A Major Project-II Report

On

HYBRID CLUSTERING FOR THE ENHANCEMENT OF NETWORK LIFETIME IN WIRELESS SENSOR NETWORK

Submitted in Partial fulfilment of the Requirement for the Degree of

Master of Technology

in

Computer Science and Engineering

Submitted By

Mikrant Sharma

2K15/CSE/09

Under the Guidance of

Mr. R.K. Yadav

(Assistant Professor)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahabad Daulatpur, Main Bawana Road, Delhi-110042

June 2017

CERTIFICATE

This is to certify that Project Report entitled **“Hybrid Clustering for the Enhancement of Network Lifetime in Wireless Sensor Network”** submitted by **Mikrant Sharma** (roll no. 2K15/CSE/09) in partial fulfilment of the requirement for the award of degree Master of Technology (Computer Science and Engineering) is a record of the original work carried out by him under my supervision.

Project Guide
Mr. R.K. Yadav
Assistant Professor
Department of Computer Science & Engineering
Delhi Technological University

DECLARATION

I hereby declare that the Major Project-II work entitled **"Hybrid Clustering for the Enhancement of Network Lifetime in Wireless Sensor Network"** which is being submitted to Delhi Technological University, in partial fulfilment of requirements for the award of the degree of Master of Technology (Computer Science and Engineering) is a bonafide report of Major Project-II carried out by me. I have not submitted the matter embodied in this dissertation for the award of any other Degree or Diploma.

Mikrant Sharma
2K15/CSE/09

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Mr. R K Yadav for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided without which the project could not have been a success.

Secondly, I am grateful to Dr. Rajni Jindal, HOD, Computer Science & Engineering Department, DTU for her immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Mikrant Sharma

Roll No – 2k15/CSE/09

M. Tech (Computer Science & Engineering)

Delhi Technological University

ABSTRACT

The convergence of the Internet, communications, and information technologies, coupled with recent engineering advances, is paving the way for a new generation of inexpensive sensors and actuators, capable of achieving a high order of spatial and temporal resolution and accuracy. The technology for sensing and control includes sensor arrays, electric and magnetic field sensors, seismic sensors, radio-wave frequency sensors, electro-optic and infrared sensors, laser radars, and location and navigation sensors. Although WSNs share many commonalities with wired and ad hoc networks, they also exhibit a number of unique characteristics which set them apart from existing networks. These unique characteristics bring to sharp focus new routing design requirements that go beyond those typically encountered in wired and wireless ad hoc networks. Meeting these design requirements presents a distinctive and unique set of challenges. These challenges can be attributed to multiple factors, including severe energy constraints, limited computing and communication capabilities, the dynamically changing environment within which sensors are deployed, and unique data traffic models and application-level quality of service requirements.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

1. WSN: Wireless Sensors Network
2. CH: Cluster Head
3. GA: Genetic Algorithm
4. LEACH: Low-Energy Adaptive Clustering Hierarchy
5. PEGASIS: Power-Efficient Gathering in Sensor Information Systems
6. SEP: Stable Election Protocol
7. GAEEP: Genetic Algorithm-based Energy-Efficient Protocol
8. GEACH: Genetic Algorithm Based Energy Efficient Clustering Hierarchy
9. TL-LEACH: Two Level LEACH
10. A-LEACH: Amend LEACH
11. GA-LEACH: Genetic Algorithm based LEACH
12. E-LEACH: Energy LEACH
13. TDMA: Time Division Multiple Access

CHAPTER 1: INTRODUCTION

1.1 Wireless Sensor Network

Wireless sensors network is a wireless system of several little 'battery operated sensing devices' commonly known as sensor nodes, stationed to watch environmental or physical conditions or other parameters. A conventional sensors network comprises of a large number of sensor devices. They are linked to each other wirelessly. The sensor devices can convey amongst themselves utilising radio beacons. The sensor device is outfitted with radio transceivers, computing and sensing accessories and energy source. The resources contained in a single WSN node are scarce and confined: they possess restrained power supply, limited radio capabilities and limited on-board computational power. Hence a WSN system consolidates a gateway that unites wireless connectivity back to the wired network. This gateway is known as the base station or the sink node that also performs most of the computational tasks of the network. The base station is assumed to have an infinite power supply. The sensor devices need to transfer their sensed data to the sink node for processing. They can undeviatingly exchange information with the base station or through some intermediary sensor nodes.

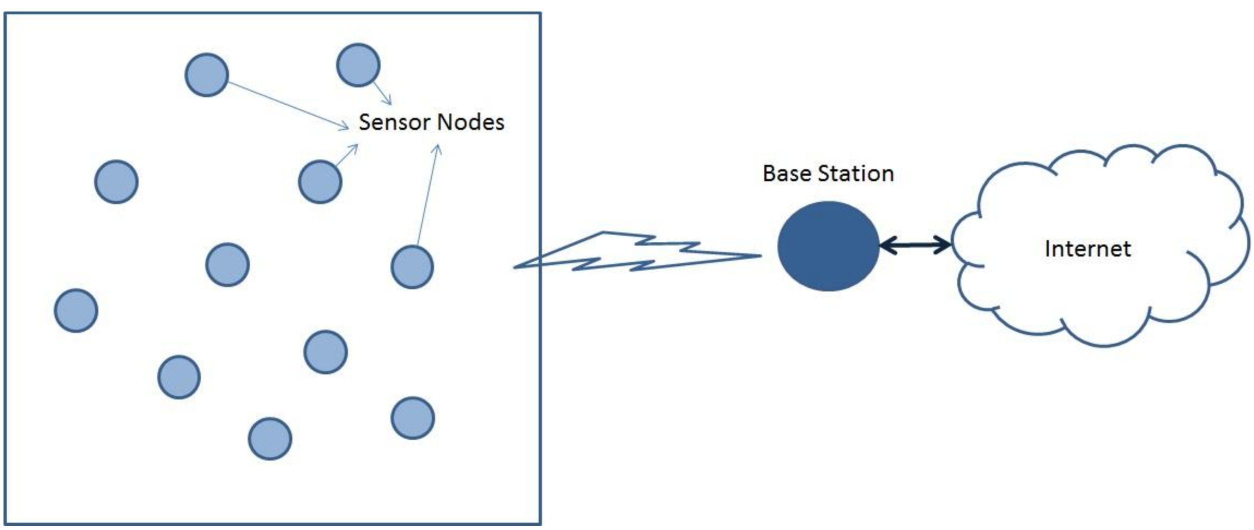

Fig. 1. Wireless Sensors Network

Communication between two nodes or with base station consumes much power, and hence the protocols and algorithms designed for communication should be energy conscious to prolong the network lifetime. Once the nodes are disposed of, they are now qualified for self-ordering into a proper system framework on a regular basis with multi-hop communication within themselves. At this point, the nodes begin gathering locally available data of importance to it. The wireless convention selected relies upon the application necessities.

1.2 Applications of WSN

Wireless sensor systems have enlarged notable recognition due to their adaptableness in the handling of concerns in various fields and can improve our livelihood in a broad span of ways. Wireless sensors networks have been successfully associated with diverse application spaces, for example:

- Region Monitoring: For examining a region, the nodes are distributed over a district where some phenomenon is to be observed. The moment of time when the sensors recognise the phenom being actuated (heat, humidity and so on), the phenom is communicated to the base stations, which at that instant takes proper activity.
- Agrarian area: Utilising a wireless system liberates the agriculturalist from the repairs of wiring in a problematic situation. Irrigation system mechanisation empowers more effective water utilisation and decreases waste.
- Wellbeing applications: Some of the usages for health monitoring using sensors networks are supporting GUIs for the restrained, coordinated patient examining, analysis, and managing medication in clinics, checking of an individual physiological report and trailing and checking physician or patients within a healthcare facility.
- Military usage: WSNs be suitably a central section of military regulation, administration, communications, figuring, insight, war zone reconnaissance and surveillance.
- Nature detecting:
  - Air contamination observing
  - Wildfire instrumentation
  - Habitat monitoring
  - Greenhouse observing
  - Landslide revealing

- Architectural Monitoring: WSNs can be employed to observe the actions inside structures and foundation, empowering engineering systems to control resources from a remote base; without physically being present at sight.
- Industrial checking: WSN networks provide notable cost saving infrastructure for machine control remotely, and there is no need of wired connections during installation of sensor and thus saving the wiring cost.
- Highway Checking: Real-time activity data is being gathered by Wireless sensors networks to later encourage transportation models and ready drivers of clog and movement issues. The sensors collect traffic flow statistics, like vehicle speeds, the volume of traffic, and highway densities, and then transmit this information through a wireless network to a base station.

1.3 Difficulties in a WSNs

A considerable number of difficulties ascend during the installation of sensors network. Sensor nodes convey over remote, unreliable lines with no foundation. An extra test is identified with the partial, generally non-sustainable power source supply of the sensor nodes. With a precise end goal to expand the lifetime of the system, the conventions should be planned from the earliest starting point with the goal of proficient administration of the vitality assets. The individual plan issues in more prominent detail.

Adaptation to internal failure: Sensor nodes are defenceless and much of the time conveyed in risky condition. Nodes can collapse because of equipment issues or environmental harm or by depleting their power supply. It is anticipated that the node crashes will be considerably greater than what is regularly viewed in traditional wireless systems or the wired networks. The conventions for such system ought to have the capacity to distinguish these disappointments as quickly as time permits and be sufficiently powerful to deal with a vast number of disappointments while keeping up the general usefulness of the system. This is particularly pertinent to the steering convention outline, which needs to guarantee that the substitute ways are accessible for rerouting of the parcels. Diverse organisation conditions posture distinctive adaptation to non-critical failure prerequisites.

Versatility: Sensor systems change in scale from a few nodes to conceivably a few hundred thousand. What's more, the sending thickness is likewise factor. For gathering high determination information, the density of nodes may increase up to a certain height when there are many nodes close to each other in the communication zone. Conventions that are being used in the sensor systems must be adaptable to certain heights and have the capacity for keeping up satisfactory execution.

Creation Costs: Because numerous arrangement models view the sensor nodes as dispensable gadgets, sensor systems can rival with conventional data collection procedures and hence the nodes could be manufactured efficiently.

Equipment Limitations: At the very least, every WSN node needs a transmission, processing and sensing system, and an energy source. Alternatively, the nodes can have implicit sensing equipment or smart gadgets, for example, a limitation framework to empower area mindful directing. Notwithstanding, every extra usefulness accompanies extra cost also, builds the power utilisation and physical size of the node. There should be a proper adjustment between the expense and low-energy specifications as per the changing functionalities.

Communication Medium: The correspondence among the sensors is ordinarily actualized utilising wireless medium in the famous industrial, scientific band. Be that as it may, some sensor systems utilise infrared or optical correspondence and former providing the obstruction free robust path.

WSN Topology: Even though WSN have developed in numerous viewpoints, these systems keep on to be with restrained assets as far as energy, processing force, memory, and interchanges capacities. Of these requirements, energy utilisation is of vital significance, which is exhibited by the expansive number of calculations, methods, and conventions that have been produced to spare energy, and along these lines expand the network life. Topology Route control is a standout amongst various critical concerns investigated for diminishing power utilisation of WSN systems.

Power Dissipation: As we have as of now observed, large portions of the difficulties of sensor systems rotate around the restricted power assets. The product and equipment configuration requires the consideration of the concerns of effective power utilisation. Just as an example, information pressure may diminish the measure of energy utilised for radio transmission, however, utilises extra energy for calculation as well as separate. The vitality arrangement likewise relies on upon the application. Sometimes, it is desired to power off a few nodes with a specific end goal to ration energy while different applications require all nodes working at the same time.

1.4 Architecture of WSN

The design of a WSN incorporates distinctive network topologies for wireless correspondences systems. The some of the network topologies which apply to distributed network systems are illustrated beneath:

Star Topology

This system employs the network topology in which an individual sink node can transmit or receive messages to/from the various sensors. Such nodes cannot transmit the messages to each other until there is establish routes among the nodes. One of the main benefits of the star network is that it can incorporate effortlessness, capacity for remote sensor systems by keeping or maintaining the remaining energy of node to the certain level so that there is an enhancement of longer period. Similarly, this topology permits less latency transmission connecting the sensor nodes and the sink node.

The limitations of star topology is that there is only single sink node at the center to which all the nodes in the network can communicate the sensed data to this node and thus leads to congestion at this node. Also, a single node cannot manage the network whenever node failures occur. For WSN, the sink node must be reachable for all the sensor nodes.

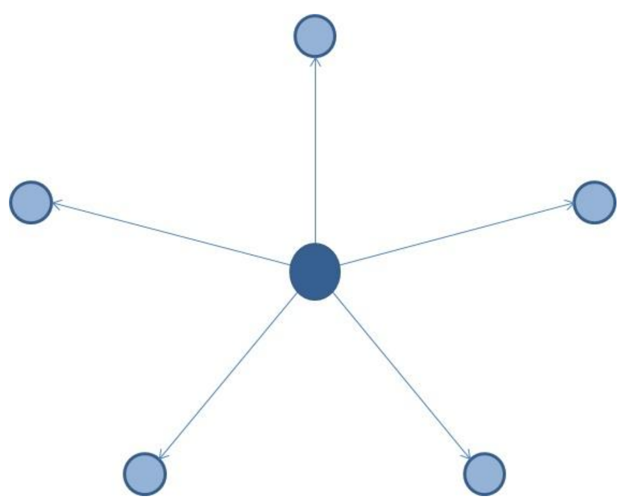

Fig. 2. Star Topology

Mesh Topology

A mesh topology supports multi-hop path communication in which multiple nodes transmit the sensed information to other nodes. In the cases where multi-hop communication is applied, if a node cannot directly transmit its sensed data to the sink node, then it is permitted to utilise in-between nodes to convey the information to the sink node. The mesh network topology supports the scalability in term of adding more number of nodes depending upon the vast usage in various applications. Of course, the transmission range is the main problem in this system. Mostly, it is noticed that multi-hop communications consume more power of the deployed nodes during data transmission and thus increases the energy depletion of nodes due to more number of hops between them. One of the main benefits is that it avoids the data redundancy while sending the data to the sink node.

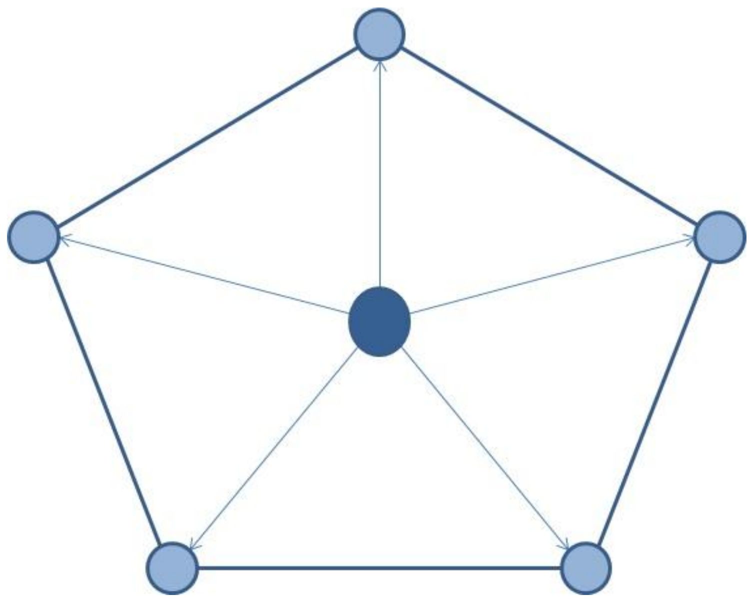

Fig. 3. Mesh Topology

1.5 Components of Wireless Sensor Node

In WSN node, there are four essential parts, for example, detecting unit, handling unit, power and handset unit as shown in figure 4. This additionally has application subordinate extra parts, for example, an area discovering framework, a power generator and a mobilizer. Detecting units are normally made out of two subunits: sensors and ADCs. The analogue signal created by the sensors are changed over to digital signal by the ADC, and after that forwarded to the processing unit. The processing unit is for the most part connected with a little stockpiling unit, and it can oversee the methods that make the sensor node work together with alternate nodes to do the doled out detecting errands. A furthermore handset unit interfaces the node to the system. One of the most critical parts of a sensor node is the power unit. Control units can be upheld by a power rummaging unit, for example, sun based cells. Alternate subunits, of the node, are application subordinate. Particular plan approach gives an adaptable and flexible stage to include the requirements of varied utilizations. Just for instance, contingent upon the conveyed sensors, the flag moulding square should be re-customized and supplanted. All sensor nodes have the capability of remote sensing about varied utilizations with the remote detecting node. Essentially, the wireless connection may be interchanged as per the requirements of the given applications' remote range prerequisite and the bi-directional requirement for efficient power enhancement. A key part of any remote detecting node is its ability to limit the power devoured by the framework.

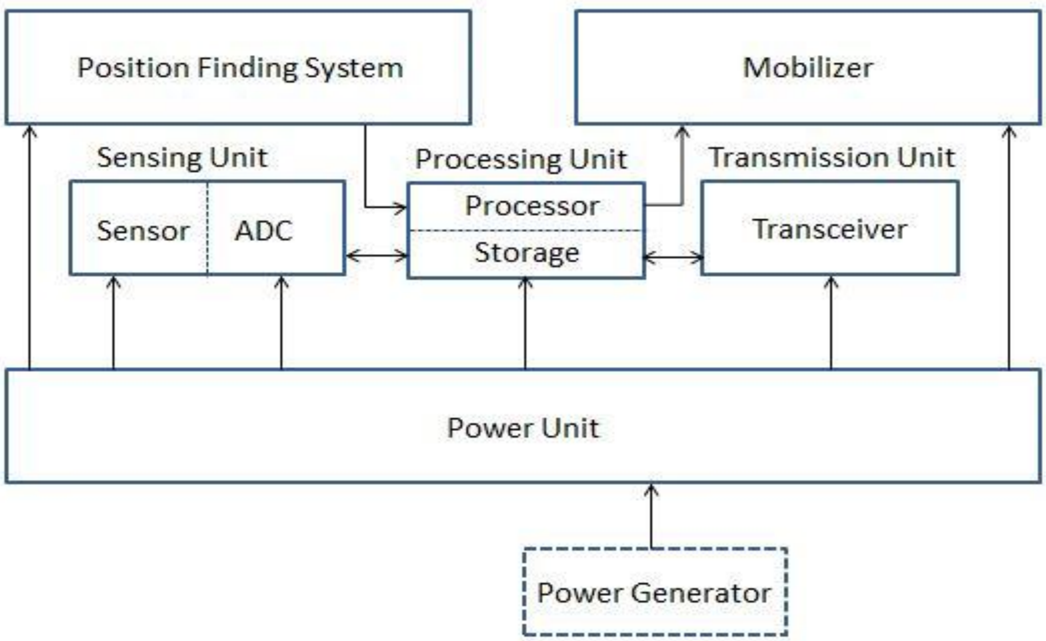

Fig. 4. Components of a Sensor Node

Generally, the wireless subsystem needs to work in measuring the energy. Along these lines, information is conveyed via the radio system whenever it is required. A calculation has to be stacked in the node to decide when to send information in view of the detected occasion. Besides, it is essential to limit the power devoured by the sensor. Consequently, the equipment ought to be intended to enable the microchip to wisely control energy to the radio, sensor, and sensor flag conditioner.

1.6 Communication Model

The sensor nodes are commonly dispersed in the sensor arena as presented in Fig. 1. All these scattered sensor nodes have the abilities to collect information and course this information to the base station. The convention stack utilised by the WSN nodes and the base station as shown in Fig. 5. This convention stack joins routing and energy mindfulness, incorporates information with systems administration conventions, conveys power effectively through the remote medium and advances helpful endeavours of sensor nodes. The convention stack comprises of the physical layer, data link layer, network layer, transport layer, application layer and three administration plane namely task administration plan, mobility administration plane, and power administration plane. Distinctive sorts of application programming can be constructed and utilised on the application layer contingent upon the detecting assignments.

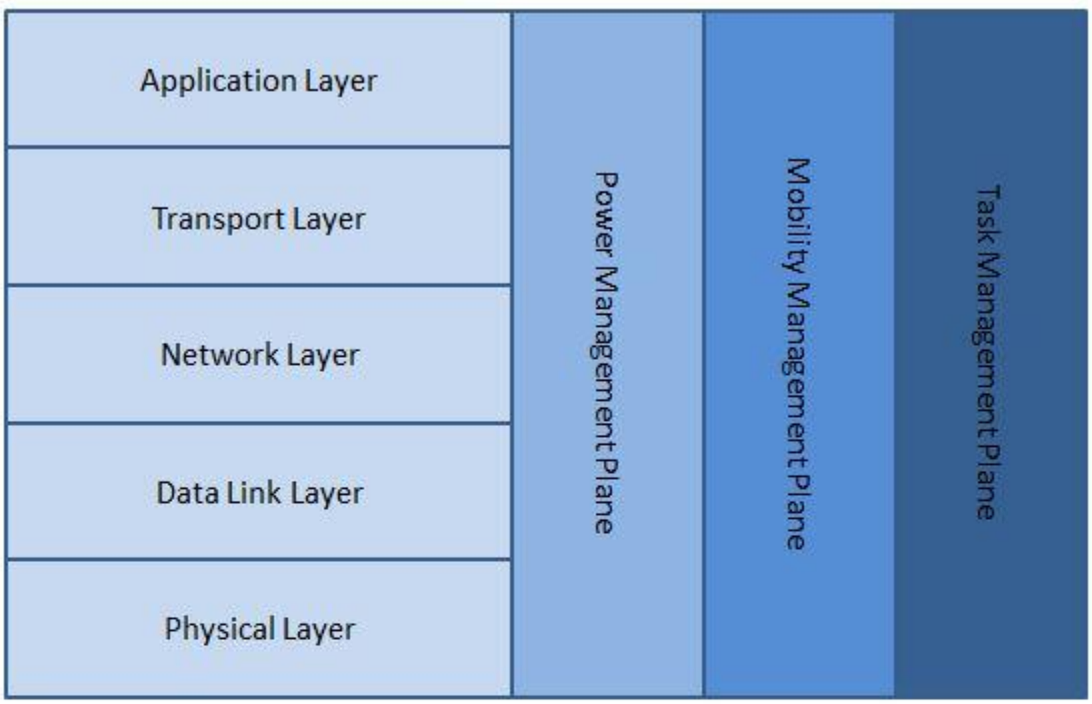

Fig. 5. Communication Model

This model focuses on making equipment and programming of the minimal layer straightforward to the end-client. The vehicle layer keeps up the information stream continuous if the sensor systems require it. The system layer deals with steering the information provided by the transport layer, particular multi-bounce remote steering conventions connecting sensor nodes and sink. The information connects the various layers such as in charge of multiplexed data streams, Media Access Control (MAC), outline location and error control. As we know that the earth is round and sensor nodes are versatile in nature. The MAC convention should be able to handle power failure and ready to limit crash with neighbors communication. Apart from this, even the physical layer includes the necessities of a basic however strong tweak, recurrence choice, information encryption, transmission and getting procedures. Moreover, the power, portability, and undertaking administration planes screen the power, development, and undertaking circulation of the sensor nodes. All the discussed plans aid the sensor nodes to organise the detecting assignment and lessen the general power utilisation.

1.7 Power Dissipation problems

Power dissipation is the most vital component to decide the life of a sensor arrange since as a rule sensor nodes are driven by the battery. Some of the time power streamlining is more convoluted in sensor systems since it included not just lessening of power dissipation additionally drawing out the life of the system much as could be expected. A sensor node mainly consists of four subsystems:

A processing subsystem: It includes a microprocessor which is in charge of the sensors and execution of communication conventions. Microcontroller units for the most part work under different modes for power administration purposes. As these working styles include dissipation of energy, the power dissipation of the different modes needs to be taken care while considering the nodes battery remaining capacity.

Correspondence Subsystem: In this, the short range radio can talk with outside world via neighbouring nodes stationed in the region. Moreover, such radios device can work under the several modes. Thus, there is need to shut down the radio device when it is not transmitting the data to any other radio placed remotely for preserving the energy of the subsystem.

Detecting subsystem: The combination of sensors and actuators mainly connects the several nodes to the outside world. Power dissipation should be lessened upon utilising low energy segments and thus saving the energy.

Energy source subsystem: This subsystem comprises a battery which provides energy to the node. It ought to be understood that the measure of energy drawn from the battery should be watched over. Because if more current is consumed from the same power source for quite a while, the battery will bite the dust quicker even in spite of the fact that it could have continued for a more drawn out time. Typically the appraised current limit of a battery capacity utilised by sensor node is not as much as the base Power dissipation. So, there are provisions for increasing life span of battery by decreasing the current continuously or by shutting down regularly.

For diminishing the power dissipation of WSN networks, distinctive sorts of conventions and calculations are concentrated everywhere throughout the whole domain. The life span of WSN networks must be expanded altogether with working framework related to the application layer. Furthermore, the system conventions are intended to be power mindful. These conventions and calculations must know about the equipment and ready to utilise exceptional elements of the small scale processors. Furthermore, handsets to limit the sensor nodes' power dissipation. This technique may forward the user defined answer for various sorts of sensor nodes plans. Distinctive sorts of sensor nodes sent additionally prompt distinctive sorts of sensor systems. This may likewise prompt the distinctive sorts of community calculations in remote sensor systems field.

CHAPTER 2: RELATED WORK

Many steering conventions in view of grouping strategy for WSNs have shown up in writing. The main low-energy versatile clustering algorithm was LEACH, which was proposed by Heinzelman et al in [1]. In LEACH convention, the cluster head gathers and aggregates detected information from sensor nodes in its individual particular group and pass the accumulated data to the BS straightforwardly. Cluster head may be situated faraway from the base station, so it utilises the vast majority of its energy for transmission, and on the grounds that it is dependably on, it will kick the bucket speedier than different nodes. The Low-Energy Adaptive Clustering Hierarchy (LEACH) clustering system is supported by two key assumptions: (1) All nodes transmit their information to a solitary sink node; and (2) All nodes have the capacity to discuss specifically with the sink node. Keeping in mind the end goal to balance the system energy utilisation, the LEACH convention actualizes a heap adjusting system that permits the person nodes to end up CH at various rounds. For each round, organise nodes select an irregular incentive in the vicinity of zero and one. The node chooses itself as a cluster head for the current round if the number is less than the threshold.

$$T(n) = \frac{p}{1-p*\left(r mod\left(\frac{1}{p}\right)\right)}, if\ n \in G \qquad (1)$$

$$T(n) = 0, otherwise \qquad (2)$$

Here, r is the current round number, p is the desired percentage of clusters, and G is the set of nodes that have not been selected as CH in the last 1/ p rounds.

LEACH beats static clustering calculations by obliging nodes to undertake to be high energy group heads and adjusting the relating clusters in view of the nodes that are cluster heads at a given time. At various times, every node has the weight of securing information from the nodes in the group, combining the information to get a total flag, and transmitting this total flag to the sink node. LEACH is completely distributed, i.e., it does not require any control information from the sink node and the nodes need no information about the global network altogether for LEACH to work. Appropriating the energy among the nodes in the system is viable in decreasing energy dispersal from a worldwide point of view and upgrading framework lifetime. In particular, our recreations demonstrate that:

- LEACH decreases correspondence energy by as much as eight times contrasted and coordinate transmission and least transmission-energy directing.

- The principal node passing in LEACH happens more than eight times later than the primary node demise in coordinate transmission, least transmission-energy steering, and a static clustering convention, and the last node passing in LEACH happens more than three times later than the last node passing in the different conventions.

In [3], a new Genetic Algorithm-based Energy-Efficient Protocol (GAEEP) has been exhibited to proficiently amplify the lifetime and steadiness time of wireless sensor network. GAEEP utilises hereditary calculation to make strides the system lifetime and solidness time of the remote sensor organises by finding the ideal number of group heads and their areas in light of limiting the energy utilisation of the sensor nodes. Matlab re-enactment comes about demonstrated that the proposed GAEEP convention is more energy proficient and more solid in clustering process as contrasted with LEACH, SEP, A-LEACH and DEU conventions in low or high thick systems and in homogeneous or heterogeneous systems. Likewise, GAEEP convention builds the dependability of clustering process since it grows the soundness period and packs the precariousness time frame. Additionally, it outflanks the past conventions regarding energy scattering rate, organise lifetime and steadiness period in both homogeneous and heterogeneous cases.

In [4], the increment in the lifetime of the WSN is accomplished in an adjusted sense utilising the proposed calculation GAECH. The exploratory outcomes have demonstrated that GAECH performs superior to anything its different partners like LEACH in three different organise situations. The increment in execution is particular when the BS is situated far away from the system, which is essentially conceivable in the greater part of the on-going applications. As the best case, GAECH indicates 21.11% expanded lifetime than its partner when the BS is found outside of the system. Additionally, GAECH is observed to be rationing energy by adjusting energy utilisation among the groups. Contingent on the application requirements, the different weight coefficients said in the wellness capacity of GAECH can be changed to get better outcomes. The future inquiries about bearing would associate with utilising different parameters like node degree and leftover energy of a node in the wellness work.

In [5], the authors proposed SEP (Stable Election Protocol)in which each sensor node in a heterogeneous two-level various levelled arrange autonomously chooses itself as a cluster head in view of its underlying energy in respect to that of different nodes. SEP is dynamic in that we don't accept any earlier dissemination of the diverse levels of energy in the sensor nodes.

Moreover, our investigation of SEP is most certainly not just asymptotic, i.e. the examination applies similarly well to small-sized systems. Dissimilar to [6], SEP does not require any worldwide information of energy at each decision round. At long last SEP is versatile as it doesn't require any information of the correct position of every node in the field. The creators have proposed to stretch out SEP to manage clustered sensor systems with more than two levels of the chain of importance and more than two sorts of nodes. They are currently executing SEP in Berkeley/Crossbow bits and looking at sending issues counting dynamic updates of weighted voting probabilities in light of current heterogeneity conditions and also the combination of SEP with MAC conventions that can give low-cost data about the conveyance of energy in the region of every node. SEP code and research come about are freely accessible at http://csr.bu.edu/sep.

In [8], the K-medoids LEACH (KLEACH) convention was portrayed to enhance the grouping and CHs determination methodology. For the first round of correspondence, in setup stage, the K-medoids calculation was utilised for cluster development, which guarantees uniform grouping.

Genetic Algorithm based (LEACH-GA) proposed in [9] decides the ideal edge likelihood for cluster arrangement in WSNs utilising GA. LEACH-GA enhances the CHS edge work, yet at the same time CHs are arbitrarily chosen, and the lingering energy of every node is not considered in cluster arrangement.

Another variant of LEACH called TL-LEACH (Two-Level LEACH) was displayed in [12]. In this convention, another level in the group is added to the regard of LEACH that considers just one level. This new level diminished the utilisation energy for transmitting particularly when the CH situated for far from BS. The work portrayed in [13] proposed V-LEACH (Vice-CH LEACH) convention. In V-LEACH other than having a CH in the group, the sensor network additionally consists of a vice-cluster head that plays the part of the cluster head and makes the cluster dependably associated with BS when the CH expires. The LEACH convention requires the client to indicate the fancied likelihood of CHs for use with the edge work in deciding if a node turns into a CH or not. Another convention called Amend LEACH (A-LEACH) was produced in [14], for choosing CHs in a conveyed design in two-level progressive WSNs. Contrasting LEACH, A-LEACH is heterogeneous conscious, as in selection probabilities are biased by the underlying vitality of a node in respect to that of different nodes in the system. This drags out the time interim before the demise of the main node, which is significant for some applications.

In [15], Enhanced LEACH (En-LEACH) convention was embraced to deal with CH disappointment and to represent the non-uniform and dynamic lingering energy of the nodes. Energy LEACH (E-LEACH) calculation in [16] presented the idea of the energy limit, and a separation calculates choosing CH. Energy limit is to decide if the node can be utilised as an essential of the CH node and the separation variable is utilised to select the way with the littlest information transmission separate.

Another Distributed Energy-efficient Unequal Clustering convention (DEU) was displayed in [17] to adjust vitality utilisation among CHs. DEU joins uneven clustering and multi-bounce steering system. Additionally, DEU convention utilises the clock to choose the nodes with high vitality as CHs to spare the vitality of nodes in the system. CHs consider competitor nodes' ideal jump number, lingering vitality, inter and intra cluster correspondence cost to choose hand-off nodes.

CHAPTER 3: ROUTING IN WIRELESS SENSORS NETWORKS

Wireless sensor systems have enlarged notable recognition because of their adaptableness in handling of issues in various fields and can change our livelihood in a widespread range of ways. Wireless sensors networks have been successfully associated with diverse application spaces. Owing to a huge number of nodes in the system and the complications of the environment, it is tough and even difficult to replace or recharge batteries for the sensor nodes. Keeping in mind the end goal to viably use wireless sensors network we have to reduce the energy dissipation while cluster formation and during the exchange of information between the WSN nodes and the sink node.

3.1 Direct Communication

It is the simplest way of communication between a sensor node and the sink node. In this communication protocol, the sensor node sends its data to the sink node directly. The direct communication between the sensor node and the base station makes for a simple communication protocol implementation, but it has a limitation. As the energy consumed during communication depends on the distance between two parties, the nodes that are greater distance away from the sink node will be drained out of power much earlier than the nodes that are at much closer distance from the sink node. This faster draining of power will cause an unbalanced network which is not desirable in WSNs.

3.2 Multi-Hop Communication

Another way to communicate is through multi-hop communication. Multi-hop communication involves transmission of data to the sink node via one or more intermediary nodes. The nodes that are greater distance away from the base station transmit their data to some other node which in turn forwards it to another node or the base station. This way of communication may occur to have overcome the limitation of the direct communication method, but it also has its own limitation. In this method, the nodes that act as intermediary nodes drain out of power faster than other nodes. Hence the nodes nearer to the sink node are more probable to drain out of power than that are at much greater distance from the base station. So there came a need for some other method for information exchange between sensor nodes and the sink node. Another problem that ascends during exchange of information between sensor nodes and the sink node is the transfer of a lot of redundant data from sensor nodes to the sink node. Most of the data sensed by sensors that are near to each other are redundant, and this data is forwarded to the sink node. Unwanted energy is wasted to transmit this data to sink node. If somehow this redundancy can be removed, network lifetime can be enhanced multiple folds.

3.3 Need of Clustering

To process the data sensed by the sensor nodes, the sensor nodes have to send this data to the base station. Two major problems come into picture while the communication between the sensor node and the base station takes place. The first problem is the unbalanced network due to uneven consumption of energy in the sensor nodes. This problem can be explained as follows. The communication between the sensor node and the base station can take place in two ways. The first way is direct communication method. It is the simplest way of communication between a sensor node and the sink node. In this communication protocol, the sensor node sends its data to the sink node directly. The direct communication between the sensor node and the base station makes for a simple communication protocol implementation, but it has a limitation. As the energy consumed during communication depends on the distance between two parties, the nodes that are greater distance away from the sink node will be drained out of power much earlier than the nodes that are at much closer distance from the sink node. This faster draining of power will cause an unbalanced network which is not desirable in WSNs. Another way to communicate is through multi-hop communication. Multi-hop communication involves transmission of data to the sink node via one or more intermediary nodes. The nodes that are greater distance away from the base station transmit their data to some other node which in turn forwards it to another node or the base station. This way of communication may occur to have overcome the limitation of the direct communication method, but it also has its own limitation. In this method, the nodes that act as intermediary nodes drain out of power faster than other nodes. Hence the nodes nearer to the sink node are more probable to drain out of power than that are at much greater distance from the base station. So there came a need for some other method for information exchange between sensor nodes and the sink node.

Another problem that ascends during exchange of information between sensor nodes and the sink node is the transfer of a lot of redundant data from sensor nodes to the sink node. Most of the data sensed by sensors that are near to each other are redundant, and this data is forwarded to the sink node. Unwanted energy is wasted to transmit this data to sink node. If somehow this redundancy can be removed, network lifetime can be enhanced multiple folds.

The solution to these problems is to group the sensors into small groups. These groups are known as clusters. This partitioning of the wireless sensors network into clusters is called as clustering. All the clusters have their leader called cluster head. Every other member of the cluster sends its data to its cluster head. The cluster heads may directly forward all the data received to the sink node. Else, it can remove the redundancy from data collected and then forward it to the sink node. This way clustering solves the problem of transfer of redundant data from the sensor node to base station. The problem of the unbalanced network remains if static clustering is used. Static clustering means the clusters once formed are not changed. The cluster head remains same for the lifetime of the network. Now since the cluster head dissipates much more energy than the other sensor nodes, it will drain out of power much faster than other nodes. Hence dynamic clustering is used in this thesis. In dynamic clustering, the clusters and the cluster heads keep changing. The cluster head should be chosen with care. The performance of the algorithm essentially depends on the formation of clusters and selecting the cluster heads.

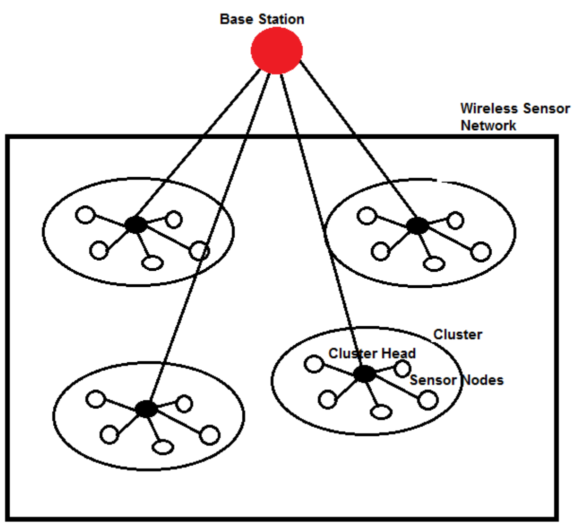

Fig. 6. Clustering in Wireless Sensor network

3.4 Major Clustering Algorithms

3.4.1 LEACH

Low-energy adaptive clustering hierarchy or LEACH is a self-organizing, clustering protocol that uses a threshold value to decide the cluster heads. The threshold for every round is determined, and then the nodes generate a random number. If the random number generated is larger than the threshold, the node gets to be a cluster head for that round. Thus LEACH employs random distribution of energy load among the sensor nodes. LEACH is the clustering algorithm that focuses on extending the network lifetime by reducing the energy consumption per round. To accomplish these targets, LEACH receives a progressive way to deal with sort out the system into an arrangement of clusters. Each cluster is overseen by a chosen group head. The cluster head accepts the accountability to complete numerous undertakings. The essential operations of LEACH are sorted out in two particular stages. The primary stage, the setup stage, comprises of two stages, group head determination and cluster development. The second stage, the enduring state stage, concentrates on information accumulation, collection, and conveyance to the base station. Toward the start of the setup stage, a series of group head choice begins. The group head choice process guarantees that this part pivots among sensor nodes, in this manner appropriating energy utilisation equally overall system nodes. To decide whether the ball is in its court to end up noticeably a cluster head, a node, n, produces an arbitrary number, v, in the vicinity of 0 and 1 and analysed it to the cluster head determination limit. The node turns into a group head if it's produced esteem, v, is not as much as T (n). The group head determination limit is intended to guarantee with a high likelihood that a foreordained portion of nodes, P, is chosen cluster heads at each round. Further, the edge guarantees that nodes which served in the last 1/P rounds are not chosen in the current round.

$$T(n) = \frac{p}{1-p*\left(rmod\left(\frac{1}{p}\right)\right)}, if\ n \in G \quad (1)$$

$$T(n) = 0, otherwise \quad (2)$$

Here, p is the desired percentage of cluster heads, r is the current round number and G is the set of nodes that have not been selected as CH in the last 1/ p rounds.

The variable G speaks to the arrangement of nodes that have not been chosen to turn into group heads in the last 1/P rounds, and r signifies the current round. The predefined parameter, P, speaks to the group head likelihood. Plainly if a node has filled in as a cluster head in the last 1/P rounds, it won't be chosen in this round. Toward the finishing of the cluster head choice process, each node that was chosen to wind up plainly a cluster head publicises its new part to whatever remains of the system. After accepting the group head notices, each residual node chooses a group to join. The fruition of the setup stage flags the start of the consistent state stage. Amid this stage, nodes gather data and utilise their dispensed openings to transmit to the cluster head the information gathered. This information gathering is occasionally performed. Recreation comes about demonstrate that LEACH accomplishes noteworthy energy investment funds.

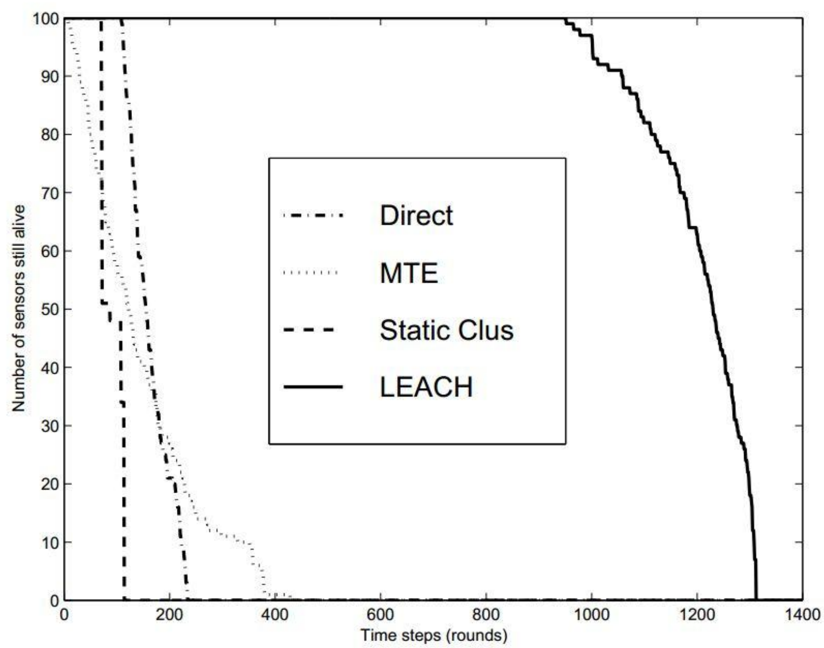

Fig. 7. LEACH Result: Number of alive nodes

These investment funds depend essentially on the information collection proportion accomplished by the cluster heads. Regardless of these advantages, in any case, LEACH endures a few weaknesses. The suspicion that all nodes can achieve the base station in one jump may not be reasonable, as capacities and energy stores of the nodes may fluctuate after some time starting with one node then onto the next. LEACH is a totally circulated calculation, requiring no control data from the base station. The cluster administration is accomplished locally, which decimates the requirement for worldwide system information. Besides, information accumulation by the cluster likewise contributes extraordinarily to energy forgiving, as nodes are never again required to send their data specifically to the sink. It has been indicated utilising recreation that LEACH beats customary directing conventions, including coordinate transmission and multi hop directing, least transmission-energy steering, also, static clustering–based steering calculations.

3.4.2 PEGASIS

The convention goes for expanding the lifetime of a system by accomplishing an abnormal state of vitality productivity and uniform energy utilisation overall system nodes. Second, the convention endeavours to decrease the hold-up that information brings about on their way to the sink. The system demonstrates considered by PEGASIS accept a homogeneous arrangement of nodes sent over a topographical range. Nodes are accepted to have worldwide information about other sensors' positions. Moreover, they can control their energy to cover discretionary reaches. The objective is to create a steering structure and a conglomeration plan to lessen energy utilisation and convey the collected information to the base station with insignificant deferral while adjusting energy utilisation among the sensor nodes. As opposed to different conventions, which depend on a tree structure or a group based progressive association of the system for information social affair and scattering, PEGASIS utilises a chain structure. Nodes speak with their nearest neighbours. The development of the chain begins with the most distant node from the sink. Arrange nodes are added to the chain dynamically, beginning from the nearest neighbour to the end node. Nodes that are as of now outside the anchor are added to the chain in a ravenous form, the nearest neighbour to the top node in the present chain to start with until all nodes are incorporated. A node inside the affix is chosen to be the chain pioneer. Its duty is to transmit the amassed information to the base station. The chain pioneer part moves in situating the chain after each round. Rounds can be overseen by the information sink, and the move starts with one round then onto the next can be stumbled by a powerful reference point issued by the information sink. Revolution of the influential position among nodes of the chain guarantees, by and large, an adjusted utilisation of vitality among all the system nodes. The chain pioneer issues a token to the last node in the correct end of the chain. After getting the token, the end node transmits its information to its downstream neighbour in the chain at the pioneer. The neighbouring node totals the information and transmits them to its downstream neighbour. This procedure proceeds until the accumulated information come to the pioneer. After accepting the information from the correct side of the chain, the pioneer issues a token to one side end of the chain, and a similar conglomeration process is completed until the information achieve the pioneer. After accepting the information from both sides of the chain, the pioneer totals the information and transmits them to the information sink. The chain-based twofold approach prompts huge vitality diminishment, as nodes work in an exceedingly parallel way. Besides, since the various levelled, treelike structure is adjusted, the plan ensures that after $\log_2 N$ steps, the collected information touch base at the pioneer. The chain-based double accumulation conspires been utilised in PEGASIS as another option to accomplishing a high level of parallelism. With CDMA-skilled sensor nodes, it has been demonstrated that the plan performs best as for the energy defer item required per round of information assembling, a metric that adjusts the energy and defers the cost.

3.4.3 SEP

In SEP, the effect of heterogeneity of nodes is studied in wireless sensors networks that are progressively grouped. In these systems, a portion of the nodes move toward becoming group heads, total the information of their cluster individuals, what's more, transmit it to the sink. We expect that a rate of the populace of sensor nodes is outfitted with extra vitality assets—this is a wellspring of heterogeneity which may come about from the underlying setting or as the operation of the system develops. We likewise expect that the sensors are haphazardly (consistently) dispersed and are not versatile, the directions of the sink and the measurements of the sensor field are known. We demonstrate that the conduct of such sensor systems turns out to be extremely precarious once the main node kicks the bucket, particularly within sight of node heterogeneity. Established grouping conventions expect that all the nodes are furnished with a similar measure of vitality, and as a result, they cannot take the full preferred standpoint of the nearness of node heterogeneity. SEP is a heterogeneous-mindful convention to drag out the time interim before the demise of the main node (we allude to as steadiness period), which is pivotal for some applications where the criticism from the sensor organise must be dependable. SEP depends on weighted decision probabilities of every node to wind up cluster go to the rest of the energy in every node. We appear by recreation that SEP dependably drags out the solidness period contrasted with (and that the normal throughput is more prominent than) the one got utilising current clustering conventions. We finish up by concentrating the affectability of our SEP convention to heterogeneity parameters catching vitality irregularity in the network. SEP yields longer steadiness district for higher estimations of additional energy brought by more intense nodes.

CHAPTER 4: PROPOSED WORK

Wireless sensor systems have enlarged notable recognition because of their adaptability in taking care of issues in various application spaces and can possibly change our lives in a wide range of ways. WSNs have been effectively connected with diverse application spaces. Clustering in wireless sensors network is a significant aspect that decides the lifetime of the system. Even many of the well-known clustering methods have been designed keeping in mind the energy consumption and network scalability. The aim of clustering has always been on finding clusters set, reducing overall power consumption in the network, and the lifetime enhancement of the network. The increase in the lifetime comes as a result of proper load balancing among the sensor nodes. In this thesis, a novel clustering algorithm, GA-MEANS, is proposed to increase the lifetime of the network and to ensure uniform distribution of cluster heads across the sensing area. The proposed algorithm is a cross of genetic algorithm and k-means clustering algorithm. The results have been compared with other algorithms and indicate the better performance of as compared to other algorithms.

4.1 Problem Statement

After a comprehensive survey of available clustering algorithms, it has been found that less focus is made on choosing optimal number of cluster while cluster formation. The number of clusters formed need not very high or very low for enhancement of network lifetime. Therefore choosing an optimal number of clusters before cluster formation is recommended to achieve better results.

4.2 Proposed Solution

Clustering is widely used to achieve balanced network stability and prolonged network lifetime. Clustering can also help in the aggregation of data to reduce the redundancy in the data received to lessen the workload of sensor nodes to spare energy and in this manner increment the general lifetime of the network. The k-means algorithm has been considered to be one of the best clustering algorithms. But the output of k-means is very much dependent on the value of k, i.e., the number of clusters and this value of K taken as input from the user. The user, most of the times, does not calculate the optimal value of k and hence the performance of the algorithm may not come out to be the best. To get best results while using k-means, we need to know the optimal value of k, i.e. the number of clusters. This problem is considered to be computationally demanding (NP-Hard problem). Therefore, Genetic Algorithm has been successfully applied in several computationally difficult optimisation problems. In this thesis, a hybrid of K-Means and Genetic Algorithm has been developed to get an optimal clustering of the sensor network.

4.2.1 Genetic Algorithm

The genetic algorithm is a meta-heuristic, nature-inspired optimisation problem-solving technique that belongs to the larger class of evolutionary algorithms (EA). It is based on natural selection, the rule that encourages biological evolution.

In the genetic algorithm, an initial random population of solutions is created, and the optimal solution is searched by evaluating each individual of the population. The population is repeatedly modified for the next generation. The population for next generation is obtained from the current population. The children for the next generation are produced by the individuals randomly selected from the present population. The population "evolves" towards an optimal solution over succeeding generations. The three main operations used to obtain next generation are:

- **Selection:** Selection operation reproduces the individuals with better fitness to be the part of next generation. Different selection techniques that can be used are Roulette wheel selection, rank, steady state, etc. The selection technique used depends on the application.
- **Crossover:** Crossover operation combines two parent individuals by exchanging their genetic information to form children for the next generation.
- **Mutation:** Mutation operation applies random variations to individual parent solution to create children. It helps in searching the solution from a new area.

The evaluation of each individual is done on the basis of fitness value. Fitness value of an individual tells how close this individual solution is to the optimal solution. The fitness value is calculated using a fitness function. The fitness function is decided after several hit and trials to get the optimum result.

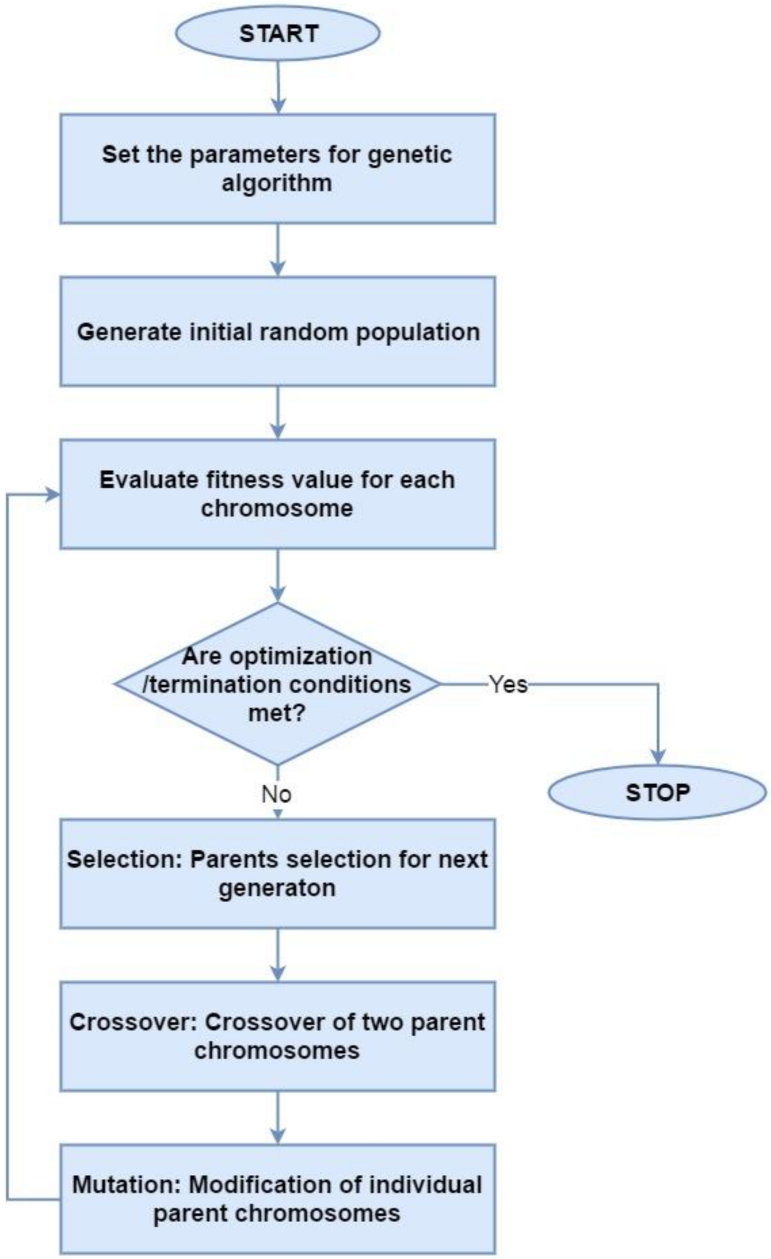

Fig. 8. Flowchart of Genetic Algorithm

4.2.2 GA-MEANS Algorithm

If a wireless sensors network is given, it is a computationally hard problem to find the optimal number of clusters to be formed. It belongs to the NP-Hard class of problems in computational complexity theory. Many heuristic algorithms have been applied to solve this class of problems.

Genetic Algorithm has been successfully applied to many of such optimisation problems. Hence we propose to use Genetic Algorithm for finding the optimal number of clusters to be formed to optimise the performance of K-Means. Once the optimal number of clusters is found, then k-means can be applied straightaway.

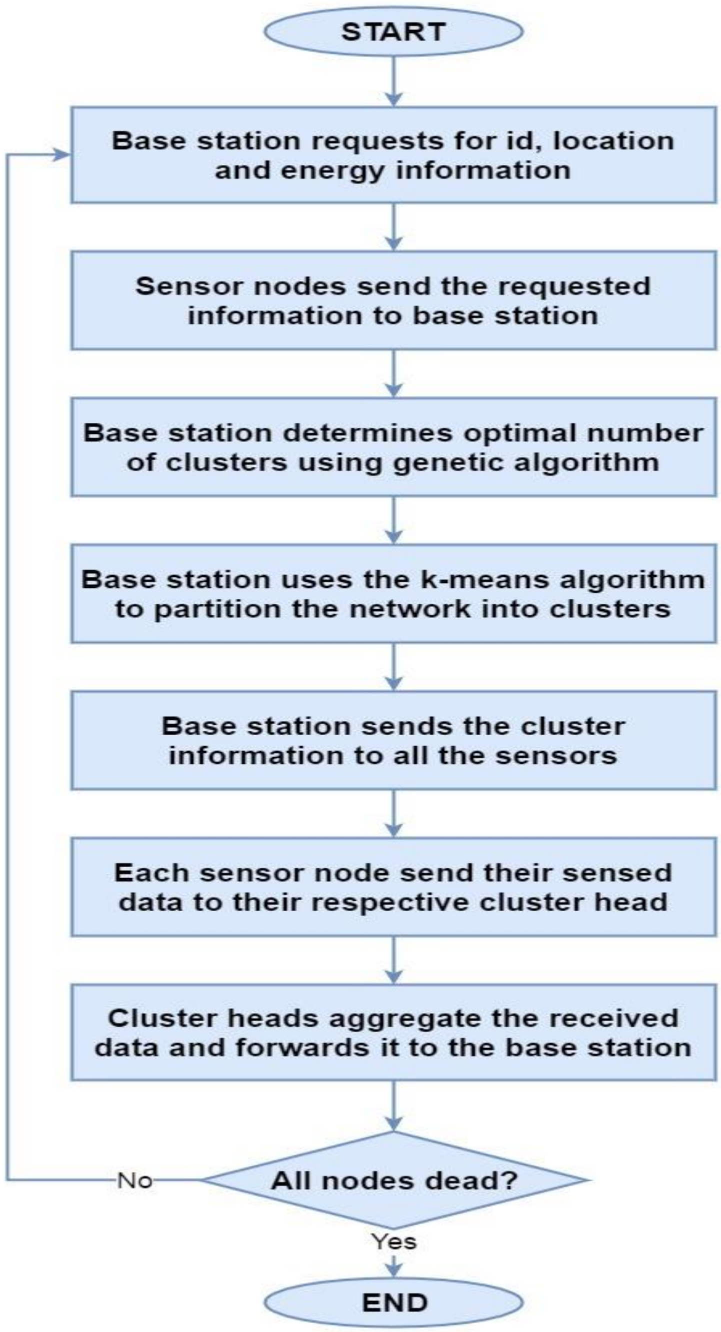

Fig. 9. Flowchart of GA-MEANS

GA-MEANS algorithm proceeds in rounds with a preparation round at the beginning of the first round. In the preparation round, the base station sends a request for location and id information to all the sensor nodes, and then all the sensor nodes respond by sending the requested information. Each round of GA-MEANS consists of two phases. In the first step, the cluster head selection phase, the base station gathers energy information from each sensor node and uses this information to form clusters and choosing the cluster head for each cluster. In the second phase, the data transmission phase, the actual sensed data is sent from sensor nodes to their respective cluster heads and then from cluster heads to the base station.

4.2.2.1 Cluster Head Selection Phase

In this phase, the optimal number of clusters is determined and then clusters are formed. The sensor nodes send their energy information as a control packet to the base station. Based on the received information, the base station applies genetic algorithm to ascertain the number of clusters for the current round to minimise the dissipation of energy during the communication. The k-means algorithm uses this number to create the clusters in the network. The k-means algorithm returns cluster centroids of each cluster. The cluster head of each cluster is selected on the basis of the distance from the cluster centroid. The sensor node nearest to the cluster centroid with its energy greater than or equal to the average energy of the cluster is chosen as the cluster head of that cluster for that round. Once the clusters are formed, and the cluster head for each cluster is determined, the base station sends control packets to cluster heads informing them about the member nodes in its cluster. The base station then informs all the other sensor nodes about their respective cluster heads. The cluster heads, based on the received information, formulates a TDMA schedule notifying its member nodes about the slots in which they can transmit the data.

i. Working of GA

The base station applies GA on a subset, S' of the sensors set, S such that sensor S_i belongs to S' if the energy of S_i is greater than equal to the average energy of all the active nodes in the sensor network. The initial population consists of bit strings each of length equal to the size of S'. The bits in the binary stream represents sensor nodes of the set S'. If the bit is '0', then the corresponding node is a common node else it is a cluster head. Each bit string in the population represents an individual solution.

ii. Determining the optimal number of clusters

Genetic Algorithm stops when the change in the fitness value of the individuals does not change much in successive generations, or a given number of generations are completed. The individual with best fitness value of the last population is returned as the solution. The returned solution is a stream of zeros and ones. The number of ones in this solution gives the optimal number of cluster heads for the given round.

iii. K-Means

The optimal number of clusters as obtained after using genetic algorithm is used as input 'k' to the k-means algorithm. K-means clustering is a non-hierarchical type of clustering that partitions the network into mutually exclusive clusters with each sensor belonging to the cluster with the nearest mean. The K-means algorithm clusters the network, but the cluster centroid returned may not be a sensor node. Hence the real cluster head is selected by the distance from the cluster centroid. The sensor node nearest to the cluster centroid with its energy greater than or equal to the average energy of the cluster is selected as the real cluster head for that round.

4.2.2.2 Data Transmission Phase

In data transmission phase, the sensor nodes sense their surrounding environment and send their data to their respective cluster heads. The data is sent by the TDMA schedule generated by the cluster heads. After receiving data from all the sensor nodes, the cluster head aggregates the data and forwards the aggregated data to the base station. After a certain time, the cluster head selection phase starts and new clusters are determined for the next round.

CHAPTER 5: SIMULATION AND RESULT

The simulation of the above algorithm has been carried out using MATLAB. The network parameters used are same as used in previous algorithms. The network area is a 50m*50m field with 100 sensor nodes randomly uniformly distributed over the network area. The base station is located at x=-25m and y=-100m coordinate. Each sensor node is given an initial energy E0=0.5J. The energy is consumed as per the radio model described.

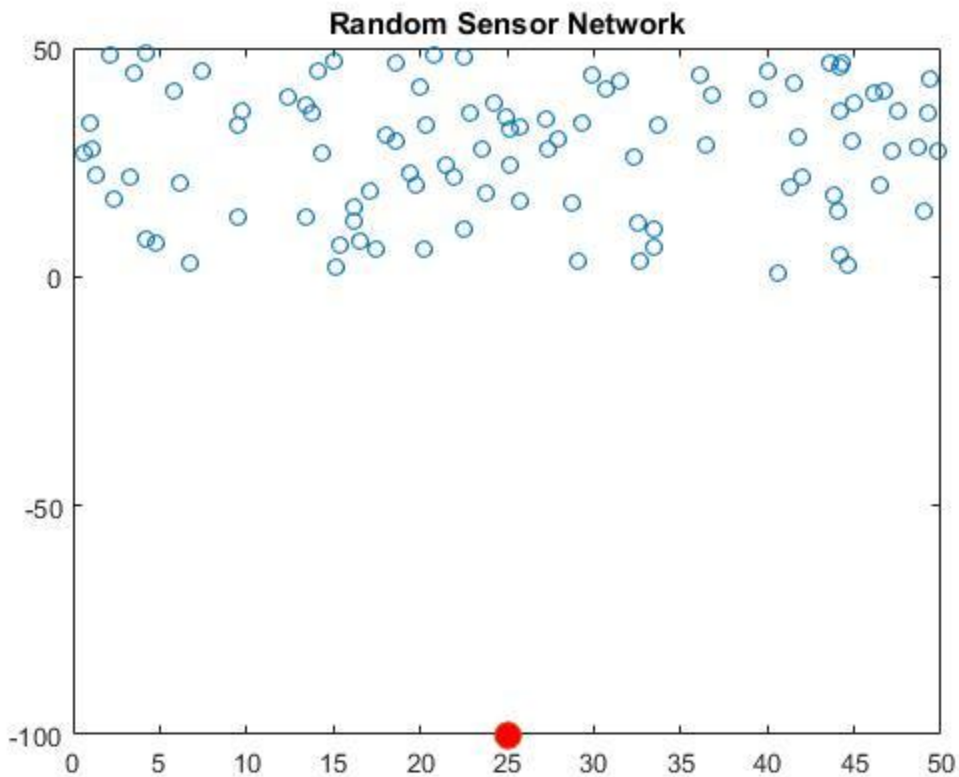

Fig. 10. Random Sensor Network

The parameters used in the genetic algorithm are described as follows. The initial population for every round is generated randomly. The population type is 'bitstring', i.e., the population consists of zeros and ones only. The population size chosen is 50. The maximum number of generations after which the GA stops is 150. The population creation function used is uniform creation function. The selection function is Roulette selection, the crossover function used is Single Point Crossovers and the mutation function used is Uniform Mutation.

Parameter	Value
Population Size	50
Population Type	'BitString'
Maximum Generation	150
Population Creation Function	Uniform Creation Function
Selection Function	Roulette selection
Crossover Function	Single Point Crossovers
Mutation Function	Uniform Mutation

TABLE I. Genetic Algorithm parameters

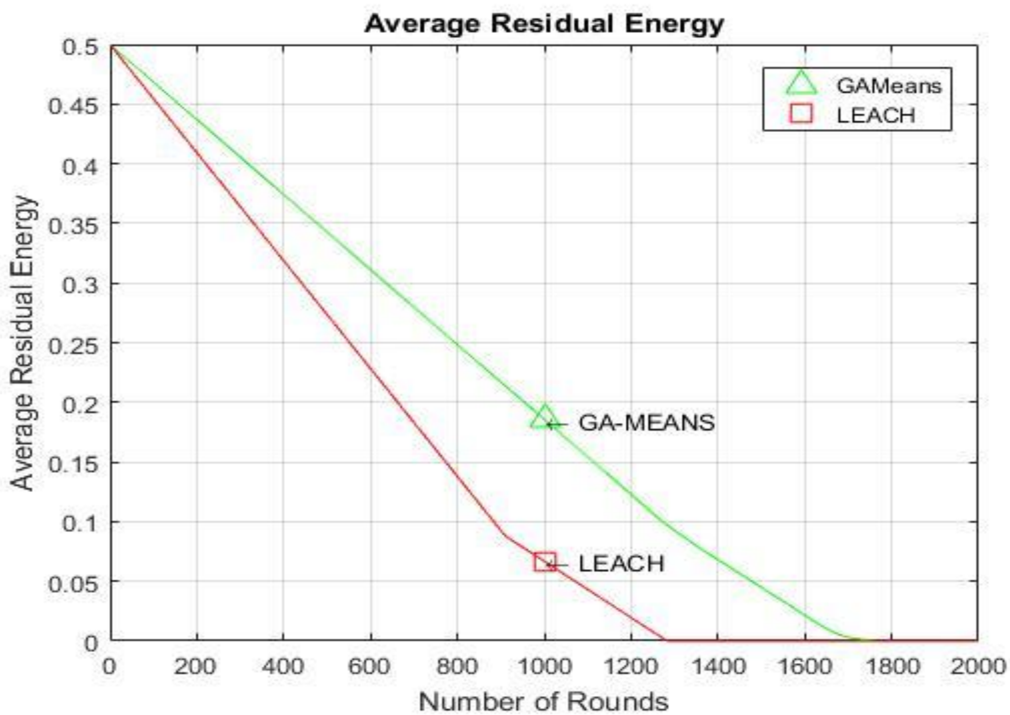

Fig. 11. Average Residual Energy after each round

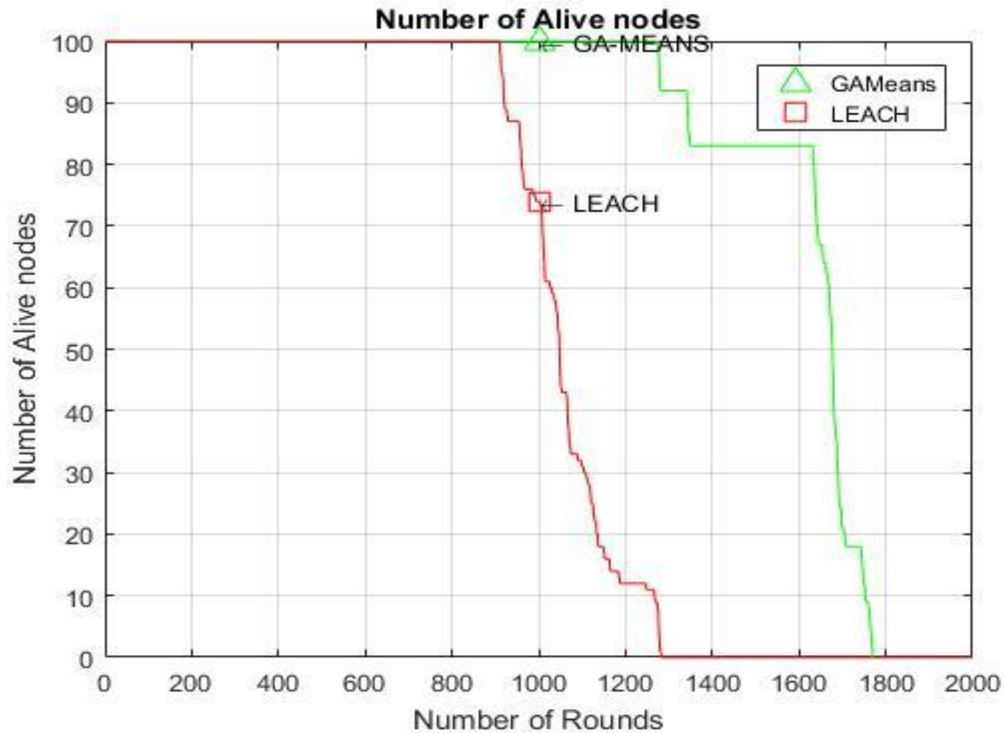

Fig. 12. Number of Alive Nodes after each round

Comparison Factor	**LEACH**	**GA-Means**
First Node Die	911	1274
All Nodes Dead	1283	1770
Average Energy Dissipated per round (in Joules)	0.000389711	0.000282486

TABLE II. Comparison of LEACH and GA-Means

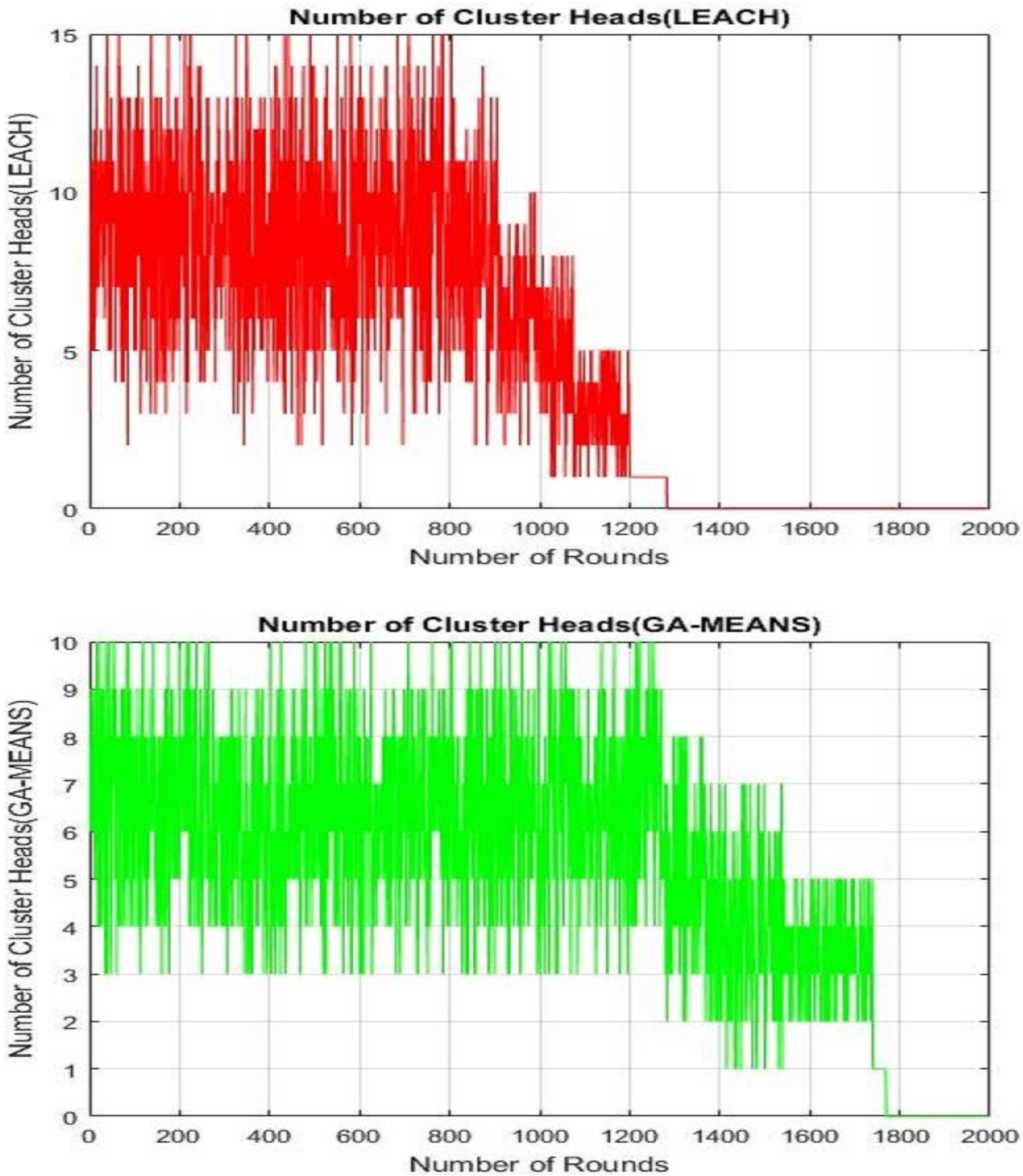

Fig. 13. Number of Cluster heads in each round

CHAPTER 6: CONCLUSION

In this thesis, a new algorithm, GA-Means clustering algorithm, is proposed with the aim to optimally cluster the wireless sensor network. The central focus of the proposed algorithm is to partition the WSN into optimal number of clusters. The algorithm successfully enhances the lifespan of the wireless sensors network. The algorithm also maximises the stability period of the network. The GA-Means algorithm uses the genetic algorithm to find an optimal number of clusters on the grounds of residual energy, number of live nodes and distance from the base station, and the k-means algorithm efficiently clusters the network. The results show that the prosed GA-Means is better than LEACH in terms of network lifespan, stability period and reliability.

CHAPTER 7: REFERENCES

[1] Wendi Rabiner Heinzelman, Anantha Chandrakasan and Hari Balakrishnan, “Energy-Efficient Communication Protocol for Wireless Microsensor Networks”, Hawaii International Conference on System Sciences, January 4-7, 2000.

[2] Lin Yu Tseng, Shiueng Bien Yang, “A genetic approach to the automatic clustering problem”, Pattern Recognition 34 (2001) 415-424.

[3] Mohammed Abo-Zahhad, Sabah M. Ahmed, Nabil Sabor and Shigenobu Sasaki, “A New Energy-Efficient Adaptive Clustering Protocol Based on Genetic Algorithm for Improving the Lifetime and the Stable Period of Wireless Sensor Networks”, International Journal of Energy, Information and Communications Vol.5, Issue 3 (2014), pp.47-72.

[4] B. Baranidharan and B. Santhi, “GAECH: Genetic Algorithm Based Energy Efficient Clustering Hierarchy in Wireless Sensor Networks”, Volume 2015, Article ID 715740, 8 pages.

[5] Georgios Smaragdakis, Ibrahim Matta and Azer Bestavros, “SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks”, Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA), (2004).

[6] Goldberg, David E., Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley, 1989.

[7] Hwei-Jen Lin, Fu-Wen Yang and Yang-Ta Kao, “An Efficient GA-based Clustering Technique”, Tamkang Journal of Science and Engineering, Vol. 8, No 2, pp. 113122 (2005).

[8] P. Bakaraniya and S. Mehta, “K-LEACH: An improved LEACH Protocol for Lifetime Improvement in WSN”, International Journal of Engineering Trends and Technology (IJETT), vol. 4, no. 5, (2013) May, pp. 1521-1526.

[9] J. L. Liu and C. V. Ravishankar, “LEACH-GA: Genetic Algorithm-Based Energy-Efficient Adaptive Clustering Protocol for Wireless Sensor Networks”, Int. J. of Machine Learning and Computing, vol. 1, no. 1, (2011) April, pp. 79-85.

[10] Eduardo Raul Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, and Andre C. Ponce Leon F. de Carvalho, “A Survey of Evolutionary Algorithms for Clustering”, IEEE Transactions on Systems, Man, and Cybernetics - part C: Applications and Reviews, vol. 39, no. 2, March 2009.

[11] Wenliang Wu, Naixue Xiong and Chunxue Wu, "Improved clustering algorithm based on energy consumption in wireless sensor networks", IET Networks, 2017, Volume: 6, Issue: 3, Pages: 47 – 53.

[12] V. Loscrì, G. Morabito and S. Marano, "A Two-Level Hierarchy for Low-Energy Adaptive Clustering Hierarchy", IEEE 62nd Vehicular Technology Conference, (2005) September 25-28, pp. 1809-1813.

[13] M. B. Young, A. Al-zou'bi, Y. Khamayseh and W. Mardini, "Improvement on LEACH Protocol of Wireless Sensor Network (VLEACH)", International Journal of Digital Content Technology and Its Applications, vol. 3, no. 2, (2009), pp. 132-136.

[14] K. Go. Vijayvargiya and V. Shrivastava, "An Amend Implementation on LEACH protocol based on Energy Hierarchy", International Journal of Current Engineering and Technology, vol. 2, no. 4, (2012) December, pp. 427-431.

[15] R. Halke and V. A. Kulkarni, "En-LEACH Routing Protocol for Wireless Sensor Network", International Journal of Engineering Research and Applications (IJERA), vol. 2, no. 4, (2012) August, pp. 2099-2102.

[16] W. Liu and L. Wang, "An improved algorithm based on LEACH protocol", Journal of Applied Mechanics and Materials, vol. 347, (2013), pp. 2725-2727.

[17] C. Zhang, F. Liu and N. Wu, "A Distributed Energy-efficient Unequal Clustering Routing Protocol for Wireless Sensor Networks", Int. Journal of Computational Information Systems, vol. 10, no. 6, (2014) March, pp. 2369-2376.

[18] Genetic Algorithm - Wikipedia, "https://en.wikipedia.org/wiki/Genetic_algorithm."

[19] Genetic Algorithm - MATLAB & Simulink, Mathworks, "https://in.mathworks.com/help/gads/genetic-algorithm.html."

[20] K-Means Algorithm - Wikipedia, "https://en.wikipedia.org/wiki/K-means_clustering."

[21] K-Means Algorithm - MATLAB - kmeans, Mathworks, "https://in.mathworks.com/help/stats/kmeans.html."