CHAPTER 3

CLUSTERING

3.1 INTRODUCTION

Sensor nodes are grouped into clusters by a process called clustering. A CH is the head of each cluster. The selection of CH is a major challenge in WSN. CH selection must be in such a way that it consumes less energy with less intra-cluster, and less inter-cluster distance. Some processing, like data aggregation, data processing, long distance communication are carried out by CH, which consumes extra energy compared to the other nodes in the network. Therefore, to extend the network's lifespan and improve its energy efficiency, optimal CH selection is necessary.

Clustering is of two types. They are:

1. Homogeneous clustering
2. Heterogeneous clustering

- **Homogeneous clustering**

In homogeneous clustering, every node in the network would have the same capabilities in terms of communication, computation, memory, energy, reliability, and other aspects.

- **Heterogeneous clustering**

In heterogeneous clustering, the network consists of different compositions of sensors with different capabilities.

3.1.1 General Framework

The main goal of clustering in WSN is to extend the network's lifespan. It is done in two main stages: cluster formation and CH selection.

- **Cluster formation** guarantees that each cluster has a few members, which lowers the burden on CHs close to the nodes. Every member node is assigned to the closest CH based on the Received Signal Strength Indication (RSSI).
- **CH selection** is used mainly for data aggregation and distribution to the nodes. The CH selection process must be used to prolong the network's lifespan. The

distances between the CH and nodes, the CH and base station, the node and base station, residual energy, RSSI, node degree, cluster density, and location are the metrics used to choose the CH. Communication between nodes and CHs, as well as between CHs and the base station, can be formed in a single-hop or multiple-hop fashion. In addition, load balancing and energy efficiency are ensured by CH.

3.1.2. Characteristics of Clustering

Certain characteristics of clustering techniques use the internal structure of the cluster to classify various clustering protocols.

- **Inter-CH connectivity:**Indicates how well a CH can communicate with the base station.
- **Cluster Count**:Refers to the number of clusters formed in every iteration; the more the quantity of CHs, the smaller the cluster distribution size and the better the energy conservation.
- **Cluster Size:** The path length between each node and the cluster's distance from CH are indicated by the cluster size. The energy consumption improves with smaller cluster sizes, and CH load and transmission distance are also efficiently decreased.
- **Cluster Density:**The quantity of common nodes inside a cluster is referred to as cluster density.
- **Message Count:**The number of messages transmitted to choose a CH is referred to as the "message count".
- **Stability:** If the cluster members are not fixed, stability indicates that the clustering schemes are adaptive; if not, the cluster count cannot be changed throughout the CH selection process, and the cluster members are regarded as fixed.

3.1.3 Challenges in Clustering

Low computational capacity, constrained bandwidth, and restricted battery life are a few problems and obstacles with communication. Similar energy levels, consistent computing power, and memory are the major issues with clustering

techniques. Since the nodes in heterogeneous networks have varying bandwidth, energy, processing, and computing capacities, this issue does not arise.

Energy: Most of the energy utilized by sensor nodes are used for computational and communicational tasks.

Node deployment: In WSN, deployment of node can be done either manually or at random. Different deployment approaches are done manually. In sensing area of WSN,the sensor nodes would be randomly positioned according to the requirement.

Coverage: It is the actual area that the deployed nodes cover, and it shows how well the sensors can monitor the intended area.

Data Aggregation: In data aggregation, when nodes aggregate data, they will send it to the CH. However, due to node density, additional nodes would send the same data, creating redundancy at the CH.

Fault Tolerance: When one node fails, it impacts the network's overall efficacy. Fault tolerance must therefore be incorporated into WSN protocols.

Location: It identifies the location of each node inside a cluster. This is an expensive approach that is not suitable for all applications.

Network Dynamics: Some applications use static nodes that cannot be relocated after they have been deployed. Some applications feature base stations and flexible nodes.

3.2 OPTIMIZATION METHODS IN CLUSTERING

The procedures for CH selection, data gathering, cluster formation, and data transfer are all optimized in the clustering process. During CH selection, it is critical to ascertain the ideal number of clusters and the cluster balancing. Since data aggregation and communication are interdependent, it is important to optimize both by choosing the right cluster size and level of inter-cluster communication distance. Figure 3.1 addresses some optimization methods in clustering.

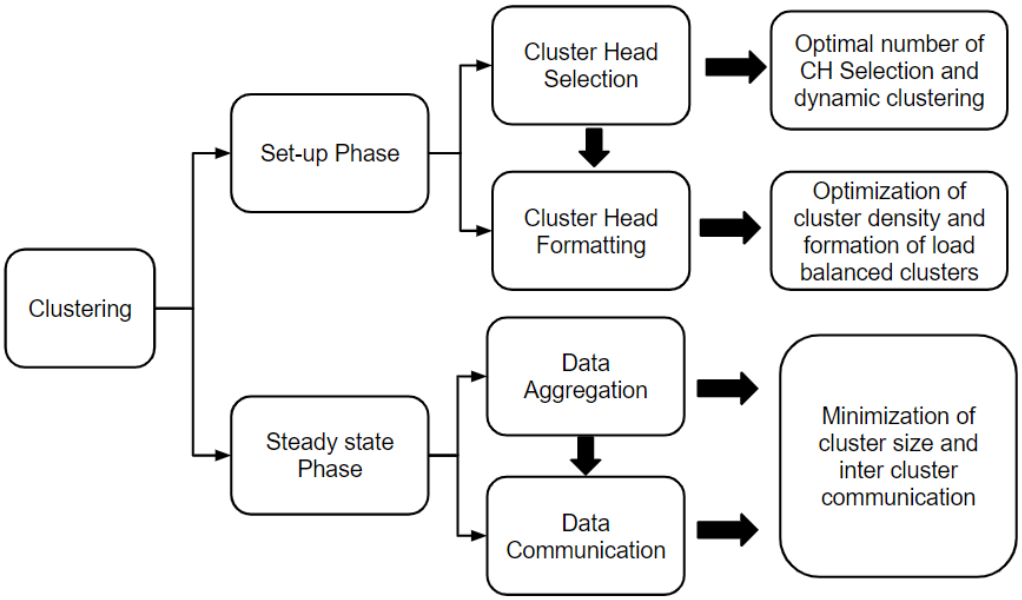

Figure 3.1 Optimization Methods in Clustering

CH selection phase: Since the CH serves as a gateway between the nodes and the base station, choosing the CH plays a vital role and it is the first phase in cluster formation. Choosing the right CHs is essential to extend the network lifespan and enhance energy efficiency. CH is selected based on some methods namely,

- Based on probability clustering optimization
- Based on non-probability clustering optimization
- Base station assisted clustering optimization
- CH assisted clustering optimization

Cluster creation phase: At the end of this phase, each node sends a join message to its optimal CH. This phase started with the newly elected CH announcing their current position. There are various types of cluster creation schemes:

- Optimal clustering
- Event clustering

Data aggregation phase: The purpose of gathering data from several sensor nodes is to eliminate redundancy during transmission and provide fused information to the base station. In data aggregation, there are three approaches:

- Tree based data aggregation
- Cluster based data aggregation
- Multipath based cluster aggregation

Data communication phase: Aggregated data is sent to the faraway base station from CH for further processing. Data would be communicated in the following ways:

- Communication within clusters
- Communication between clusters

3.3 NATURE INSPIRED OPTIMIZATION TECHNIQUES

Natural phenomena and certain evolutionary techniques provide inspiration for nature-inspired clustering algorithms, which also offer computational intelligence for tackling practical issues. Evolutionary computation is a unique method derived from natural abilities, while swarm intelligence is derived from the behaviours of insects, birds, and other small animals. Inspiration from the behavior of nature is forced to enhance many complicated algorithms. Some of the nature inspired optimization techniques are discussed below.

3.3.1 Ant Colony Optimization (ACO)

ACO determines a dependable path from the source to the destination. This technique emulates the foraging behaviour of real ants. Ants first exhibit random movement patterns when foraging for food. Ants deposit pheromones as they go, and these chemicals dissipate over time. The pheromone serves as a trail marker for ants to follow, attracting those with the highest pheromone levels. This method identifies the near-optimal solution and the shortest path for transmission in WSN.

3.3.2 Particle Swarm Optimization (PSO)

PSO promotes the behaviour of birds flocking together. They look for food haphazardly and locate the nearest food source. Birds move in groups without colliding due to each member following their group leader and adjusting its position and pace accordingly. Birds also communicate with each other regarding their position and the whereabouts of food. Each bird within the solution space is referred to as a particle. Every particle is assigned a fitness value that assesses the solution's quality.

3.3.3 Artificial Bee Colony Optimization (ABC)

The ABC algorithm is a meta-heuristic that draws inspiration from honey bee foraging behaviour. Three groups make up the colonies of artificial bees: workers, scouts, and observers. Once they have located a food source, employed bees look for food and interact with other bees. In the dancing area, onlooker bees are waiting for food. The amount of nectar affects how long the dance lasts. High nectar content positions will draw more bees, and many other bees will follow that trail. The ABC algorithm uses the same behaviour.

3.3.4 Bacterial Foraging Optimization (BFO)

BFO is modeled after a community of foraging bacteria. This approach imitates how bacteria travel in pursuit of nourishment. The method of gradually looking for nutrients in the surroundings is known as "chemo taxis." The bacteria in BFO represent the solution, and the amount of nutrients reflects the fitness value. This behaviour is implemented in the same way.

3.3.5 Firefly Optimization Algorithm (FFO)

The novel optimization method known as FFO imitates how actual flies are drawn to one another by flashlight. Brightness and attractiveness are inversely related. The distance grows as the light drops. Attractiveness is related to the objective function. The Euclidian distance between two flies and their residual energy are the basis for fitness.

3.4 Low Energy Adaptive Clustering Hierarchy (LEACH)

Clustering and routing methods have been inspired by the fundamental concepts of LEACH.At the end of each iteration, LEACH aims to find the CH among all nodes.Therefore, every node in the network experiences considerable energy dissipation when communicating with the base station. Data is sent to the CH using standard sensor nodes. To lower the cost of redundant data transmission, CH combines the data sent by the regular node and transmits it to the sink. The LEACH operation involves multiple rounds.In each round, there are two distinct phases, the setup phase and the steady state phase. Clusters are arranged during the setup phase,

and data is sent to the base station during the steady state phase. During the setup phase, every node decides whether to be the CH for the current round based on how many CHs are in that network and how many times it has been the CH previously. Each node selects a random number from 0 to 1. If the selected number is below the threshold, the node becomes the CH for this round. Equation 3.1 represents the threshold value

.

$$T(n) = \begin{cases} \frac{P}{1-P\left[r \bmod \frac{1}{P}\right]}, & if n \in G \\ 0, & otherwise \end{cases} \quad (3.1)$$

Upon being elected as CH, a node transmits an advertisement message to the remaining nodes. Nodes are assigned to a specific cluster based on the strength of the received signal, and then they transmit a membership message to the CH. CH is chosen for each round to evenly spread the energy demand among the sensor nodes. During the steady-state phase, sensor nodes detect the data and send it to the CH. CH combines the data and transmits it directly to the base station. LEACH utilizes Time Division Multiple Access (TDMA) or Code Division Multiple Access (CDMA) to minimize collisions inside and across the clusters.

Advantages of LEACH:

- Each node shares its load as a CH equally.
- The TDMA scheduling minimizes collisions in the CH.
- Member nodes activate or deactivate communication during specific time windows to reduce excessive energy consumption.

Disadvantages of LEACH:

- Not suitable for broad area networks due to single-hop inter-cluster routing.
- LEACH does not guarantee actual load balancing because CHs are chosen based on probabilities rather than residual energy.

- It is hard to distribute the CHs uniformly in a predetermined manner since the CH selection is based on probabilities.

3.5 PROPOSED METHOD

Transmission of data in WSN requires more energy than processing. Nodes engaged in long-distance communication spend more energy than those involved in short-distance communication. Consequently, the finite energy resources provide challenges in sustaining energy levels across the network. Efficient transmission requires selecting the optimal route and nodes with ample energy. This study introduces a GA-based energy optimization approach that overcomes the mentioned limitation and identifies an appropriate CH, resulting in cluster formation.

3.5.1 Genetic Algorithm Overview

Darwin's theory of natural evolution is adhered to by the heuristic search algorithm known as the GA. Parents in GAs are chosen depending on their ability to have kids. The offspring with the lowest fitness level in the population is replaced by a new one. Three operators are used by the GA: crossover, mutation, and selection.

- **Selection:** Classification of chromosomes is based on their fitness function and the one with the best fitness is selected.
- **Cross Over:** The genetic qualities of two parents are combined to provide a new offspring.
- **Mutation:**Search space is diversified by randomly modifying new individuals

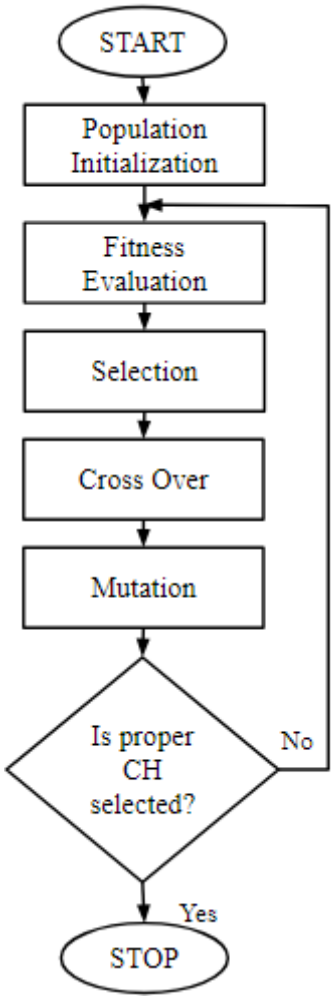

Figure 3.2 Flow chart of Genetic algorithm

As soon as the new generation is generated, the algorithm looks for a termination state. When the termination condition is met, the algorithm stops. If not, the cycle will persist until the termination condition is met. Figure 3.2 illustrates how the GA operates. GA is widely recognized as a suitable approach for addressing optimization problems in various applications. Cluster creation and CH selection are crucial in reducing the energy consumption of the network in WSN.

3.5.2 Model of the Network

Sensor nodes are distributed randomly in the target area for data acquisition and must exhibit specific features.

1. Nodes are not movable.
2. The base station is situated either within or outside the field.

3. Power in the radio model is adjusted based on the transmission distance.

Figure 3.3 represents the radio model used in WSN. The radio model utilized is the free space model, as stated below. The energy dissipated to run the circuit is 60nJ per bit ($Eelec$), while the energy dissipated by the transmission amplifier is 10pJ per bit per square meter (φamp).

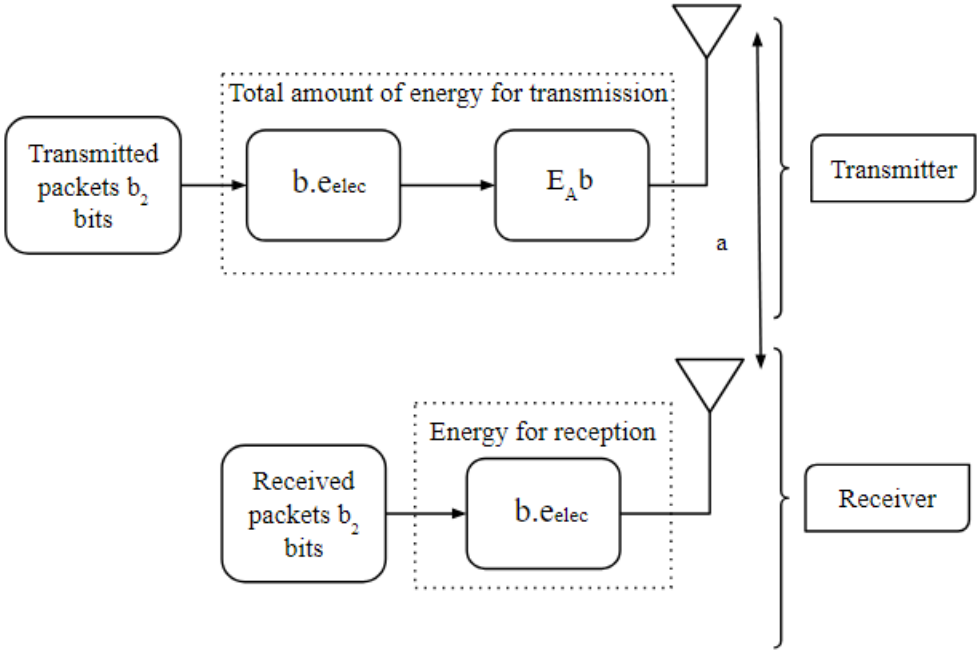

Figure 3.3 Radio model for WSN transmission

Energy for transmitting and receiving data is calculated using equations 3.2 and 3.3 and equation 3.4 represents the energy required for acknowledgement of packet exchange.

$$ETr(m,d) = Eelec * m + \varphi amp * m * d2 \quad (3.2)$$

$$ERe = Eelec * m \quad (3.3)$$

$$EACK = \tau ACK\ (ETre(m,d) - ERe) \quad (3.4)$$

where$\tau ACK = \eta ACK/n$ is the ratio of packet length to packet acknowledgment.

The resultant energy required is calculated using the equation 3.5,

$$Totalenergy = Etotal = ETr\,(m,d) + ERe - EACK \quad (3.5)$$

$Eelec$ and $\varphi amp$are constants, while m denotes the number of bits transmitted or received,$d$represents the distance between the sending and receiving antennas, whereas $EACK$ stands for the energy expenditure associated with the acknowledgment packet exchange.

3.5.3. Proposed Algorithm

This algorithm combines the GA and Algorithm for Cluster Establishment (ACE) for selecting the CH and establishing the clusters with minimal energy consumption. The hybrid algorithm would give the chance for all the nodes to become CH based on the fitness value and the number of neighboring nodes. The overall process of the algorithm is given in Figure 3.4.

Step 1: Random generation of the initial population.

Step 2: With respect to energy, distance, and the number of alive nodes the fitness is calculated.

i) The selection of the parent node is obtained based on the probabilisticmethod.
ii) A single-point cross over operator is applied over it.
iii) If necessary, a mutation operator would be applied.

Step 3: The new population is updated based on the generation of new offspring.

Step 4: Depending on the best-fit chromosome, CH selection and cluster formation are carried out.

Step 5: End.

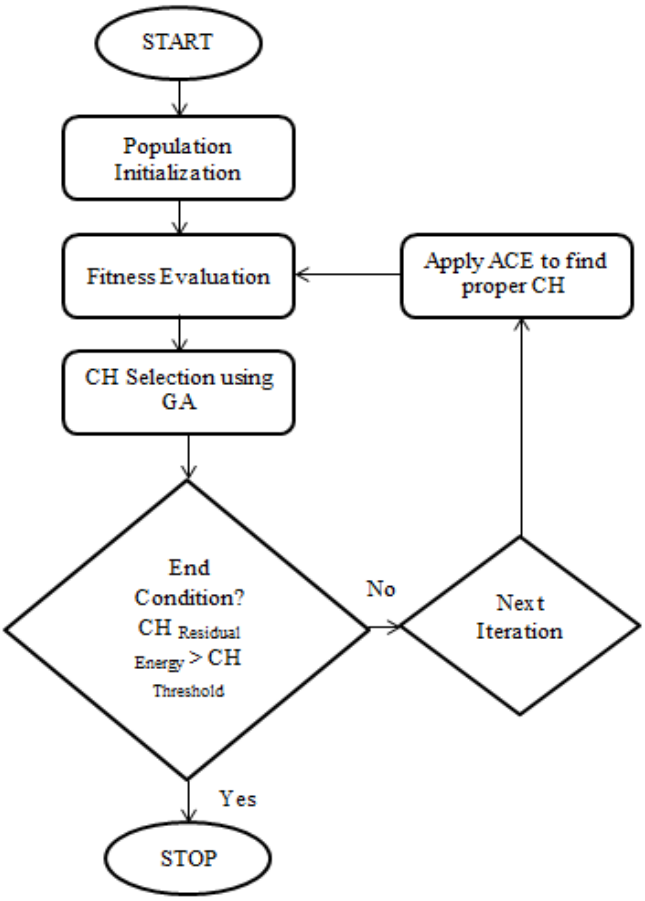

Figure 3.4 Flow chart of the proposed work

- **Initialization of Population:**

The total number of nodes is regarded as the initial population, which is created using a random number technique, where 0 denotes a member node in the population and 1 represents the CH chromosome.

Fitness calculation:

Fitness is determined by residual energy, distance, and the number of active nodes.

Equation 3.6 represents the fitness function.

$$F(CH) = \{E * sin(mA)\}/\{En + Dn\} \quad (3.6)$$

$En =$ Node energy

$Dn =$Node-to-node distance

$mA =$ Active nodes quantity

$E =$ Total energy of all active nodes

The chromosome with the highest fitness value is chosen as the candidate solution for the current iteration. The number of CHs is determined depending on the energy considerations. Once a CH is chosen, member nodes join the clusters based on their proximity to the CH. Data is captured by nodes and transferred to CHs for aggregation and delivery to the base station.

Parent Selection:

Two chromosomes are chosen for the next generation based on probabilistic factors. The Roulette wheel approach is used to undertake crossover operations on the selected chromosomes to generate new progeny.

- **Cross over:**

From the selected parents, a single point cross over is applied to generate the new population. The size of the parents and offspring are the same.

Example: PARENT 1: 11010|111

PARENT 2: 01101|001

↑

Cross over point

First off-spring: 11010001

Second off-spring: 01101111

- **Mutation:**

Mutation is performed to change the value of a gene to preserve genetic diversity. It is utilised to preserve and introduce variation in the genetic population and is often implemented with a low likelihood.

Example: Offspring after cross over: 11010111

| |

Offspring after mutation : 11110110

In every iteration, the ACE would select the new CH. ACE utilises feedback on the number of neighbouring nodes, with each node functioning as a CH just once.

3.6 RESULTS AND DISCUSSION

Work is done using the simulation tool MATLAB 2020a. Nodes are randomly distributed in a square space $(N * N)m^2$ as shown in figure 3.5. Simulation is

performed using 100 nodes in an area of $(100\ x\ 100)m^2$. Sensor nodes are distributed throughout the target region, forming 10 clusters, each led by a CH. After each repetition, the value of CH is altered, resulting in the formation of a new cluster. Table 3.1 contains the simulation parameters.

Table 3.1 Simulation Parameters

Parameter	**Value**
Number of nodes	100
Field Area	(100 x 100)m^2
Optimal number of clusters	10
Number of iterations	1400
Cross over percentage	0.5
Mutation rate	0.01

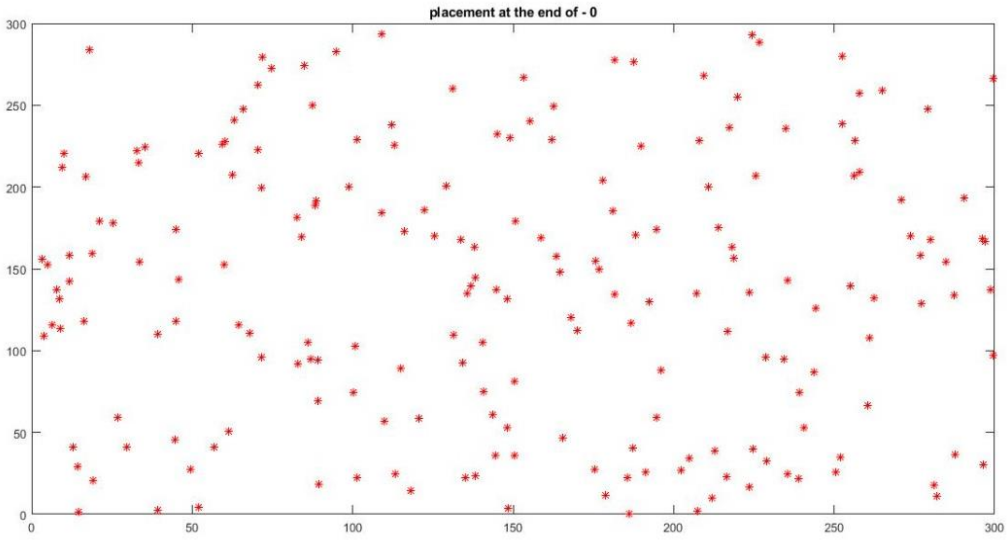

Figure 3.5 Random distributions of sensor nodes in the target area

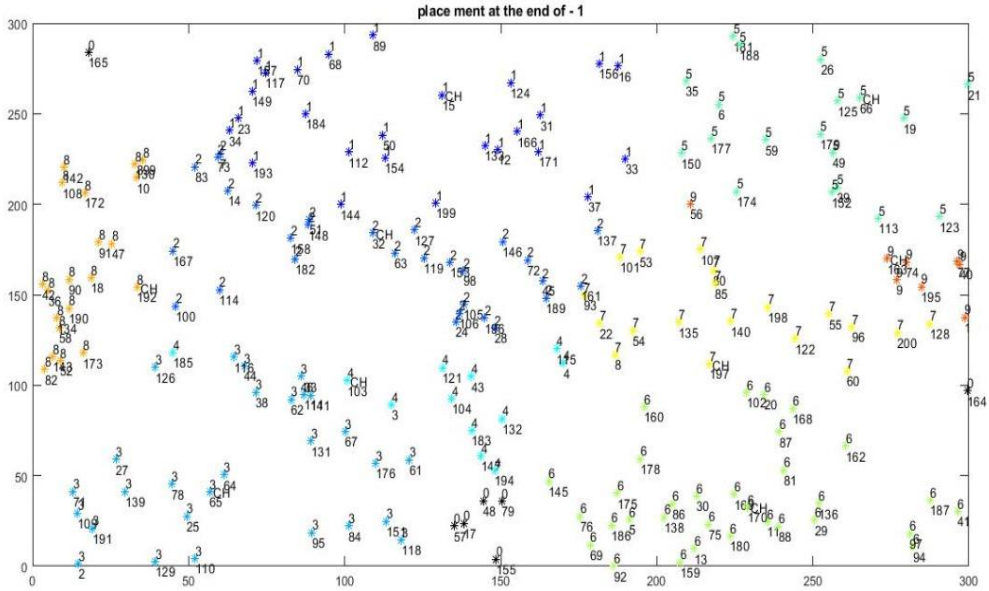

Figure 3.6 Cluster formation and CH selection after the initial round

The sink node is situated at a considerable distance from the member nodes. GA-based CH selection and cluster formation are contrasted with the LEACH protocol. Figure 3.6 illustrates the cluster development and selection of CH after the initial iteration.

Table 3.2 gives the comparison of the residual energy of the network with the LEACH protocol and the GA based clustering. Simulation is carried out for 1400 rounds. The initial energy is 50J. In 200 iterations, the residual energy using the LEACH protocol is 44J whereas the residual energy using the GA is 45J. It gradually reduces and at iteration 1200, the residual energy using the LEACH protocol is 1J whereas the residual energy using the GA is 6J which is around 12% of the initial energy and in the case of LEACH protocol, the residual energy is around 2%. Figure 3.7 represents the comparison of the residual energy graph.

Table 3.3 compares the number of active nodes with the LEACH protocol and Genetic Algorithm for calculating the network lifetime. Initial number of nodes are 100. Finally, at the end of 1200 iterations, 83% of nodes are alive in using the GA whereas only 53% of nodes are active in LEACH protocol.

Figure 3.8 demonstrates that the proposed approach consistently has a greater number of active nodes compared to LEACH at each iteration. The proposed algorithm has a more consistent lifespan compared to LEACH. Choosing the appropriate CH and forming clusters enhances the system's stability. GA-based approaches involve 156 additional rounds compared to LEACH.

Table 3.2: Comparison of residual energy with the LEACH protocol and the proposed work

Iterations	Residual Energy (J)	
	LEACH	Proposed work (GA based clustering)
200	44	45
400	33	35
600	25	27
800	15	20
1000	9	12
1200	1	6

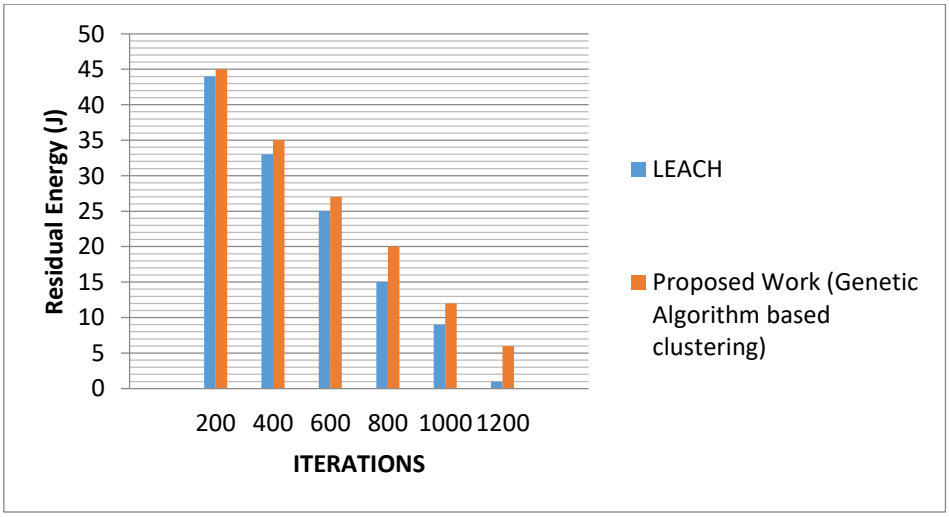

Figure 3.7 Comparison of residual energy with the LEACH protocol and the proposed work

Table 3.3 Comparison of Network lifetime (Number of alive nodes)

Iterations	Network Life (Number of Alive Nodes)	
	LEACH	Proposed work (GA based clustering)
0	100	100
200	100	100
400	100	100
600	100	100
800	99	100
1000	89	100
1200	70	99
1400	53	83

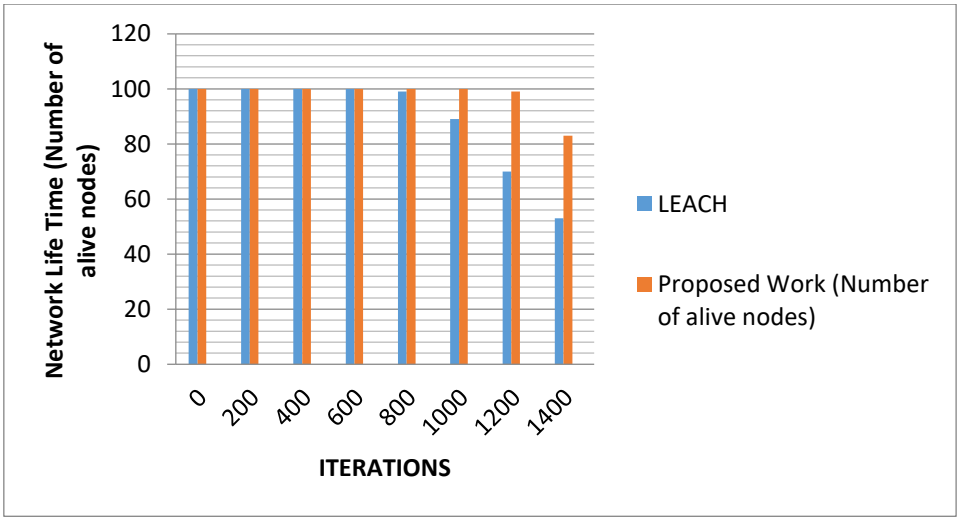

Figure 3.8 Comparison of Network lifetime (Number of alive nodes)

Table 3.4 compares the intra-cluster throughput. Initially, 2000 packets are transmitted and at the end of 1400 iterations, 300 data packets are delivered to the CH while using the GA whereas while using the LEACH protocol only 100 packets reach the CH. As per the results, more than 10% of packets reach the destination by the proposed work when compared with the existing method.

Table 3.4 compares the intra-cluster throughput. Initially, 2000 packets are transmitted and at the end of 1400 iterations, 350 data packets are delivered to the CH while using the GA whereas while using the LEACH protocol only 100 packets reach the CH. At the end of 1400 iterations, 17.5% of total packets are transmitted using the Genetic Algorithm.

The intra-cluster throughput and inter-cluster throughput are higher than LEACH because of reduced overhead, as shown in Figures 3.9 and 3.10.

Table 3.4 Comparison of intra-cluster throughput

Iterations	Intra Cluster Throughput (Number of packets delivered to the CH)	
	LEACH	Proposed work (GA based clustering)
0	2000	2000
100	1500	1500
200	1100	1200
400	1000	1100
600	900	1000
800	700	900
1000	600	700
1200	400	500
1400	100	300

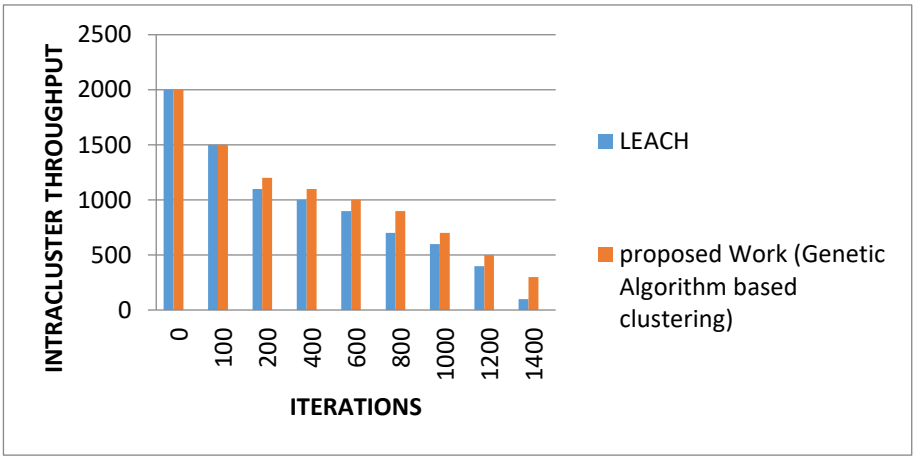

Figure 3.9Number of packets delivered to CH (Intra cluster Throughput)

Table 3.5 Comparison of inter cluster throughput

Iterations	Inter-Cluster Throughput (Number of packets delivered to the base station)	
	LEACH	**Proposed work (GA based clustering)**
0	2000	2000
100	1500	1500
200	1500	1500
400	1400	1500
600	1300	1400
800	1000	1200
1000	600	700
1200	400	500
1400	100	350

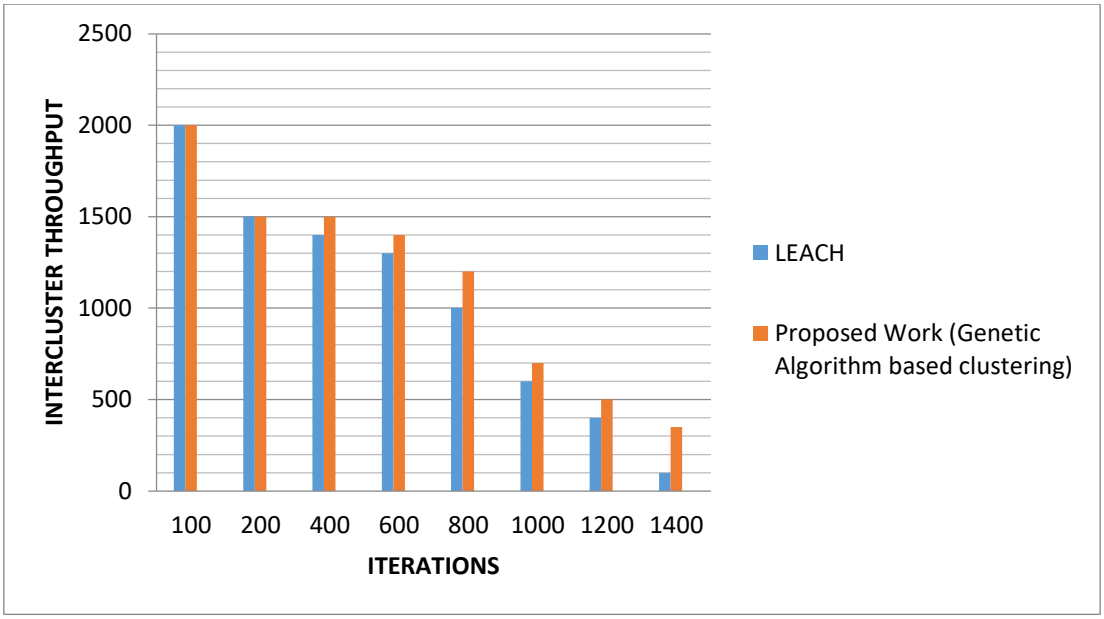

Figure 3.10 Number of packets delivered to Base Station

3.7 CONCLUSION

The Algorithm for Cluster Establishment (ACE) and Genetic Algorithm (GA) are used in this work, and the outcomes are compared with the LEACH procedure. Based on the results, the proposed algorithm for clustering performs better compared to the LEACH technique.