protocols in Mobile Adhoc Networks‖, International Journal of Innovation, Management and Technology, Vol. 1, No. 3, August 2010.

[7] Azzedine Boukerche, Athanasios Bamis, Ioannis Chatzigiannakis, Sotiris Nikoletseas. A mobility aware protocol synthesis for efficient routing in ad hoc mobile networks. The International Journal of Computer and Telecommunications Networking, Vol. 52, Issue 1, pp 130-154, Jan, 2008.

[8] Geon yong Park, Heeseong Kim, Hwi Woon Jeong, and Hee yong youn, "A Novel cluster head selection method based on k-means algorithm for energy efficient wireless sensor network", IEEE 27th international conference on advanced information networking and applications workshops,pp.910-915,2013.

[9] Seifemichael B.Amsalu*, Wondimu K.Zegeye, Dereje Hailemariam, Yacob Astatke, "Design and performance evaluation of an energy efficient routing protocol for wireless sensor networks". IEEE Annual conference on information science and systems, 2016.

[10] Krishnakumar A, Dr. Anuratha V, "An energy-efficient cluster head selection of LEACH protocol for wireless sensor networks", IEEE, International conference on Nextgen Electronic Technologies, pp.57-61, 2017.

[11] Hairong Zhao, Wuneng Zhou, Yan Gao , "Energy Efficient and cluster-based routing protocol for WSN", IEEE Eight International conference on computational intelligence and security, pp.107-111, 2012.

[12] Yang yang, Qian liu, Zhipeng gao, Xuesong Qui, and LanlanRui, "Data clustering-based fault detection in WSNs", IEEE 7th International conference on advanced computational intelligence, pp.334-339, 2015.