Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
Department of Computer Science
College of Science for Women

An Energy-Efficient Data Reduction by using Similarity and Data Compression Techniques in IoT

A graduation project is submitted to the Department of computer science in partial fulfillment of the requirements for the degree of Bachelor of Computer Science.

By

Naba'a Ali Turki

Supervisor

Lecturer: Ahmed Mohammed Hussein

Babylon, Iraq

2022-2023

يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ
وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ وَاللَّهُ
بِمَا تَعْمَلُونَ خَبِيرٌ

صدق الله العلي العظيم

سورة المجادلة(11)

الاهداء

إلهي لايطيب الليل إلا بشكرك ولايطيب النهار إلا بطاعتك ولا تطيب اللحظات إلا بذكرك .. ولا تطيب الأخرة إلا بعفوك ولا تطيب الجنة إلا برؤيتك

الله جل جلاله

إلى من بلغ الرسالة وأدى الأمانة .. ونصح الأمة .. إلى نبي الرحمة ونور العالمين

سيدنا محمد (صلى الله عليه واله وسلم)

إلى من كلله الله بالهيبة والوقار .. إلى من علمني العطاء بدون انتظار .. إلى من أحمل أسمه بكل افتخار .. كلماتك نجوم أهتدي بها اليوم وفي الغد وإلى الأبد...

والدي

إلى ملاكي في الحياة .. إلى معنى الحب وإلى معنى الحنان والتفاني .. إلى بسمة الحياة وسر الوجود إلى من كان دعائها سر نجاحي وحنانها بلسم جراحي إلى أغلى الحبايب

أمي الحبيبة

الى رفيقات المشوار اللاتي قاسمني لحظاته رعاهم الله ووفقهم ....

الي كل من سعى جاهدا كي اصل لهذا اليوم واختم مشوراري الدراسي

إلى منارة العلم والعلماء إلى الصرح الشامخ .... الى الذين حملوا أقدس رسالة في الحياة إلى الذين مهدوا لنا طريق العلم والمعرفة ..

أساتذتنا الأفاضل

Supervisor Certificate:

I certify that the preparation of this project entitled "Installment Sales System" was made under my supervision in the **Department of computer science, college of science for women, university of Babylon** as a partial fulfillment of requirements of the B.Sc. Degree in Computer Science.

Signature:

Name:

Title:

Date: / /

ABSTRACT

Nowadays, the number of Internet of Things (IoT) devices has rapidly increased due to their increasing use in different real-world applications. The sensor devices represent the basic element of the IoT network because they gather data from various environments and situations, while the sink node serves as the network's brain because it processes the data and makes decisions. However, the large amount of data that the sensor devices gather and send to the gateway toward the sink, on the one hand, causes the sensor's limited energy to be depleted and, on the other hand, makes it more difficult to achieve the decisions using these data at the sink. Therefore, before sending data to the gateway, it is important to get rid of any duplicate data while maintaining a high level of data quality. In this project, an Energy-Efficient Data Reduction by using Similarity and Data Compression Techniques in IoT (EERSDC) for saving power in IoT networks is suggested. The (EERSDC) makes periodic decisions on sending the data to the gateway. It uses the Jaccard similarity method in each period to find the similarity between the current period with the next period's data and decide whether the next data should be sent to the gateway or not. When the decision is made to send the data to the gateway, an effective compression approach is used by (EERSDC) to get rid of the duplicate data. It uses Run Length Encoding (RLE) to compress the data before sending it to Gateway. Simulation results based on real-world data show that the energy consumption of the (EERSDC) method is better than the energy consumption of data before transmission, and the (EERSDC) reduced the redundant data ratio and transferred data size.

Table of Contents

List of Figures

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The fundamental aspect of the Internet of Things (IoT) is to allow the communication of virtual and physical things with each other [1]. IoT systems include embedded intelligence, wireless sensor networks, and cloud computing. Sensors, cameras, radio frequency identifiers (RFID), and other devices are used by IoT systems to gather environmental data [2]. These systems are capable of providing sophisticated services such as remote management, online analytics, and real-time remote monitoring. The IoT is used in a range of remote monitoring applications, including healthcare, smart manufacturing, smart homes, smart cities, and smart agriculture, with the objective of improving productivity and decreasing costs [3, 4]. In the future, periodic sensor networks (PSNs) will be one of the most critical parts of the Internet of Things (IoT), and they will play a key part in people's lives due to their extensive use in a variety of applications. These types of networks have received a lot of attention from researchers in the past 4 years. The characteristics of these PSNs differentiate them from other ad hoc wireless networks. Furthermore, several limitations due to these characteristics are imposed and lead to many challenges in the PSNs. Some of these challenges are routing, data aggregation, topology control, security, and coverage. One of the most fundamental research challenges in PSNs is to gather and consolidate huge amounts of data in an energy-efficient manner regularly, then transport them to the sink to extend the network lifetime. Because sensor batteries have a limited lifetime, energy-efficient data collection and data reduction methods for frequent data gathering are required for energy optimization.

Transmitting/receiving the data by sensor devices is an expensive process, while in-network computations are much less expensive from the energy consumption point of view and are sometimes ignored as insignificant [5, 6].

However, as shown in Figure 1 [6] the computation requires much less energy than the data transmission/receiving. For instance, the energy required to transmit a 1 KB data message over a distance of 100 m is nearly similar to the execution of nearly three million instructions on a normal microprocessor. As a result, any extra processing that lowers the data size even by one data bit would save energy. This dramatic difference between transmitting/receiving and computation highlights the importance of in-network data processing in reducing consumed energy in the network [5].

1.2 Contributions of Project

This project introduces the following contributions:

- **i.** An Energy-Efficient Data Reduction by using Similarity and Data Compression Techniques in IoT (EERSDC) is proposed.
- **ii.** The proposed approach executes the Similarity measurement to find the similarity between the current period and the next period and then make a decision about sending or not the currently collected data to the Gateway.
- **iii.** An efficient compression approach is proposed and implemented by the proposed approach to compress the data before sending them to the Gateway. This approach is based on the RLE compression approach to minimize the size of the data that have been sent to the Gateway and to preserve the power of sensor batteries thus extending the network's life span.
- **iv.** A custom simulator based on the Python programming language and based on real observed data from sensor nodes put at the Intel Berkeley Research Lab [11] is used to run several simulation experiments.

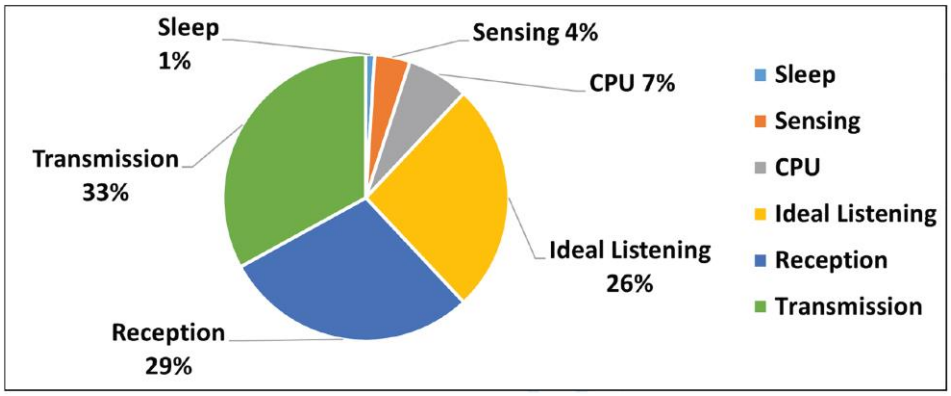

Figure 1.2: Energy Consumption inside the Sensor Device

1.3 Aim of Project

A project aims to deploy a network of small, low-cost sensors that can gather data from their environment and transmit it wirelessly to a central hub or database for analysis. These sensors can be used for a wide range of applications, including environmental monitoring, industrial process control, smart agriculture, and home automation. Other important considerations for a wireless sensor network project include ensuring data security and privacy, managing the network to optimize performance and reliability, and providing a user-friendly interface for accessing and visualizing the data collected by the sensors. Overall, the goal of a wireless sensor network project is to leverage the power of IoT technologies to enable more efficient and effective data collection and analysis in a wide range of fields.

1.4 Related Works

The authors in [7, 8] proposed a modified k-nearest neighbor algorithm for data redundancy removal in the sensor node to save energy and extend the lifespan of the network. Then, they extend their work to include data redundancy elimination on the second level of the network (Gateway). The received data vectors of sensor nodes are gathered into groups of similar data vectors and then one representative vector is sent for each group. In [9], the authors introduce a new data reduction method for saving energy in the IoT network. They implemented two algorithms to remove the redundant data at both the sensor and aggregator levels. The divide and conquer method is implemented at the aggregator level while the clustering is used with the sensor nodes. The proposed work in [10] presents an energy-aware data transmission reduction in fog computing-based IoT networks. The method is activated at both sensor nodes and the fog Gateway. In the sensor devices, they implement combined grouping and easy encoding methods to remove unnecessary data before forwarding them to the fog Gateway. In the fog Gateway, they proposed a clustering algorithm based on the "Dynamic Time Warping" (DTW) that combines with the simple encoding method to further remove the redundant data before transmitting them towards the cloud data center.

The comprehensive resume of related literature mentioned above indicates many methods for data reduction using different techniques. Nevertheless, the proposed methods in existence cannot adequately eliminate redundant data while keeping a high level of data accuracy. Furthermore, some of these methods are more complicated and require high time and memory complexities, and they cannot be implemented inside the constrained resources sensor devices.

CHAPTER TWO

THEORETICAL BACKGROUND

2.1 Introduction

This project focuses on the data reduction problem, in which saving energy is an essential condition. To address this problem, this project suggests one level data reduction approach for reducing the data redundancy and reducing the power consumed while keeping good data accuracy in the gateway. In this section, some theoretical background about some employed techniques will be presented.

2.2 Similarity Measurement

Similarity measurement is a technique used to determine how similar or dissimilar two entities are. It is a process of comparing two objects based on their features or attributes. There are various types of similarity measurement techniques available. In this project, we introduce the Jaccard Similarity method as in present in the next section.

2.2.1 Jaccard Similarity

The Jaccard similarity method is an algorithm used to determine the similarity between two sets of data or documents. It is commonly used in computer science, data analysis, and information retrieval. The Jaccard similarity method calculates the relative similarity between two sets by measuring the ratio of the common elements to the total number of distinct elements in both sets. The formula for Jaccard similarity is [13]:

$$J(A,B) = |A \cap B| / |A \cup B|$$

Where $|A \cap B|$ is the number of elements common to sets A and B, and $|A \cup B|$ is the total number of distinct elements in both sets.

The Jaccard similarity index ranges from 0 to 1, where 0 indicates no common elements between the sets and 1 indicates that the sets have identical elements. The Jaccard similarity method is used in various fields, such as machine learning, text analysis, information extraction, and automatic classification. It is used to determine the similarity between datasets, documents, or text.

2.3 Run-Length Encoding (RLE)

Run-length encoding (RLE) is a lossless compression method where sequences that display redundant data are stored as a single data value representing the repeated block and how many times it appears in the image. Later, during decompression, the image can be reconstructed exactly from this information. This type of compression works best with simple images and animations that have a lot of redundant pixels. It's useful for black-and-white images in particular. For complex images and animations, if there aren't many redundant sections, RLE can make the file size bigger rather than smaller. Thus it's important to understand the content and whether this algorithm will help or hinder.

This technique was first patented by Hitachi in 1983. It's less popular today because there are other more advanced options available, but you will still find it in use for color for fax machines, icons, line drawings, and simple animations. You'll also find it in TIFF and PDF files. However, this compression was regularly used when transmitting analog television signals back in 1967.

The RLE method is a simple form of data compression used to reduce the size of a file or data set. It works by analyzing a data stream and replacing sequences of repeated values with a count and a single instance of the value. The basic idea behind RLE is to replace runs of identical values with a single

value and a count of how many times that value appears. For example, the sequence "AAAAABBBBCCCC" could be encoded as "5A4B4C". The advantage of this method is that it can reduce the amount of data that needs to be stored or transmitted without losing any information.

RLE is often used in image and video compression, where it can be particularly effective for encoding simple or repetitive images. It is also used in some audio and text compression algorithms. While RLE is a relatively simple and effective compression method, it has some limitations. It works best on data with long runs of identical values and may not be as effective on more complex data sets. In addition, decoding an RLE-encoded file requires a bit of processing overhead, which can slow down the decompression process [10].

2.4 Energy Consumption

Energy consumption can be represented as time series data, where the energy consumption is measured at different time intervals. Time series analysis of energy consumption data can provide insights on patterns, trends, and seasonality of energy usage over time. This information can be useful in several ways, such as predicting future energy demand, identifying energy usage patterns of specific sectors or regions, and designing efficient energy management strategies. To perform a time series analysis of energy consumption data, we need to follow the following steps:

1. ***Data collection:*** Collect the energy consumption data for a significant period of time, preferably over multiple years.
2. ***Data pre-processing:*** Clean the data by removing any missing or inconsistent values and correcting any anomalies.

3. ***Data exploration:*** Explore the data by plotting it over time to identify any trends, seasonal patterns or outliers.
4. ***Model selection:*** Select appropriate time series models based on the identified patterns in the data. For instance, if there are seasonality and trends, then models like ARIMA or Seasonal ARIMA can be used.
5. ***Model evaluation:*** Evaluate the model's performance by comparing the predicted values against the actual values. Use evaluation metrics like Mean Squared Error (MSE) or Root Mean Square Error (RMSE) to determine the accuracy of the model.
6. ***Forecasting:*** Use the selected model to forecast future energy consumption.

By performing these steps, we can effectively analyze energy consumption data as a time series and derive useful insights to aid in making informed decisions.

2.5 Programming Language of Python

Python is a high-level interpreted programming language. It was created by Guido van Rossum and first released in 1991. Python's design philosophy emphasizes code readability and simplicity, making it an excellent language for beginners and experienced programmers. Python has a large standard library and a vast number of external libraries, making it versatile for many applications. It is commonly used for web development, scientific computing, data analysis, artificial intelligence, machine learning, and automation, among others [12].

Python is easy to learn and has a straightforward syntax. It uses indentation to indicate blocks of code, making it easy to read and maintain. Python supports multiple programming paradigms, including object-oriented, procedural, and functional programming. Python's popularity is due to its simplicity, ease of use, and versatility. Its popularity has earned it a wide range of use cases and many job opportunities for programmers who are proficient in it.

CHAPTER THREE

PROPOSED APPROACH

3.1 Data Gathering from IoT Sensors

The gathering task of data is periodic, in which new data from sensor i are taken at every time s. Then, by node i, in each period a new data vector was generated as $R = [r_1, r_2,\ldots, r_T]$, where T is the total number of data that were taken throughout each time. When s is really short or there has not been any modification to the surveillance area for a while, the node will record the same (or substantially similar) data. As a result, within each period, the node will be able to send enormous amounts of data to the gateway.

3.2 Proposed EERSDC Approach

To decrease data transmission in IoT networks, An Energy-Efficient Data Reduction by using Similarity and Data Compression Techniques in IoT (EERSDC) approach based on similarity and compression is presented. Each sensor employs the EERSDC method. It is seen as a sensible strategy to save energy and transport fewer data packets, maintaining the accuracy of the data received at the gateway while prolonging the network's lifespan. The flowchart for the proposed EERSDC technique is shown in Figure 3.1. The proposed EERSDC approach is distributed at each sensor device and works in periods. In each period, the sensor device collects a fixed number of readings from the surrounding environment in the area of interest. The collected readings of the first period will be transmitted. Every time the sensor device decides to send its collected data, it reduces them by removing the redundant ones, compressing them, and then sending them to the gateway. The process of reducing the data and compressing it starts by applying the similarity technique Jaccard method, which finds the similarity between the current period with the next period, if the similarity exists then the data will not send to the gateway, and if else then will compress the data before sending to the gateway.

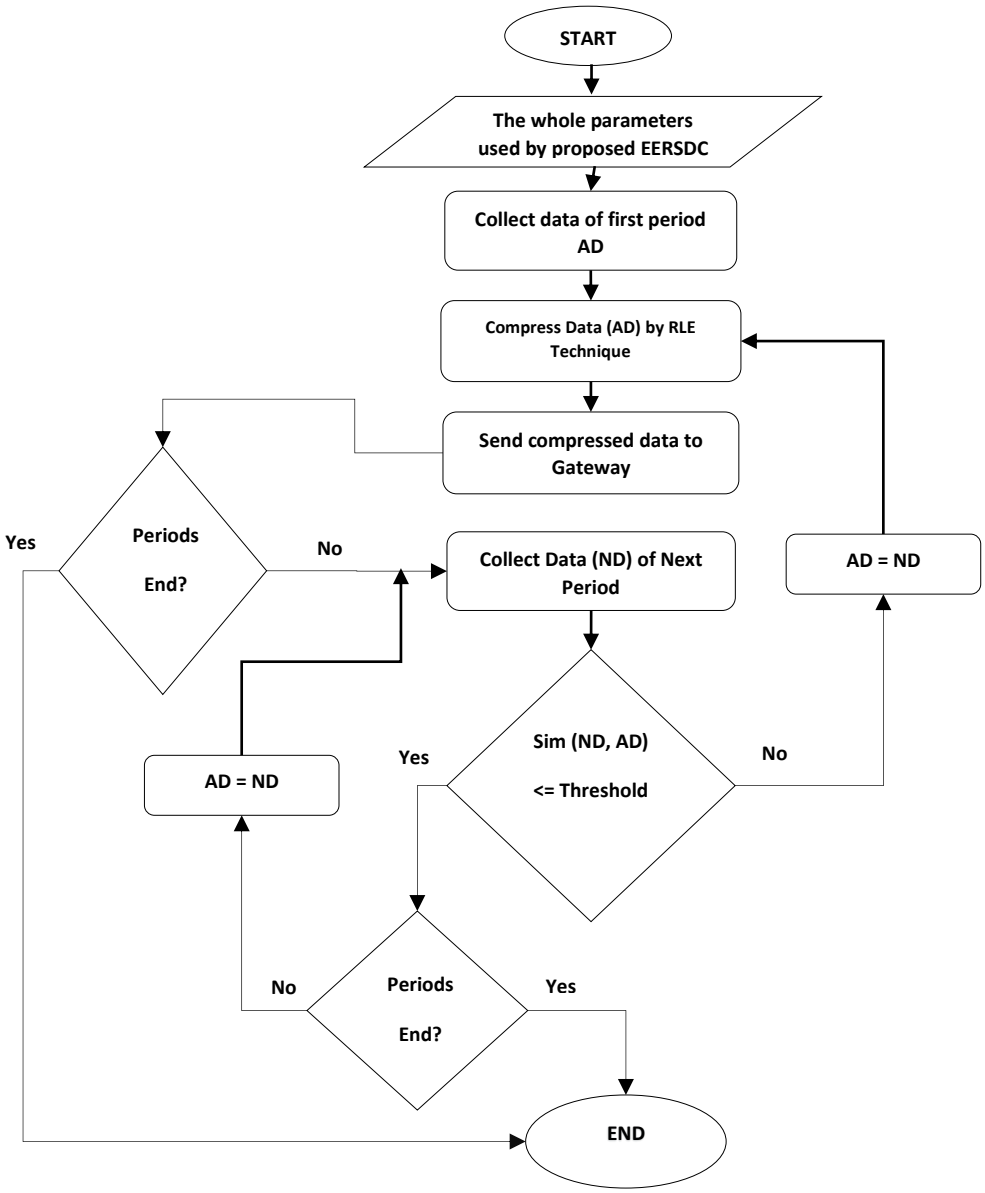

Figure 3.1: Flowchart of EERSDC Proposed

3.3 Results and Discussion

This section achieved several experiments to evaluate the effectiveness of the EERSDC technique. These simulation results are conducted using a custom simulator built in Python. Actual data from sensor nodes placed at the Intel Berkeley Research Lab are used in these simulation studies [11]. The Berkeley database comprises information on voltage, humidity, temperature, and light that were collected every 31 seconds from 54 sensors, as shown in Figure 3.2. In this project, we used one sensor, and 10,000 sensed data are taken from the sensor, with the emphasis on only one kind of sensor measurement: humidity, for the sake of simplicity (Sensor1).

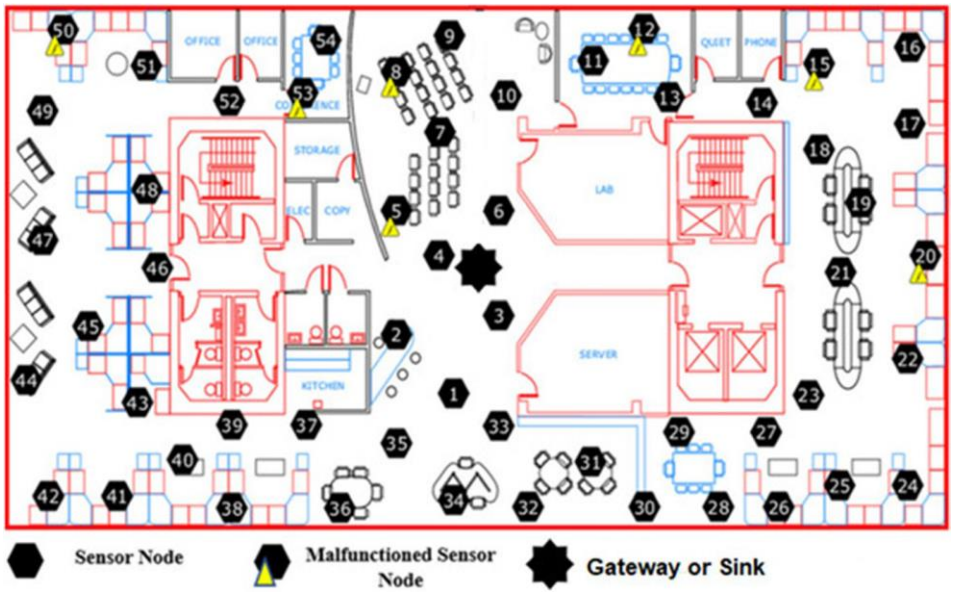

Figure 3.2: Intel Berkeley Research Lab

Heinzelman's radio model is used as an energy consumption model during the simulation [9]. It is shown in Figure 3.3.

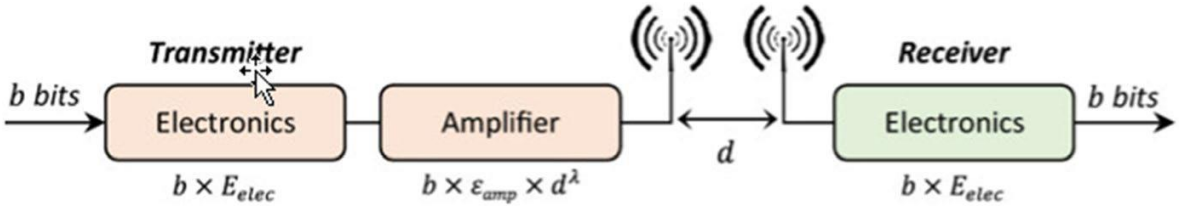

Figure 3.3: Heinzelman's radio model

3.3.1 Data Reduction

The data reduction was calculated by calculating the amount of data that was reduced after the data compression process using the RLE method, where the obtained results were compared with the original amount of data. Figure 3.3 illustrate the data reduction.

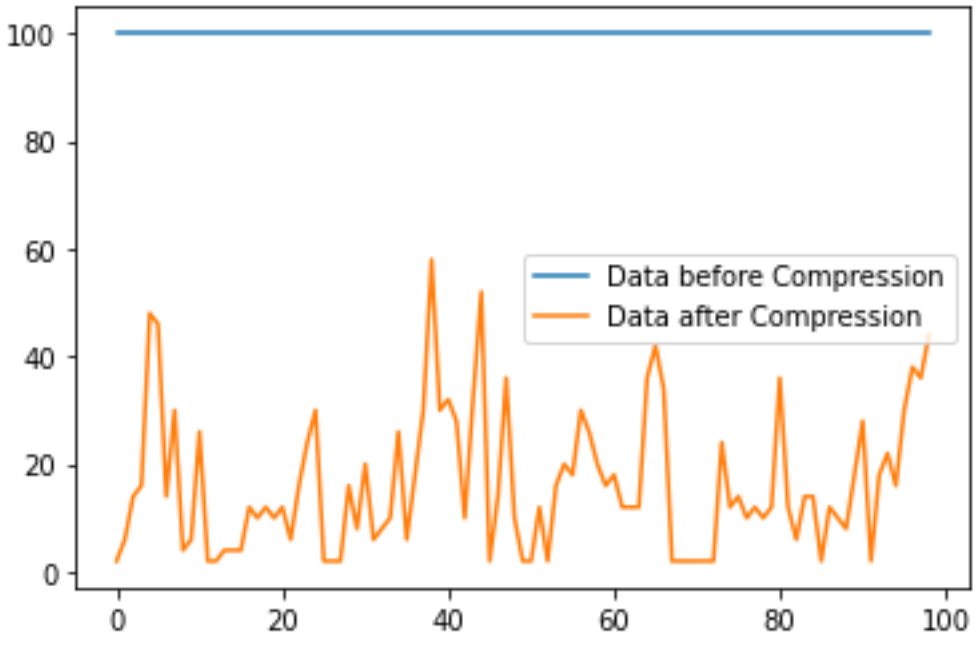

Figure 3.4: Data Reduction

3.3.2 Energy Consumption

Power is a crucial component of sensor devices due to the constrained resources of the sensor nodes. Figure (3.4 and 3.5) represents the amount of energy used by the sensor device for all periods.

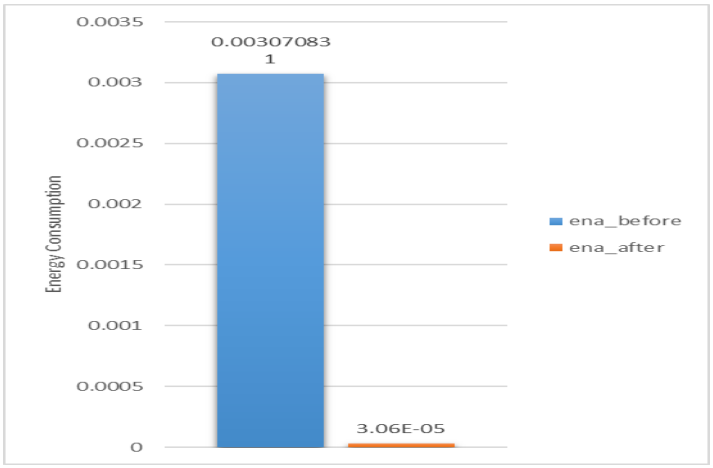

Figure 3.5: Energy Consumption 1

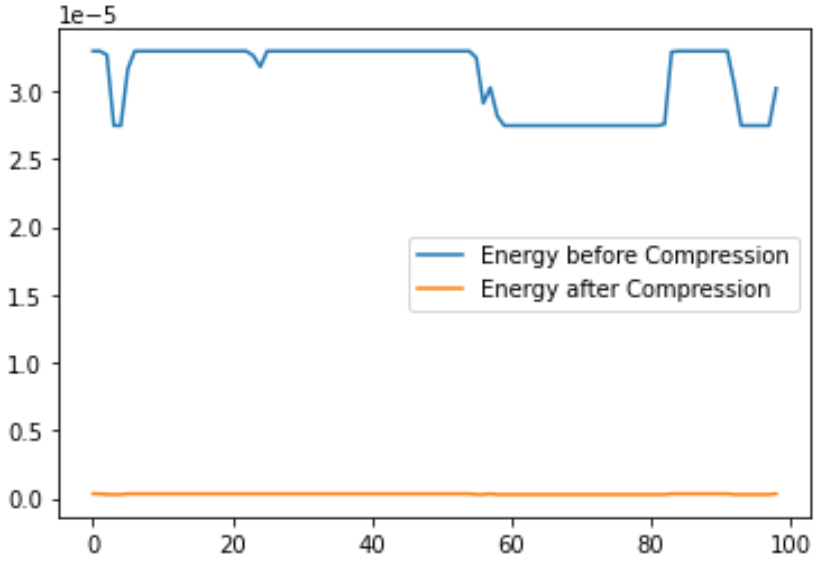

Figure 3.6: Energy Consumption 2

As we can see in the Figures the energy consumption reduced from 0.0031 up to 0.000031 compared with the normal situation.

3.4 Conclusion and Future Work

To lower data communication in the IoT, an Energy-Efficient Data Reduction by using Similarity and Data Compression Techniques in IoT (EERSDC) for saving energy in IoT networks is suggested in this project. The EERSDC works in periods. At each period, we used the similarity function to remove the redundant data. Based on the conducted results, the EERSDC has better performance than other similar methods in terms of reducing data, sending fewer data, using less energy, and getting more accurate data. The EERSDC reduced the depleted power by the sensor device (in joules) from 0.0031 to 0.000031. Future work will focus on applying machine and deep learning approaches to minimize the size of data carried from the gateway to the cloud. Finally, we want to think about the noise data that are created during the communication process. We plan to add a white noise removal algorithm that is based on EERSDC and to test how efficient the DiPCoM approach is with or without this algorithm.

References

1. Adu-Manu KS, Tapparello C, Heinzelman W, Katsriku FA, Abdulai JD (2017) Water quality monitoring using wireless sensor networks: current trends and future research directions. ACM Trans Sens Netw (TOSN) 13(1):1–41.

2. Akyildiz IF, Vuran MC (2010) Wireless sensor networks. John Wiley & Sons.

3. Alhmiedat T et al (2015) A survey on environmental monitoring systems using wireless sensor networks. J. Netw 10(11):606–615.

4. Almeida FR Jr, Brayner A, Rodrigues JJ, Maia JEB (2017) Improving multidimensional wireless sensor network lifetime using pearson correlation and fractal clustering. Sensors 17(6):1317.

5. Bagaa M, Challal Y, Ksentini A, Derhab A, Badache N (2014) Data aggregation scheduling algorithms in wireless sensor networks: solutions and challenges. IEEE Commun Surv Tutor 16(3):1339–1368.

6. Bahi JM, Makhoul A, Medlej M (2014) A two tiers data aggregation scheme for periodic sensor networks. Adhoc Sens Wireless Netw 21(1):77.

7. Harb H, Makhoul A, Abou Jaoude C (2018) En-route data filtering technique for maximizing wireless sensor network lifetime. In: 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 298–303. IEEE

8. Harb H, Makhoul A, Couturier R, Medlej M (2015) Atp: An aggregation and transmission protocol for conserving energy in periodic sensor networks. In: 2015 IEEE 24th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 134–139. IEEE

9. Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, pp. 10–pp. IEEE.

10. Idrees AK, Abou Jaoude C, Al-Qurabat AKM (2020) Data reduction and cleaning approach for energy-saving in wireless sensors networks of iot. In: 2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1–6. IEEE.

11. Madden S (2004) Intel berkeley research lab. http://db.csail.mit.edu/labdata/labdata.html

12. https://creativecomputing.ca/02/2_1_The_Python_Language.html .

13. https://www.learndatasci.com/glossary/jaccard-similarity/ .