Journal of Communications Technology, Electronics and Computer Science, Issue 34, 2021
ISSN 2457-905X

A Hybrid Method for Energy Efficient Data Storage in the Internet Of Things

Shahram Jamali[1], Negar Taheri[2], Mohammad Esmaeili[3]

[1,2,3]Department of Computer Engineering
[1]University of Mohaghegh Ardabili, [2,3] Science and Research Branch, Islamic Azad University
[1,2,3]Ardabil, Iran
Author Correspondence: esmaeili.cse@gmail.com

Abstract— **In the Internet of Things (IoT), to increase the volume of stored data, distributed systems have replaced by centralized systems, and high volume data is divided into smaller sections and each section is stored in a data registration center. The design of the system should be such that even with a number of unavailable data centers, all information is still retrievable, therefore it is necessary to store multiple copies of the information in the system so that the initial information is not lost, despite the loss of part of the data registration center. In the distribution data storage, node energy balance and reduce the cost of access to data is a major problem between sensor nodes. Sensor node clustering is a viable solution to solve this problem. In this paper, a distributed data storage method using PSO and the K-means clustering mechanism organized by the binary decision tree C4.5 in the Internet of Things environment was proposed. As the simulation results show, the proposed method has been able to reach increasing availability, decreasing communication costs, and decreasing energy usage.**

Keywords— The Internet of Things; Distributed Storage; Energy Efficiency; PSO; K-Means; Decision Tree.

I. INTRODUCTION

The Internet of Things is a new concept that has emerged in recent decades [1] and is one of the thousands of the spread of the Internet and of course the development of wireless technologies. The most important feature of the new generation of the internet (IoT), connect all electrical objects to the Internet. Thus, various home appliances (such as watches, light bulbs, refrigerators, etc.) can be controlled, turned on, and off via the Internet and remotely. The collection of objects involved in the Internet of Things includes several related systems such as radio frequency identification (RFID) [2], Machine to Machine (M2M) [3], and wireless sensor networks (WSNs) [4] are. The emerging Internet of Things networking, wireless sensor networks play an important role in sensing and collecting data [5] in the surrounding environment. Because wireless sensor networks tend to stay for a long time without protection, distributed data storage [6] must be fault-tolerant and in fact, if one or more sensors fail, there wasn't any problem with the stored data. WSNs for IoT monitoring systems include free nodes for sensing the environment and usually a sink node for gathering data and a gateway to the internet. Connections among sensor nodes and sink nodes are not instant namely in isolated WSNs in which sink node is not always present [7]. As a result, distributed storage with data replication was proposed. Because wireless sensor networks tend to remain unprotected for long periods of time, distributed data storage must be such that it is error-prone, and In fact, if one or more sensors fail, there wasn't been any problem with the stored data. Thus, in order to present an efficient distributed storage with a data replication scheme, communicational optimization, and optimize energy usage are required. The structure used in the proposed approach is a distributed data storage scheme with the C4.5 decision tree mechanism as well as the combined algorithm K-means and PSO for data clustering ،which makes it accessible and improves energy consumption in data storage. The rest of this paper is organized as follows: Related works are reviewed in Section II. In Section III the proposed method is given. Section IV expresses simulation results of the proposed methods and their performance compared with the related works and finally, Section V concludes the paper.

II. RELATED WORKS

In recent years, various schemes to distribute and replicate data in WSNs have been proposed [8]. Data replication strategies have been proposed in the literature, mainly to overcome the problem of node failures. Authors in [9] propose ProFlex, a distributed data storage protocol for replicating data measurements from constrained nodes to more powerful nodes. The protocol benefits from the higher communication range of such nodes and uses the long link to improve data distribution and replication against the risk of node failures, but its disadvantages are data security. in [10] Supple data publishing protocol is provided. A possible flexible data dissemination protocol for wireless sensor network that considers sink to be static or mobile. The supple protocol has three stages: A tree structure, weight distribution, and data replication. This protocol has been introduced to overcome the disadvantages of RaWMS [11] and Deep [12] protocols. Unlike Proflex, Supple uses a single multiplication structure using tree topology. Its structure is created by a central sensor node in the assessment area in the form of a tree. The disadvantages are high energy consumption and traffic congestion in nodes close to the central

node. A data storage approach based on a Multi-level mapping Index is proposed in Park's paper [13] in a large-scale wireless sensor network. Considering the constraint of response time for the periodic queries. This method divides the whole network into three layers. Each layer uses local storage, data-center storage, and external storage respectively. This method will save huge energy compared with local storage, exterior storage, and data-centric storage. In [14] the proposed replication-based distributed data storage algorithm is greedy, in which the uncontrolled number of copies leads to the loss of information, in addition to the lack of access to data and consumes more energy. All of the above algorithms, despite some problems, are basically acceptable in terms of performance, but these methods are not suitable for storing data in a pervasive environment such as the Internet of Things, given the challenges mentioned. So in this research to find a method and optimal solution for these challenges are addressed. A distributed data storage mechanism in the Internet of Things environment with an energy efficiency approach was proposed. In general, according to studies, our approach has been exceeded. First, a mechanism of distributed storage with minimal repeat, that the decision tree clustering topology objects expanded in the Internet environment. First, a distributed storage mechanism with minimal repetition, which has been expanded by tree topology to cluster decision-making in the Internet of Things environment. As an efficient way to reduce the number of transmissions among objects, the balance of load distribution on the Internet of Things is to use efficiently the available energy resources and increase the lifespan of the network. As an efficient way to reduce the number of transmissions among objects, the balance of load distribution on the Internet of Things is to use efficiently the available energy resources and increase the lifespan of the network.

III. Proposed Method

Energy saving is an important issue in large-scale systems such as the Internet of Things. Collect data from the sensor nodes is a big challenge unless proper management of sensor data flow is approved. Data collection plays a vital role in the Internet of Things. In the proposed method, the data collection mechanism is a combination of tree structure and clustering, which are used as an efficient method to reduce the number of transmissions among objects and are an effective solution to this problem. A clustering algorithm combination based on K-means and particle optimization (PSO) has been proposed to achieve efficient energy management in the Internet of Things. The structure used in the proposed approach is a distributed data storage scheme with a combined clustering mechanism K-means and PSO organized by the binary decision tree C4.5 based on the weight of the clusters.

A. Optimal cluster selection by combining K-means and PSO

In the proposed model, the K-means algorithm [15] is used for clustering. But to choose the optimal cluster, a particle optimization algorithm [16] has also been used. In this method, when the particle pattern was defined and the initial parameters of the particle swarm optimization algorithm were determined, the K-means algorithm was used to generate the initial population to produce the required number of clustering solutions, which is chosen with the logic of the distance the headers and delivered to the PSO. The work of PSO tries to move the particles towards optimization by changing the position of the particles (optimizing the choice of headers) and then using the fit function, the fit of each method is determined using four parameters and optimized by PSO algorithm and Finally, the optimal solution is chosen.

B. Fitness function

In this research, for achieving the ultimate goal, 4 parameters have been defined which attempt to optimize in structure. The first parameter is the cluster's distance from the sink that should be less. To calculate and evaluate it, the usual geometric spacing given in (1) is used.

$$dis_{h(i)}^{BS} = \sqrt{(x_{h(i)} - x_{BS})^2 + (y_{h(i)} - y_{BS})^2} \quad (1)$$

In which i is an index of cluster head, h (i) means the cluster-head i, $^{x}{}_{h\,(i)}$ represents the longitudinal coordinates of the cluster head and $^{x}{}_{BS}$, the coordinates of the longitudinal sink or the central station and $^{y}{}_{h\,(i)}$, $^{y}{}_{BS}$ are respectively the cross-coordinates of the cluster head and sink and $dis^{BS}{}_{h\,(i)}$ represent the cluster heads distance from the sink. The second selected parameter for applying in the evaluation function is the energy used to transfer data between nodes, which is dependent on the distance which is obtained from the following equation.

$$\text{Energy}_{h(i)} = \sum_{j=1}^{m} \text{E}_{tr} * dis_{h(i)}^{s(j)} \qquad (2)$$

Where m is the number of cluster members, E_{tr} is the base energy level defined for the data unit transfer, $dis_{h(i)}^{S(j)}$ which shows the distance of the header from the members of the cluster, which is similar to (1) and $Energy_{\boldsymbol{h(i)}}$ represents the amount of energy consumed in the desired cluster. The next parameter in the proposed method is the amount of space occupied, in other words, the amount of space required to store the data of a cluster in the header. Therefore, reversing the amount of space required in Equation (3) has been used.

$$Storage_{h(i)} = \frac{m_{h(i)}}{M_{h(i)}} \qquad (3)$$

Where $M_{h(i)}$ is the amount of memory the header and $\boldsymbol{m_{h(i)}}$ number of cluster members and $Storage_{\boldsymbol{h(i)}}$ also represents the index of memory. The last parameter to determine the fitness is the distance between the members of a cluster of cluster heads which is obtained from Equation (4).

$$Dis_{h(i)} = \sum_{j=1}^{m} dis^{s(j)}{}_{h(i)} = \sum_{j=1}^{m} \sqrt{(X_{h(i)} - x_{s(j)})^2 + (y_{h(i)} - y_{s(j)})^2} \quad (4)$$

To determine the overall fit function of the weighted average, the four defined parameters were used after their normalization and the weight of all of them was considered to be 0.25. The function used to normalize the relationship is (5).

$$norm(f(t)) = \frac{(f(t) - \min(f))}{(\max(f) - \min(f))} \quad (5)$$

As a result, the general Fit function is defined as follows:

$$Fit_{h(i)} = w_{dis}.norm(dis^{BS}_{h(i)}) + w_{Energy}.norm(Energy_{h(i)}) + W_{Storage}.norm(Storage_{h(i)}) + W_{Dis}.norm(Dis_{h(i)}) \quad (6)$$

where in:

$$w_{dis} = w_{Energy} = w_{Storage} = w_{Dis} = 0.25 \quad (7)$$

The weight of each of the four parameters is. The goal of the proposed method is to realize the relation (8) and find the appropriate response to it.

Object = minimize (Fith(i)) (8)

C. The formation of cluster binary tree

The c4.5 decision tree [17], based on cluster proper routing architecture, can ensure the distance of data transfer between measurement nodes, scalability and balance of communication load, reduce energy consumption and increase the longevity of measurement nodes and ultimately data availability in the Internet of Things network. The condition for making a tree is that first the function of the average weight of each cluster is calculated then the best feature (weight) is found and finally divide and conquer is done recursively. The tree is built from top to bottom and the root node of the tree starts with more weight and selects the best feature based on greedy search and the rest of the levels are formed superficially. Each decision node has two children. To form a tree indicator as a CH capability, equation (9) to determine the level tree nodes defined and calculated. This indicator is used in order to select the roots of the leaves of the tree.

$$HDR\_ability_{h(i)} = W_s * \left(\frac{M_{h(i)}}{\max(M_h)}\right) + W_E\left(\frac{E_{h(i)}}{\max(E_h)} + \frac{E_{h(i)}}{dis_{h(i)} * \max(E_{h(\dots)})}\right) \quad (9)$$

Where in $\frac{M_{h(i)}}{\max(M_h)}$ Normalized amount ofmemory, $\frac{E_{h(i)}}{\max(E_h)}$ Normalized value of absolute energy, $\frac{E_{h(i)}}{dis_{h(i)} \times \max(E_{h(\dots)})}$ Normalized value of energy relative to distance and W_E, the weight is determined for the effect of energy whose value is determined in simulation.

D. Distributed data storage mechanism

Wireless receiver network sensors collect data and store it in headers. Storage and copying are done in the headers, which is related to the process of sending results and answering queries. In total, two phases exist: In the first phase, the sensors sense and send their data to the header in turn, and they are erased from their memory, and the headers copy the received information in three places, and when they can't send it to the header, then the doesn't copy and the data remains in the sensors themselves. Thus, the unique data (new sense) is in both the headers and the sensors. Figure (1) shows the proposed model.

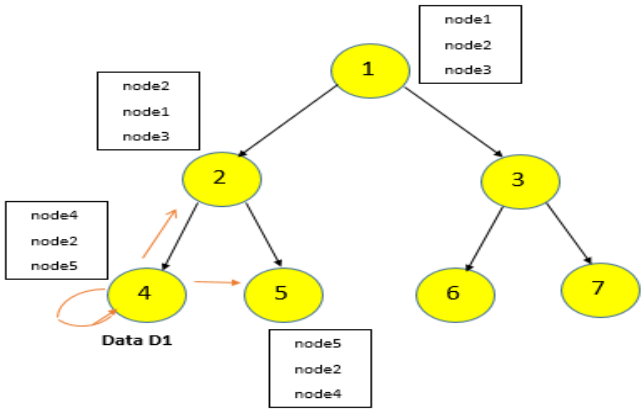

Fig.1: Replication in the case of R = 3 [18]

E. Troubleshooting energy hole in data recovery with fixed sink

Fixed sink node in distributed data storage to retrieve data, reduces network lifetime, as fixed sink has energy hole problem; Nodes near the sink typically consume more energy than other sensor nodes in the network due to the data relay and hence disconnection from the network may occur. Due to the problem of the energy hole, energy is considered on the structure so that the problem does not occur again. The data recovery period is 600 seconds, and when the time comes, all the data are sent to the sink and if there is a lack of energy, the network is rebuilt and clustering is re-formed otherwise, data are collected with the same structure.

IV. Evaluation of Simulation Results

The topology and model used in the simulation consist of a base station outside the sensing field, and sensing nodes internet of things fixed and uniform, with 120, 60, 180 nodes, are distributed on a large-scale IoT system monitoring environment. The dimensions of the IoT network are 600 x 600 m, which is a homogeneous network and the nodes are connected to each other as a form grid. The buffer size of the nodes is constant and equal to 100 bytes and the initial energy is limited for all nodes and the sense distance between all the sensors is [31-33 s]. The proposed method with three scenarios

of 60, 120, and 180 nodes is compared with the previous method [14] of 60 nodes.

A. Evaluation of memory used (Stored data)

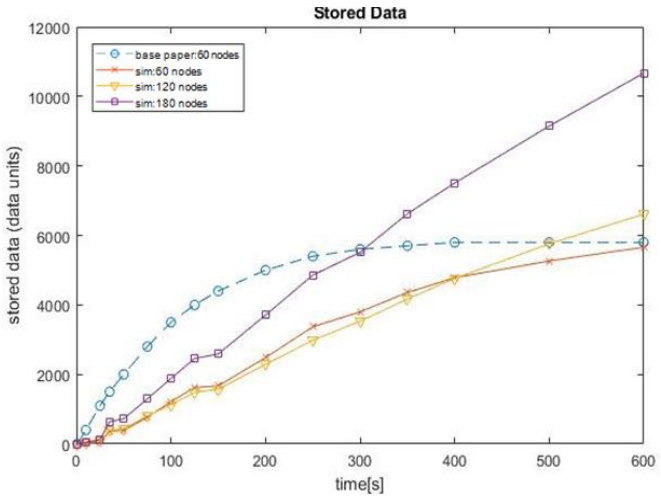

Fig.2: Stored data (memory consumption)

Due to the fact that in the previous method compared, the triple copy operation is performed by all nodes, the data exchange rate is high and consumes a lot of energy, and also initially fills the memory severely, but in the proposed approach, copy operations are performed on a limited number of cluster nodes after aggregation. Low memory consumption and uniform growth in the amount of memory consumption in the proposed method are quite evident. In the first scenario, the proposed method with 60 nodes, the amount of data stored in it is comparable to the previous method and its superiority is clear. The percentage recovery rate of the proposed method was 3.4% compared to the baseline method.

B. Evaluation of data loss due to energy limitation

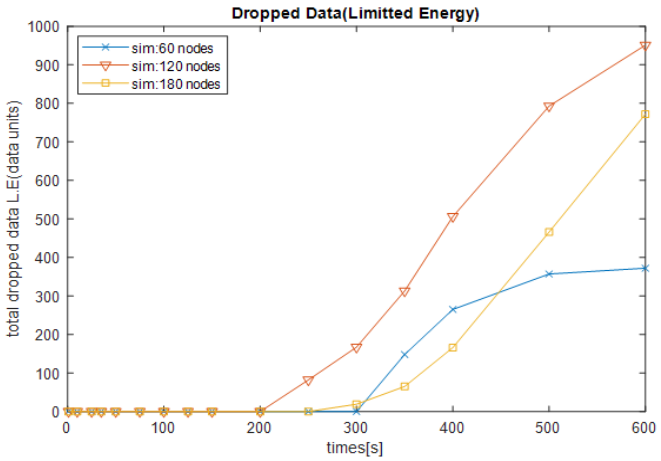

Fig.3: The amount of data deleted due to energy constraints

Energy consumption is very important in networks where nodes use a battery source and affect the life of the network. The previous method did not address energy consumption, but in theory, its energy runs out soon and the number of data drops is high because the number of transfers is higher. As shown in Figure (3), due to the increase in the volume of data collected compared to the increase in nodes and due to the fact that in all three scenarios, the initial energy levels of the nodes are considered constant and the same. The important point in this diagram is that as the number of nodes increases by 2 or 3 times, the amount of deleted data does not increase by 2 or 3 times. The reason for this is because of the low cost and low power consumption data transfer due to congestion and proximity of the nodes is.

C. Evaluation of unique stored data

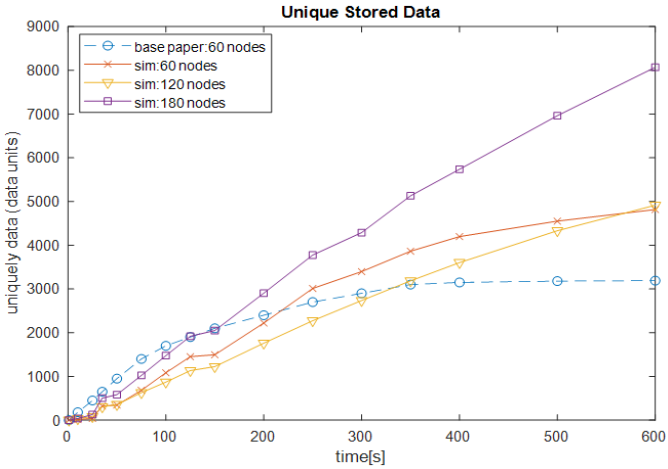

Fig.4: Unique stored data amount (regardless of copy)

Figure (4) shows that as long as the sensors do not die and are sensed, it is considered unique data. Comparing the previous method [14] with the proposed method in 60 nodes is that the previous method because it has used more of the entire network storage space, has caused a small amount of pure data to be stored in the network while the proposed method with better memory usage has caused a lot of data to be sensed and stored in the network. To keep network conditions stable, 60 knots are better than 120 knots and 180 nodes.

D. Evaluation of Network Lifetime and Energy Consumption

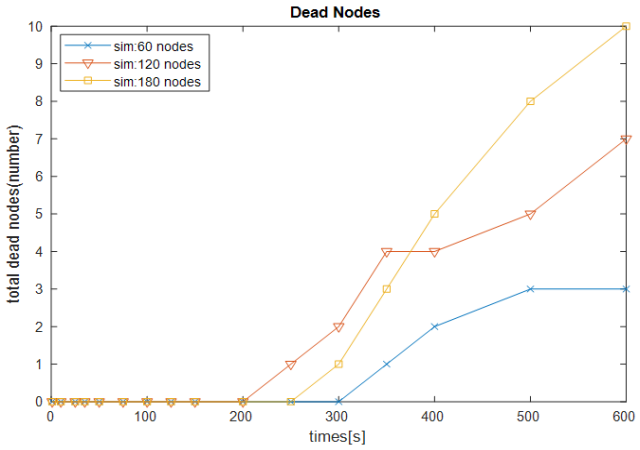

Fig.5: The number of nodes lost relative to time

Figure (5) shows the number of nodes lost due to energy depletion during simulation. As the number of nodes increases and keeping other conditions constant, the number of dead nodes increases. To reduce node death and increase network life, it is possible to increase the period of data collection and recovery period.

Table (1) shows the amount and percentage of energy consumed in all 3 scenarios of the proposed method. As it turns out, total energy consumption increases with increasing nodes, which is normal. As the number of nodes increases, energy consumption is high, but the third column shows, the percentage of energy consumed in the total energy of the network; From the point view of energy consumption, if the distances are close to each other, energy will be consumed less and this will be the percentage of energy consumption of the whole network. Also, the fourth column, which shows the ratio of energy consumption to the number of nodes (average per capita energy consumption), indicates that with the increasing number of nodes, per capita energy consumption decreases.

TABLE I. AMOUNT PERCENTAGE AND ENERGY CONSUMPTION IN VARIOUS SCENARIOS

The proposed scenario(nodes)	***Energy consumption (energy unit)***	***Percentage of consumption***	***The ratio of energy consumption to the number of nodes***
Sim: 60	47,2455	15.75	0.787425
Sim: 120	88.9543	14.83	0.7412858
Sim: 180	119.5969	13.29	0.6644272

V. CONCLUSION

The simulation of the proposed method has been simulated with MATLAB software. The goal is optimal distributed storage in the Internet of Things with low energy efficiency. In this paper, the results of the research evaluation showed that the particle swarm optimization algorithm, which is a meta-exploration algorithm, has a good ability to reduce energy consumption and optimal memory usage. The use of the K-means algorithm has increased its accuracy and capability and has led to the optimal selection of clusters and headers. The process of using the binary tree graph has caused to be made in energy saving. The results show the improvement in storage performance compared to the previous method is provided. The results showed that the optimization of the results was directly related to the ratio of the number of nodes, the dimensions of the network, and also the size of the memory of the nodes in addition, due to the high consumption of memory and energy in the knots, especially the knots close to the roots in the binary tree, increasing the memory and energy of nodes can increase the life and network capability.

REFERENCES

[1] Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. Computer Networks, 54(1), 52787-52805. Bradley, J., Barbier, J., & Handler, D. 2013.

[2] E. Vahedi, R. K. Ward, and I. F. Blake, “Performance analysis of RFID protocols: CDMA versus the standard EPC Gen-2,” IEEE Trans. Autom. Sci. Eng., 2014.

[3] M. Hasan, E. Hossain, and D. Niyato, “Random access for machine-to-machine communication in LTE-Advanced networks: Issues and approaches,” IEEE Commun. Mag., 2013.

[4] M. Esmaeili, S. Jamali, and H. S. Fard, "Energy-Aware Clustering in the Internet of Things by Using the Genetic Algorithm," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 12, no. 2, pp. 29-37, 2020.

[5] B. Sheng, Q. Li, W. Mao, Data storage placement in sensor networks, in: 7th ACM International symposium on Mobile Ad Hoc Networking and Computing, 2006.

[6] L. Yan, W. L. Tao, K. Liu and Y. Xu, “The Research of Data Storage and Retrieval Scheme for Wireless Sensor Networks”, International Journal of Intelligent Engineering and Systems, 2009.

[7] Corke, P., Wark, T., Jurdak, R., Hu, W., Valencia, P., & Moore, D.. Environmental wireless sensor networks. Proceedings of the IEEE, 98(11), 1903–1917. 2010.

[8] Neenu M. Nair, J. Sebastian Terence. Survey on Distributed Data Storage Schemes in Wireless Wireless Sensor Networks. Indian Journal of Computer Science and Engineering (IJCSE),2014.

[9] G. Maia, D.L. Guidoni, A.C. Viana, A.L. Aquino, R.A. Mini, A.A. Loureiro, “A distributed data storage protocol for heterogeneous wireless sensor networks with mobile sinks”, Ad Hoc Networks, 2013.

[10] C. Viana, T. Herault, T. Largillier, S. Peyronnet, F. Zaı¨di, Supple: a flexible probabilistic data dissemination protocol for wireless sensor networks, in: 13th ACM International Conference on Modeling, analysis, and simulation of wireless and mobile systems,2010.

[11] Z. Bar-Yossef, R. Friedman, G. Kliot, RaWMS – random walk based lightweight membership service for wireless ad hoc networks, ACM Transactions on Computer Systems 26 ,2008.

[12] M. Vecchio, A.C. Viana, A. Ziviani, R. Friedman, Deep: density-based proactive data dissemination protocol for wireless sensor networks with uncontrolled sink mobility, Elsevier Computer Communication 33 (8) ,2010.

[13] J. Park, D. Seong, H. Kim, et al., "A data-centric storage scheme for high storage utilization in wireless sensor networks," *Cluster Computing*, vol. 18, no. 1, pp. 247-257, 2015.

[14] Pietro Gonizzi ,Gianluigi Ferrari ,Vincent Gay b, “Data dissemination scheme for distributed storage for IoT observation systems at large scale” Information Fusion, 2013.

[15] Youguo Li, Haiyan Wu, ‘‘A Clustering Method Based on K-Means Algorithm”, 2012.

[16] Raghavendra V. Kulkarni, Senior Member, ‘‘Particle Swarm Optimization in Wireless Sensor Networks:” A Brief Survey, 2008.

[17] S. Ruggieri,Dipartimento di information, Universita di Pisa. ‘‘Efficient C4.5”,2010.

[18] Taheri, Negar, AND Jamali, Shahram. "Distributed Data Storage in the IoT: A Performance and Reliability Approach" Networking and Communication *Engineering*, 2020.