

# CLUSTERING BASED ON HYBRIDIZATION OF GENETIC ALGORITHM AND IMPROVED K-MEANS (GA-IKM) IN AN IOT NETWORK

Moez Elarfaoui <sup>1</sup>, Nadia Ben Azzouna <sup>2</sup>

<sup>1</sup> University of Tunis, ISG Tunis, SMART LAB, 2000 Bardo, Tunisia

<sup>2</sup> University of Tunis, ESSECT Tunis, SMART LAB, Tunisia

## ABSTRACT

*The continuous development of Internet infrastructures and the evolution of digital electronics, particularly Nano-computers, are making the Internet of Things (IoT) emergent. Despite the progress, these IoT objects suffer from a crucial problem which is their limited power supply. IoT objects are often deployed as an ad-hoc network. To minimize their consumption of electrical energy, clustering techniques are used. In this paper, a centralized clustering algorithm with single-hop routing based on a genetic algorithm and Improved k-means is proposed. The proposed approach is compared with the LEACH, K-means and OK-means algorithms. Simulation results show that the proposed algorithm performs well in terms of network lifetime and energy consumption.*

## KEYWORDS

*IoT – Sensor Network - Clustering - LEACH - Genetic algorithm - K-means - Optimization - Energy*

## 1. INTRODUCTION

The Internet of Things is a field with a full effervescence, it has invaded many domains. Its applications are found in agriculture, the military, transport, and many other domains. The number of IoT devices is continuously growing and billions of them will be linked to the Internet [1], [2]. These equipments present serious challenges such as deployment, security, mobility, and especially the consumption of electrical energy. Indeed, the majority of them are often battery-powered and it is almost impossible to replace the battery in harsh environments [3]. When they are deployed as a Wireless Sensor Network (WSN), IoT objects suffer from a lifespan limited by the capacity of the batteries. Among the approaches used to minimize the energy consumption of these devices and therefore the lifetime of the entire network is clustering [4]. This technique consists of dividing all the nodes of the network into clusters. Each of these clusters is managed by a particular node, which is often called the Cluster Head (CH). This CH takes care of the organization of the communication inside the cluster, the collection of data from the member nodes, and finally the sending of these packets, after aggregation, to the base station. The network of IoT objects will therefore be reduced to a network of CH nodes. Increasing the lifetime of such a network amounts to forming clusters that optimize energy consumption in the entire network.

Clustering techniques as well as routing techniques have a considerable influence on energy consumption. Routing consists of finding the optimal path passing through the CH nodes to send a packet from a node to the destination (Base Station).

The clustering procedure can be carried out within the base station (centralized clustering) or by cooperation between the nodes (distributed clustering). Similarly, the routing procedure, in turn, and according to the same principle, can be centralized or distributed [5], [6], [7].

Also, the establishment of the path, for the routing of the packets, can be in mono-hop or multi-hop. In mono-hop routing, the packets are sent directly from the CH node to the base station without any intermediate nodes. While in multi-hop routing, the packets are emitted from a CH node to another until the BS.

The clustering procedure is an NP-hard problem [8]. As a result, meta-heuristics and artificial intelligence techniques are often used to provide approximate solutions [9].

In this paper, the problem of centralized clustering in an IoT network with a mono-hop routing is invoked. The purpose of the proposed solution is to maximize the lifetime of the network. Therefore, our contribution focuses on the following points:

- a) Calculating the number of clusters to be built: the number of CH nodes influences the overall energy consumption in the network. A smaller number of CH nodes generate more communication load between the CH and its member nodes, while a very large number generates more traffic in the network.
- b) Election of suitable CHs: The election of the CH node must take into account the amount of energy, its distance from the base station, and the load balancing between the CH nodes.
- c) Studying the BS coordinates in the area to be monitored: In many WSN studies, the choice of position is arbitrary and not well argued [12,13,14,15]. In this work, the influence of the base station position on the network lifetime extension is studied.

The remainder of the paper is structured as follows. Section 2 presents the related works in clustering and some kinds of hybridization. Section 3 details the proposed approach then, in Section 4, the simulation results are discussed. Finally, in the last section, the conclusion is elaborated and different perspectives are presented

## **2. RELATED WORKS**

The clustering procedure is a NP-problem and can encompass one or more objects simultaneously. As a result, several techniques have been developed. These techniques depend on the objectives and the specificity of the network nodes; homogeneous, heterogeneous, mobile, etc. Due to the energy limitation of sensor nodes, cluster-based approaches have gained immense research interest in the last few decades. Clustering is an effective way of conserving WSN's energy and organizing a large number of nodes efficiently [10].

The clustering process includes some characteristics that enable the construction of a suitable cluster: Cluster properties, cluster head properties and the cluster formation process [30].

Cluster properties are concerned with size in terms of the number of member nodes, the number of clusters, inter-cluster and intra-cluster communication [30,31].

Cluster head properties cover mobility (mobile or stationary), its role (relay or fusion), and nature (nodes are heterogeneous or homogeneous). As for the clustering process, cluster head selection can be centralized or distributed. Also, several CH selection methods have been studied in the literature. They are classified into different types: probabilistic, attribute-based, weighted probabilistic, optimization-based, etc. [31]. The technique used can also categorize the clustering

process, which can be metaheuristic or non-metaheuristic (exact) [31]. Metaheuristic methods consist of converting a theory inspired by nature or bio-inspired into mathematical calculations to solve optimization problems such as Particle Swarm Optimization (PSO), Firefly Algorithm (FA), Genetic Algorithm (GA), Ant Colony Optimization (ACO), etc. [33]. Grouping nodes into clusters aims to optimize energy consumption and maximize network lifetime. The performance of a clustering procedure depends on the number and distribution of CHs in the network. Since the clusters created are profoundly affected by the CH selection method, this is an important step of the clustering process [32].

In the following Table.1, we present some works using a centralized clustering process and mono-hop routing. The election of CH nodes is based on both metaheuristic and non-metaheuristic techniques.

Table.1. Some related works

| Ref N° | Algorithm                                 | Main idea  |
|--------|---|--|
| [11]   | GA  | The fitness function is based on:<br>-The distances between the node members and the CH<br>-The distance between the CH and the BS   |
| [12]   | Genetic-Algorithm (GAEEC)                 | The fitness function for cluster-head election is based on:<br>-The amount of residual energy,<br>-The distance from the base station -The density of nodes around specific node   |
| [13]   | GA  | Fitness based on:<br>-Sum of the distances of the nodes to the base station<br>-Distance from nodes to the CH and the distance from CH to BS<br>-The residual energy of nodes and the number of clusters                         |
| [14]   | Improved k-means                          | -Using weight of CH and weight of cluster for CH election<br>-If the weight of the node is equal to the weight of the cluster, it will be selected as a CH node  |
| [15]   | OK-K means                                | CH election uses:<br>-The node closest to the centroid with sufficient energy<br>- A node elected as CH during the previous round, must not be selected  |
| [16]   | Energy-saving clustering algorithm (ESCA) | -Centralized solution: K-mean is executed on the BS<br>-CH election is based on the residual energy of nodes and the distance of their cluster regard to others nodes.   |
| [17]   | Improved K-means                          | -Choosing initial cluster centers by applying a probabilistic method<br>-Recalculates the sum of distances between each node and its closest cluster center  |
| [18]   | Hybridization between the GA and K-means  | -K-means used to create the initial population<br>-Then genetic algorithm enforced to create the clusters<br>Fitness function is defined based on intra and inter cluster distances and the number of alive nodes in the network |
| [19]   | Improved LEACH                            | GA used to select the CH with more residual energy   |

In the previous works, the objective function is based on the amount of residual energy and the distance between the node and BS or centroid. This distance is static (stationary nodes) and the same node can be elected in many rounds, it results that these nodes dead quickly. Adding the amount of residual energy as another parameter to elect CH node avoids this issue. Nevertheless, the same node can be also elected in many rounds. The election procedure of the CH node must be fair for all eligible nodes in order to extend their lifetimes. The frequency of election in the previous rounds should be taken into consideration to select CH.

In many works, the start points of K-means are chosen randomly. Since K-means is sensitive to the K starting points, it is suitable to find a solution to select these K starting points in an optimal way. Moreover, the position of the BS is fixed arbitrarily and not justified. In the case of random deployment of the nodes and with single hop routing, the position of BS has an impact on the network lifetime.

### 3. PROPOSED APPROACH

The proposed solution consists of a hybridization of Genetic Algorithm and Improved K-Means (GA-IKM). As K-means is sensitive to the  $k$  starting points [20], in terms of number and nature, the main idea is to choose the  $k$  starting points in an efficient way for K-means operation.

The selection of the starting points is entrusted to a genetic algorithm while the formation of the clusters is carried out by k-means with other improvements that take into account the constraints specific to IoT networks. Firstly, we present the procedure of the genetic algorithm. Secondly, we explain the improvements of k-means (section 3.3.3).

#### 3.1. Network Model

The network model consists of a set of nodes, which are randomly deployed in a field. The base station (BS) is deployed in the center of the field.

The solution is centralized and it is run in the Base Station (BS). Therefore, it is necessary to set the following assumptions:

- All nodes are homogeneous. So they have the same technical characteristics; computing capacity, storage capacity and amount of energy.
- The nodes are stationary (not mobile).
- The base station is able to determine the geographic coordinates of the nodes.
- A node can change its transmission range according to its needs.
- All nodes use the same packet size

#### 3.2. Energy Dissipation Model

The architecture of an IoT object is essentially based on three parts. Sensing module, processing module and communication module (radio module) [21]. The communication module is the greediest in terms of energy consumption [22]. It depends on the structure of the network and the routing protocols. In the rest of this work, we focus on the communication module. There are different radio energy models in the literature [34]. The most common energy model used is that of the first order Fig.1.

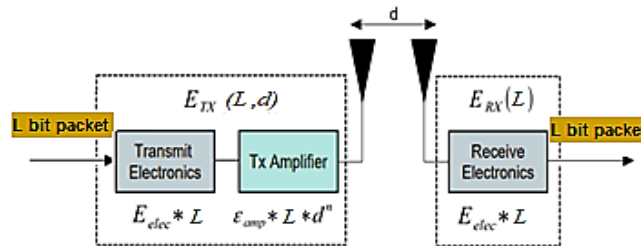


Fig.1. Radio Energy Dissipation Model [23]

During a communication procedure, a node can play the role of a transmitter or a receiver. The consumed energy by a node N to send a packet of  $L$  bits over a distance  $d$  is given by the formula (1) [24].

$$E_{TX}(N) = E_{elec} * L + E_{amp} * L \quad (1) \text{ where } E_{amp} = \begin{cases} L * d^4 & \text{if } d \geq d_0 \\ L * d^2 & \text{if } d < d_0 \end{cases}$$

$E_{elec}$  : is the electronic energy required for signal processing such as filtering, modulation.

$d_0$ : a threshold distance.  $d_0 = \sqrt{\frac{E_{fs}}{E_{amp}}}$

$E_{fs}$  and  $E_{amp}$  are the amplification energy in the free space and multi-path fading careers, respectively.

The consumed energy by a node N to receive a packet of  $L$  bits is given by the formula (2).

$$E_{RX}(N) = E_{elec} * L \quad (2)$$

### 3.3. Algorithm Flow

The proposed algorithm begins with an initialization phase of the network parameters. The data transmission mechanism from the sensor nodes to the base station takes place in the form of rounds [35]. In our case, each round proceeds in these steps:

- (i) Computing the K number, (ii) Election of the CH nodes, (iii) Creation of clusters and (iv) Broadcasting the clustering scheme to all nodes. Once this is done, the data transmission begins. At the end of each round the nodes communicate their amounts of residual energy to the base station which proceeds to a new clustering scheme for the next round. Fig.2 resumes all the steps of the solution.

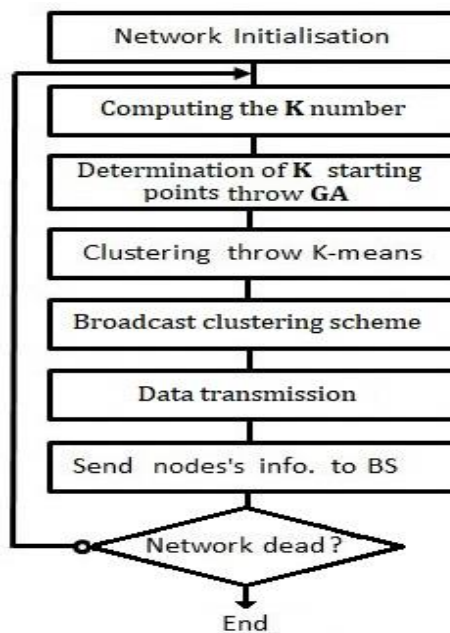


Fig .2. Flowchart of GA-IKM

### 3.3.1. Computing of the Number **K**

The number **k** must be chosen in a rigorous way in the goal to minimise the energy consumption in the network. Several techniques were used like rule of thumb, information Criterion approach, the Silhouette, the elbow method, Cross-validation, and Information Theoretic Approach [25, 26].

Most of these methods need to execute the clustering algorithm several times to fix the adequate value of **K**. In this study, we launch GA with different values of **K** starting from 2, and we note the **K** that minimizes the cost function.

This function corresponds to the total energy consumed by the network and it is calculated as:

$$E_{\text{Total}} = \sum_{i=1}^k [ E(\text{CH}_i) + \sum_{j=1}^{S_i} E(N_j) ] \quad (3)$$

Where:

$E(\text{CH}_i)$  is the consumed energy by  $i^{\text{th}}$  CH and

$E(N_j)$  is the consumed energy by the  $j^{\text{th}}$  node, closest to  $\text{CH}_i$ .

$E(N_j) = E_{\text{TX}}(N_j)$  (given by formula (1) in section 3.2)

$E(\text{CH}_i) = E_{\text{RX}}(\text{CH}_i) + E_{\text{TX}}(\text{CH}_i) \quad (4)$

$E_{\text{RX}}(\text{CH}_i) = (E_{\text{RX}} + E_{\text{DA}} * L) * S_i \quad (5)$

$E_{\text{DA}}$  is the energy consumed in data aggregation

$S_i$ : the number of the  $\text{CH}_i$ 's closest nodes (taking into account the distance)

This objective function is very important because it improves the behaviour of **K**-means. This later is sensitive to starting points that are often chosen randomly. Using GA with this objective function allows to better orient **k**-means when calculating the centroids.

In addition, this function is used at the first round of clustering to compute the initial **K** optimal number of clusters. For the rest of rounds, the number **K** is calculated by the projection of the number of alive nodes on the interval  $[0, K_{\text{initial}}]$ .

For example, if in the first round we have found an optimal value for **K** which is 9 for 100 alive nodes, in this case  $K_{\text{initial}}$  is 9. For subsequent rounds, **K** is chosen by resizing the number of alive nodes over the interval  $[0, K_{\text{initial}}]$ . The resizing data between a set of arbitrary values  $[a, b]$  is given by formula (6)

$$\text{Transformed\_Value} = \text{Round} \left( a + \frac{(\text{CurrentValue} - \text{Minimum}) * (b-a)}{(\text{Maximum} - \text{Minimum})} \right) \quad (6)$$

In our case  $[a, b] = [0, K_{\text{initial}}]$

$$K_{\text{initial}} = \text{Round} \left( 0 + \frac{(N_i - 1) * (K_{\text{initial}} - 0)}{(N - 1)} \right)$$

where:

**K<sub>i</sub>**: the value of **K** to be computed at round **i**.

**N<sub>i</sub>**: Number of alive nodes at round **i**.

**N**: The initial number of alive nodes (Maximum number of alive nodes).

The final number of alive nodes is 1 (Minimum number of alive nodes).

**Example:**

If we have  $N=100$  then, the initial number of alive nodes  $K_{initial}=9$

If  $N_i= 80$ , the number of alive nodes at round  $i$ , then  $K_i=Round\left(0+\frac{(80-1)*(9-0)}{(100-1)}\right) = 7$

If  $K_i=0$ , then the alive nodes communicate directly with the BS.

**3.3.2. Determination of Starting Points for K-Means**

The Determination of starting points for k-means is performed by the GA [27]. Starting points correspond to some  $k$  points chosen randomly as the initial centroids for k-means.

An individual from the initial population represents a potential solution [28]. This individual is represented by one chromosome that has  $k$  genes. Each gene represents a node among the alive nodes. The chromosomes are built in a random way and gene coding is numeric and uses decimal format. Each gene represents an  $id$  of a node in the network. Fig.3 gives an example of a chromosome with size = 10

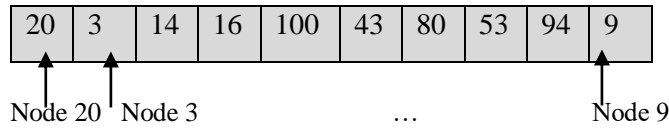


Fig .3. Representation of a chromosome

**a) Evaluation**

The population is evaluated to take only the individuals which have the best value of the fitness function. The way to calculate this function depends on the amount of energy that is consumed during a round ( $E_{total}$  in section 3.2). The energy consumption is much higher for CH nodes because this type of nodes has more load than its members. The CH collects data from its members, operates the aggregation and finally transmits the messages to the base station. Energy consumption depends also on the number of member nodes. To optimize the consumption, we have to build clusters that are balanced in terms of the number of member nodes. As a result, the objective function depends on the energy consumed by a CH node and the balancing of the sizes of clusters. To measure the balance, the standard deviation of CH neighbours ( $CH_{ngb}$ ) is used.  $\sigma(CH_{ngb}) = 0$  if all CHs have the same number of neighbours.

$$Fitness = W_1 * E_{Total} + W_2 * \sigma(CH_{ngb}). \quad (7)$$

$W_1$  and  $W_2$  represent weights.

The total energy consumed by the network ( $E_{Total}$ ) is given by the formula (3) in 3.3.1

**b) Selection**

The selection operator is used to select the parents that will participate in the crossover step. In our case, we used the Truncation selection because it is simple to implement and it conserves the best parents for the crossover.

**c) Crossover and Mutation**

The crossover is performed by dividing the chromosomes into sub-sequences. The crossover operator can be one-point crossover, Two-point crossover, K-point or uniform crossover [36]. The first operator is the common used. It allows conserving the most part of the genetic information. The mutation consists of randomly choosing some genes, which will be replaced by others, adding or deleting them. The rate depends on the size of the population, the number of genes or the type of the problem. Typically, in the literature, it is chosen at 10%.

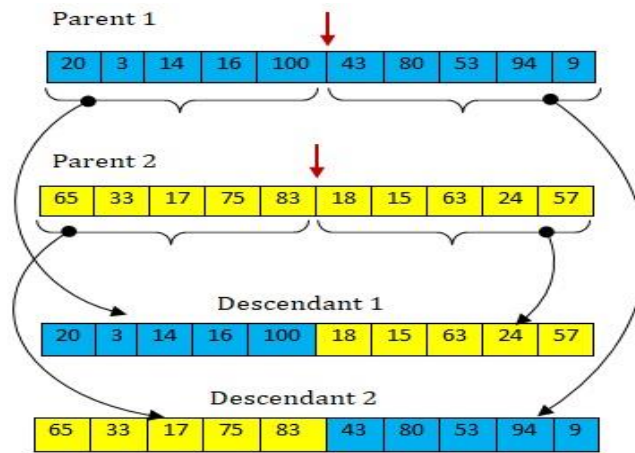


Fig .4 crossover procedure

Once the crossover is done, we randomly take genes from each descendant to be muted with respect to the mutation rate. Also, the mutation should not result in duplicate genes in the chromosome fig .5.

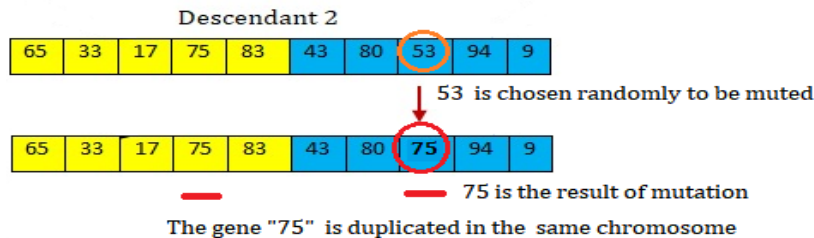


Fig .5 mutation procedure

Example: For the descendant 2, if we the gene 53 to be muted and the result gives 75. In this case we obtain two occurrences of the gene 75. So the gene taken randomly from the set of  $N$  nodes except those existing in the chromosome. The output of the genetic algorithm is a vector of  $K$  nodes, which will serve as starting centroids for k-means to perform the procedure of clustering.

**3.3.3. Clusters Building Throu K-Means**

Once we have selected the starting points (nodes) by GA, we form the clusters throu k-means algorithm. These starting points will guide the algorithm to work efficiently. Once k-means was executed, we proceed to the selection of the final CHs.

### CH nodes' election in k-means algorithm

In the basic version of k-Means [24], the CHs election is based on the Euclidean distance. Then we choose the node closest to the centroid of the cluster. In our study, the CH node is elected by taking into consideration some criteria which are:

- Distance from the candidate to the BS, the amount of residual energy and the frequency of the election as CH in the previous rounds.

-Let  $\mathbf{A}$  be the set formed by the  $p$  nodes in the cluster.  $\mathbf{A} = \{n_1, n_2, n_3, \dots, n_p\}$ ,  $p$  is the number of member nodes in the cluster.

-Let  $\mathbf{C}$  be a subset of candidate nodes likely to be chosen as CH node.  $\mathbf{C} = \{c_1, c_2, \dots, c_q\}$ ,  $\mathbf{C} \subset \mathbf{A}$ .  $q \leq p$

-The set  $\mathbf{C}$  of candidates is formed by the nodes, which have (i) the maximum amount of residual energy, (ii) the lowest election frequencies and (iii) The lowest distances to BS.

For this purpose, each node is assigned a frequency of participation as CH per round.

Initially, all the nodes have the same frequency (zero) and the same residual energy (100%).

At each round, the frequency of participation for each node is updated as well as the amount of residual energy. At the end of each round, we recalculate  $\mathbf{C}$  as follows.

-Let  $\mathbf{SE}$  be the set of nodes which have amount of residual energy greater than a threshold. The threshold corresponds to the amount of energy required to transfer a packet to the BS.

-Let  $\mathbf{SF}$  be the set of nodes with the minimum of election frequency in the previous rounds.

-Let  $\mathbf{SD}$  be the set of nodes with the minimum distances from BS.

So,  $\mathbf{C} = \mathbf{SE} \cap \mathbf{SF} \cap \mathbf{SD}$ . If  $\text{length}(\mathbf{C}) \geq 1$  we take a randomly node. Otherwise we choose the node that has the max of residual energy.

After the formation of the clusters some member nodes are closer to the BS than to their CH nodes (fig.6). In this case there will be more energy dissipation because the packets are transmitted from the member node to the CH and then from the CH to BS. The energy dissipated will be less significant if these member nodes communicate directly with the BS. As a result, nodes closest to BS can communicate directly with it (fig.7). For that, the BS will be considered as an additional CH.

Table.2. Pseudo code of Genetic Algorithm and Improved K-means

| Genetic Algorithm  | IKM   |
|--|---|
| <p><b>Input :</b><br/> <math>k</math> , Network , <math>N</math> /* Total node number*/<br/> <b>Output:</b> StartingPoints /*vector of starting points*/<br/> <b>Begin</b><br/>                     MaxIteration =100, SizePop = 40 , It = 0<br/>                     Population = [1..SizePop] of Chromosome<br/>                     Population = <b>CreateRandomPop</b>(<math>K</math>, Network, <math>N</math>)<br/>                     Population = <b>Evaluate</b> (Population)<br/> <b>Repeat</b><br/>                     NewPopulation=<b>Crossover</b>(Population)<br/> <b>Mutation</b> (NewPopulation, RateMut)<br/>                     NewPopulation = Evaluate (NewPopulation)<br/>                     Population=<b>SelectBest</b>(NewPopulation,50%)<br/>                     It=It+1<br/> <b>Until</b> It=MaxIteration<br/>                     StartingPoints=Population<br/> <b>End</b></p> | <p><b>Input:</b> Network, <math>K</math>, <math>N</math><br/> <b>Output :</b> <math>K\_Clusters</math> /* matrix of <math>K</math> Clusters*/<br/> <b>Begin</b><br/>                     CHs = [1..<math>k</math>] of nodes<br/>                     StartingPoints=<b>GA</b>(<math>K</math>, Network, <math>N</math>)<br/>                     [Centroids, Members]=<b>K- Means</b>(<math>K</math>, Network, StartingPoints)<br/>                     CHs= <b>CH_Elections</b>(Centroids, Members, Network)<br/>                     ClustersModel= <b>ClustersFormation</b>(CHs, Network, <math>N</math>)<br/> <b>End</b></p> |

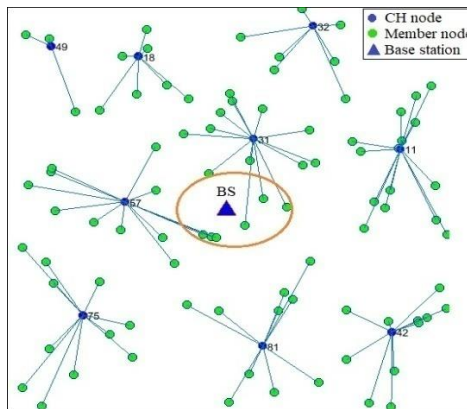


Fig.6. Nodes closest to BS

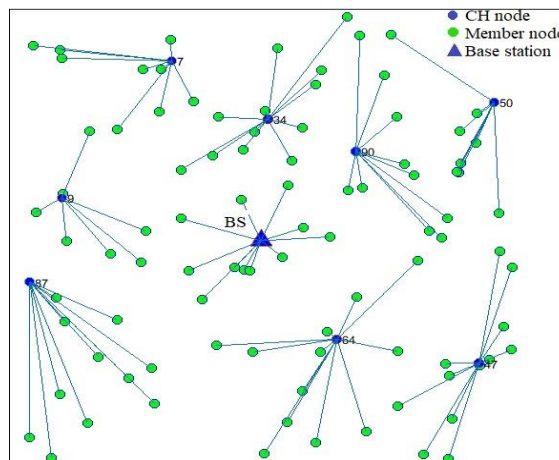


Fig.7. BS as a cluster head (CH)

Therefore, our contribution can be resumed in three steps.

**Step 1:** We use a genetic algorithm to select the  $k$  starting points for k-means.

**Step 2:** The election of the CHs is based on the amount of residual energy of the node, the distance to BS of this node and its frequency of election in the previous rounds.

**Step 3:** We include the BS as a supplementary CH.

## 4. SIMULATION RESULTS AND DISCUSSION

### 4.1. Simulation

The simulation is performed on a set of 100 nodes which are randomly deployed on a square surface of 100x100m. All nodes are homogeneous. In addition, the phenomena of disturbance and signal interference are not taken into account. The evaluation criteria in our simulation focus on:

**FND** (First dead node): Corresponds to the round number at which the first node is exhausted.

**LDN** (Last Dead Node): Is the round number at which all nodes are exhausted and therefore the whole network is dead.

**Residual Energy:** The amount of residual energy in the whole network.

**PDR** (Packet Delivery Rate): This factor is very important because it tells us about the robustness of the network and the guarantee of packet reception. It is defined as the ratio of the number of received packets to the number of packets sent [29].

Our proposed solution is compared to three algorithms that are; LEACH, K-means [24] and OK-means [15]. Leach is the reference algorithm for the study of WSN, K-means is very used for its convergence speed and OK-mean witch still recent and give interesting results in terms of extending the network lifetime.

#### 4.1.1. Simulation Environment

The simulation environment used in our work is MATLAB 2020 from Math works.

The Simulation parameters for the genetic algorithm and network are mentioned in Table.3.

Table.3. Parameters of simulation

| G.A parameters               | value | Network parameters                  | Value       |
|------------------------------|-------|-------------------------------------|-------------|
| Initial population size      | 40    | Region size                         | 100 x 100 m |
| Maximum number of iterations | 100   | Number of nodes                     | 100         |
| Mutation rate                | 0.1   | BS position                         | (50,50)     |
|                              |       | Data Packet size                    | 4000 Bytes  |
|                              |       | Control Packet size                 | 100 Bytes   |
|                              |       | Initial Energy for a node           | 0.5 j       |
|                              |       | Transmitter/Receiver electronics    | 50 nj       |
|                              |       | Efs                                 | 10pj        |
|                              |       | Eamp                                | 0.0013pj    |
|                              |       | EDA                                 | 5nj         |
|                              |       | d0                                  | 87m         |
|                              |       | Probability of CH election ( LEACH) | 0.5         |

In the literature of genetic algorithm, there is not one initialization method that works correctly on all problems. The choice of these values is empirical to have a reasonable running time with an acceptable result.

## 4.2. Results and Discussion

The simulation results given below use single hop routing and the same energy dissipation function presented by equation (3) in section number (3.3.1).

To better understand the advantages of our proposal we present the four use cases below.

- **Case 1:** (IKM) we executed k-means with the improvement of the CH election that is explained in section 3.3.1. In this case the FDN took place at round 1546.
- **Case 2:** We used GA to select the k starting points for K-means. Here we got the FDN at the round 1875.
- **Case 3:** We launch IKM with taking into consideration the BS as a CH node. The FDN took place at round 1900.
- **Case 4:** In this case we combine all possibilities. We used the GA to select the starting k points for IKM and we used BS as a CH node. So the FDN occurred at round 1988. Table.4 resumesthe results of these use cases.

Since the positions of the nodes are chosen randomly and we have the random aspect of GA, Referring to [29], a number of 32 executions is sufficient to prove the superiority of our solution over the others.

Table.4. The four improved versions of k-means

| Use cases              | FDN  | LDN  |
|------------------------|------|------|
| Improved K-Means (IKM) | 1546 | 2057 |
| GA-IKM                 | 1875 | 2054 |
| IKM with BS as CH      | 1900 | 2340 |
| GA-IKM with BS as CH   | 1988 | 2354 |

We compared our proposal to the k-means, LEACH and Ok-means by considering the FDN, LDN, Total residual energy and PDR metrics.

- **First Dead Nodes (FDN)**

Fig.9 indicates the number of dead nodes per round. This parameter represents the steady phase of the network. The more this value is delayed the more is efficient the clustering algorithm.

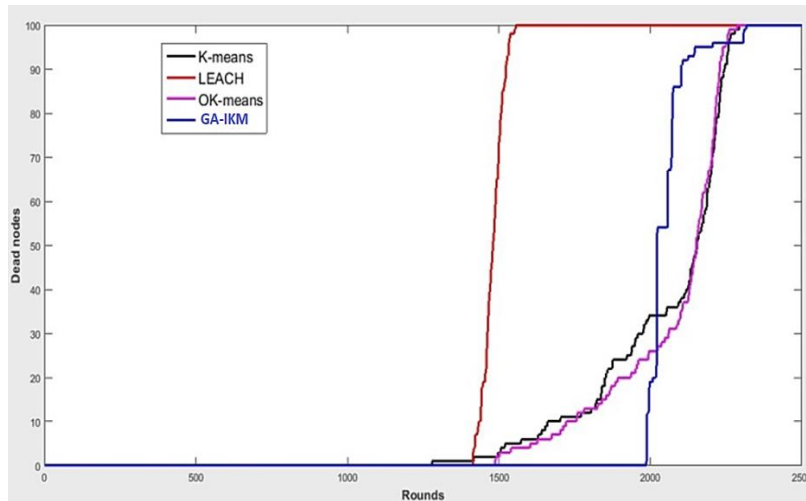


Fig.8. First dead nodes

Table.5 shows clearly the FDN and LDN for the four algorithms

Table.5. FDN and LDN

| Algorithms | LEACH | K-means | OK-means | GA-IKM      |
|------------|-------|---------|----------|-------------|
| <b>FDN</b> | 1413  | 1288    | 1489     | <b>1988</b> |
| <b>LDN</b> | 1554  | 2294    | 2287     | <b>2354</b> |

Generally, the nodes that will be exhausted first are the CH nodes because they are the ones that have more loads and therefore consume more energy. Therefore, the election of the appropriate CH is a crucial operation. With the same simulation parameters, we note that our proposed solution gives better results compared to LEACH, K-means, OK-means and even ESCA [16].

The election of CH in LEACH is based on probability and therefore the factors influencing the energy consumption are not taken into account in an exhaustive way. In the basic version of the k-means algorithm, the election of the CH node is based on the minimum distance from the centroid. For the case of OK-means, which resembles ours, the election of the CH node takes into account the node closest to the centroid with sufficient energy and which was not selected during the previous round. Our solution improved k-means by introducing starting points (using GA) and in addition during the election of the CH nodes, the distance from the BS and the frequency of previous elections are taken into account.

- **Residual Energy**

Fig.9 shows the amount of residual energy in the whole network. Here we notice again that our proposed solution gives most improvements even with a slight improvement over OK-means.

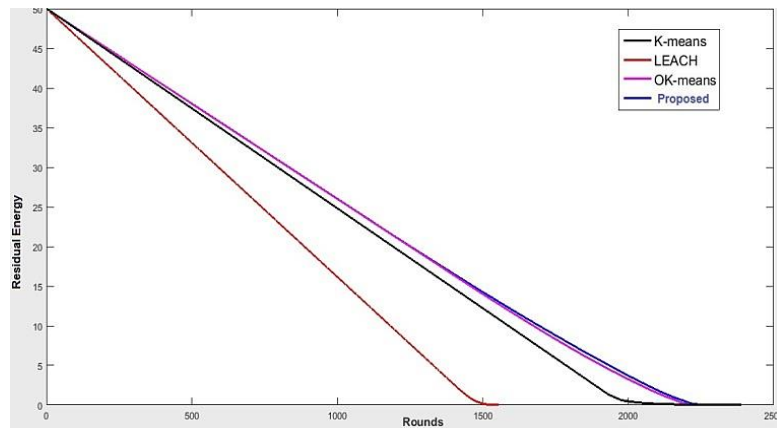


Fig.9. the total residual energy

• **Packet Delivery Rate (PDR)**

Fig10 shows the PDR generated by the four algorithms. We constant that our proposed solution gives the best value. This is because our algorithm has the most important FDN which makes it possible to guarantee the correct transmission and reception of the packets in the network.

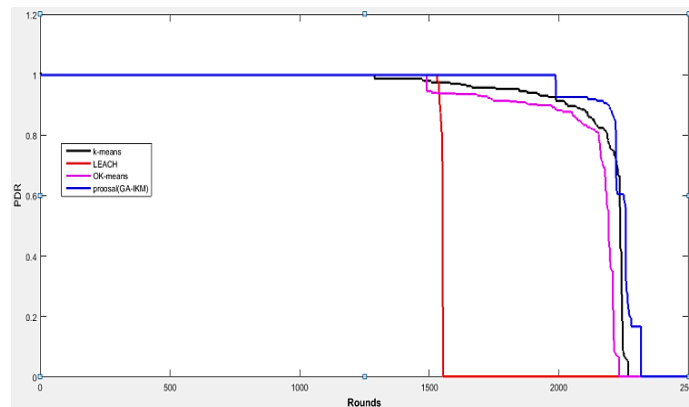


Fig.10. Packet Delivery Rate

• **BS position**

Now we discuss the influence of the choice of the BS position on the consumption of energy. Table.6 shows the impact of changing BS location.

Table.6. the impact of the BS position

| <b>BS position</b> | (0,0) | (0,100) | (50,100) | (50,0) | <b>(50,50)</b> |
|--------------------|-------|---------|----------|--------|----------------|
| FDN                | 1815  | 1643    | 1869     | 1920   | 1988           |
| LDN                | 2369  | 2341    | 2360     | 2389   | 2354           |

When the base station is at the end of the field, the FDN weakens rapidly and therefore the lifetime of the network too. Whereas, if the BS is placed in the center of the area to be supervised, the FDN will be delayed. This is due to single hop routing. For large jumps the CHs will consume more energy. The last column in Table.6 shows the importance of deploying the BS

in the center. We notice a slight deterioration in performance for the LDN but despite this deterioration, they remain good. This is due to the random deployment of nodes. Nodes closest to the BS consume less energy and therefore their lifespan is longer, which gives a larger LDN.

- **The deployment area**

Table.7 shows the impact of the deployment area. By increasing the deployment area and for a constant number of nodes, we notice a drop in the FDN value. This is because we used single-hop routing, the distance between the CHs and the BS becomes more and more important.

Table.7. The impact of the deployment area

| Number of nodes: 100, BS position: in center of area |      |      |
|--|------|------|
| Deployment Area                                      | FDN  | LDN  |
| 100 x 100 m  | 1988 | 2354 |
| 200 x 200 m  | 1185 | 2280 |
| 300 x 300 m  | 325  | 2265 |
| 400 x 400 m  | 168  | 2000 |

## 5. CONCLUSION AND PERSPECTIVES

In an IoT network, the optimization of energy consumption is a major challenge to be solved. The clustering technique is an efficient means of optimization. In this work, an improvement of the k-means algorithm by hybridization with a genetic algorithm is proposed in order to elect, efficiently, then CH node. This suggestion gives significant results mainly for FDN which increases the lifetime of the network but it does not adapt to large networks and it is sensitive to the base station position. This is due to the use of a single-hop routing. Since the communication phase is the greediest of energy consumption, in the future research we can focus on more energy-efficient routing protocols. To deal with this issue the integration of reinforcement learning is a current trend for optimization in IoT networks. This technique allows adaptation to dynamic changes in the network. In addition, optimization in such an IoT network depends on several parameters like, residual energy, distance, reliability of nodes ... etc. Therefore, a multi-objective solution will be much more interesting whether for clustering or for routing.

## REFERENCES

- [1] Rani, S., Ahmed, S.H. & Rastogi, R. Dynamic clustering approach based on wireless sensor networks genetic algorithm for IoT applications. *Wireless Netw* 26, 2307–2316 (2020).
- [2] HUSSAIN, Muhammad Zunnurain et HANAPI, Zurina Mohd. Efficient Secure Routing Mechanisms for the Low-Powered IoT Network: A Literature Review. *Electronics*, 2023, vol. 12, no 3, p. 482.
- [3] Abdulzahra, Ali Mohammed Kadhim, Ali Kadhum M. Al-Qurabat, and Suha Abdul Hussein Abdulzahra. "Optimizing energy consumption in WSN-based IoT using unequal clustering and sleep scheduling methods." *Internet of Things* 22 (2023): 100765.
- [4] DEL-VALLE-SOTO, Carolina, RODRÍGUEZ, Alma, et ASCENCIO-PIÑA, Cesar Rodolfo. A survey of energy-efficient clustering routing protocols for wireless sensor networks based on metaheuristic approaches. *Artificial Intelligence Review*, 2023, p. 1-72.
- [5] Del-Valle-Soto, C., Rodríguez, A. & Ascencio-Piña, C.R. A survey of energy-efficient clustering routing protocols for wireless sensor networks based on metaheuristic approaches. *ArtifIntell Rev* 56, 9699–9770 (2023). <https://doi.org/10.1007/s10462-023-10402-w>.
- [6] Yuste-Delgado, Antonio-Jesus, Juan-Carlos Cuevas-Martinez, and Alicia Triviño-Cabrera. "A distributed clustering algorithm guided by the base station to extend the lifetime of wireless sensor networks." *Sensors* 20.8 (2020): 2312.

- [7] Shahraki, Amin, et al. "A survey and future directions on clustering: From WSNs to IoT and modern networking paradigms." *IEEE Transactions on Network and Service Management* 18.2 (2020): 2242-2274.
- [8] Wohwe Sambo, Damien, et al. "Optimized clustering algorithms for large wireless sensor networks: A review." *Sensors* 19.2 (2019): 322.
- [9] Iwendi, C., Maddikunta, P. K. R., Gadekallu, T. R., Lakshmana, K., Bashir, A. K., & Piran, M. J. (2021). A metaheuristic optimization approach for energy efficiency in the IoT networks. *Software: Practice and Experience*, 51(12), 2558-2571.
- [10] Rostami, Ali Shokouhi, et al. Survey on clustering in heterogeneous and homogeneous wireless sensor networks. *The Journal of Supercomputing* 74 (2018): 277-323.
- [11] MERAH, Malha, ALIOUAT, Zibouda, HARBI, Yasmine, et al. Machine learning-based clustering protocols for Internet of Things networks: An overview. *International Journal of Communication Systems*, 2023, p. e5487.
- [12] Singh, Santar Pal, and S. C. Sharma. "Genetic-algorithm-based energy-efficient clustering (GAEEC) for homogenous wireless sensor networks." *IETE journal of research* 64.5 (2018): 648-659.
- [13] Heidari, Ehsan, et al. A novel approach for clustering and routing in WSN using genetic algorithm and equilibrium optimizer. *International Journal of Communication Systems* 35.10 (2022): e5148.
- [14] Razzaq, Madiha, Devarani Devi Ningombam, and Seokjoo Shin. Energy efficient K-means clustering-based routing protocol for WSN using optimal packet size. 2018 International Conference on Information Networking (ICOIN). IEEE, 2018.
- [15] El Khediri, Salim, et al. "Improved node localization using K-means clustering for Wireless Sensor Networks." *Computer Science Review* 37 (2020): 100284.
- [16] NEDHAM, WisalBassim et AL-QURABAT, Ali Kadhum M. An improved energy efficient clustering protocol for wireless sensor networks. In: 2022 International Conference for Natural and Applied Sciences (ICNAS). IEEE, 2022. p. 23-28.
- [17] Ahmad, Waseem, et al. "Optimizing Energy Efficiency in Wireless Sensor Networks using Enhanced K-Means Cluster Head Selection." *International Journal of Communication Networks and Information Security* 16.3 (2024): 565-573.
- [18] Bhushan, Shashi, Raju Pal, and Svetlana G. Antoshchuk. "Energy efficient clustering protocol for heterogeneous wireless sensor network: a hybrid approach using GA and K-means." 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). IEEE, 2018.
- [19] Bhola, Jyoti, Surender Soni, and Gagandeep Kaur Cheema. "Genetic algorithm based optimized leach protocol for energy efficient wireless sensor networks." *Journal of Ambient Intelligence and Humanized Computing* 11 (2020): 1281-1288.
- [20] Hassan, A. A. H., et al. "Clustering approach in wireless sensor networks based on K-means: Limitations and recommendations." *Int. J. Recent Technol. Eng* 7.6 (2019): 119-126.
- [21] Obeid, Abdulfattah Mohammad, et al. "A survey on efficient power consumption in adaptive wireless sensor networks." *Wireless Personal Communications* 101 (2018): 101-117.
- [22] Singh, Jaspreet, Ranjit Kaur, and Damanpreet Singh. "A survey and taxonomy on energy management schemes in wireless sensor networks." *Journal of Systems Architecture* 111 (2020): 101782.
- [23] SINGH, Shashank et ANAND, Veena. Load balancing clustering and routing for IoT-enabled wireless sensor networks. *International Journal of Network Management*, 2023, vol. 33, no 5, p. e2244.
- [24] Ray, Anindita, and Debashis De. "Energy efficient clustering protocol based on K-means (EECPK-means)-midpoint algorithm for enhanced network lifetime in wireless sensor network." *IET Wireless Sensor Systems* 6.6 (2016): 181-191.
- [25] Marutho, Dhendra, Sunarna Hendra Handaka, and EkapranaWijaya. "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news." 2018 international seminar on application for technology of information and communication. IEEE, 2018.
- [26] Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." *International Journal* 1.6 (2013): 90-95.
- [27] Kramer, Oliver, and Oliver Kramer. *Genetic algorithms*. Springer International Publishing, 2017.
- [28] Seo, Hyun-Sik, Se-Jin Oh, and Chae-Woo Lee. "Evolutionary genetic algorithm for efficient clustering of wireless sensor networks." 2009 6th IEEE Consumer Communications and Networking Conference. IEEE, 2009.

- [29] Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1), 3-18.
- [30] BAQERI, Javad. Increase the Lifetime of Wireless Sensor Networks Using Hierarchical Clustering with Cluster Topology Preservation. *International Journal of Wireless & Mobile Networks (IJWMN) Vol*, 2021, vol. 8.
- [31] RAJ, Bryan, AHMEDY, Ismail, IDRIS, Mohd Yamani Idna, et al. A survey on cluster head selection and cluster formation methods in wireless sensor networks. *Wireless Communications and Mobile Computing*, 2022, vol. 2022, p. 1-53.
- [32] AL-SULAIFANIE, Adnan Ismail, AL-SULAIFANIE, Bayez Khorsheed, et BISWAS, Subir. Recent trends in clustering algorithms for wireless sensor networks: A comprehensive review. *Computer Communications*, 2022, vol. 191, p. 395-424.
- [33] DEL-VALLE-SOTO, Carolina, RODRÍGUEZ, Alma, et ASCENCIO-PIÑA, Cesar Rodolfo. A survey of energy-efficient clustering routing protocols for wireless sensor networks based on metaheuristic approaches. *Artificial Intelligence Review*, 2023, p. 1-72.
- [34] ROY, Nihar Ranjan et CHANDRA, Pravin. Energy dissipation model for wireless sensor networks: a survey. *International Journal of Information Technology*, 2020, vol. 12, p. 1343-1353.
- [35] Gülbaş, Gülşah, and Gürcan Çetin. "Lifetime Optimization of the LEACH Protocol in WSNs with Simulated Annealing Algorithm." *Wireless Personal Communications* 132.4 (2023): 2857-2883.
- [36] Alhijawi, B., Awajan, A. Algorithmes génétiques : théorie, opérateurs génétiques, solutions et applications. *Évol. Intel.* (2023). <https://doi.org/10.1007/s12065-023-00822-6>

## AUTHORS

**MOEZ Elarfaoui** Obtained his computer engineering degree in 2001 from ENSI, Tunisia. He is university lecturer at the Zaghouan Higher Institute of Technological Studies. He is a member of SMART laboratory, Tunisia since 2022. His current research focuses on the Internet of Things.



**Nadia Ben Azzouna** is currently an associate professor at ESSECT, Tunisia. She is a member of the SMART laboratory, Tunisia since 2009. She received her master and engineer degrees in computer science and networking in 2001 from the IFSIC and the ENSTB, France respectively. She obtained her Ph.D. in computer science from the university Pierre et Marie Curie, France in 2004. Her current research interests include ubiquitous and pervasive computing, Internet of things, access control, privacy and trust management.

