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

An improved method for energy conservation in Wireless Sensor Network

A Thesis

Submitted to the Council of College of Science for Women at University of Babylon in a Partial Fulfilment of the Requirements for the Degree of Master in Science\ Computer Science

By

Wisal Basim Nedham Al-Marzoog

Higher Diploma in Computer Science/College of Science for Women/University of Babylon 2019

Supervised By

Asst. prof. Dr. Ali Kadhum M. Al-Quraby

2023 A.C. 1444 A.H.

بِسْمِ اللَّـهِ الرَّحْمَـٰنِ الرَّحِيمِ

قَالَ ٱجْعَلْنِى عَلَىٰ خَزَآئِنِ ٱلْأَرْضِ ۖ إِنِّى حَفِيظٌ عَلِيمٌ

صدق الله العلي العظيم

سورة يوسف
آية (55)

CERTIFICATION OF THE EXAMINATION COMMITTEE

We are the members of the examination committee, certify that we have read this thesis entitled (**An improved method for energy conservation in Wireless Sensor Network)** and after examining the master student **(Wisal Basim Nedham Al-Marzoog)** in its contents in / /2023 and that in our opinion it is adequate as a thesis for the degree of Master in Science \ Computer Science with degree (**Excellent**).

Committee Chairman
Signature:
Name: **Saif Mahmood Khalaf**
Scientific Order: **Assist Prof. Dr.**
Address: University of Babylon\College of Science for Women.
Date: / /2023

Committee Member
Signature:
Name: **Furkan Hassan Saleh**
Scientific Order: **Assist Prof. Dr.**
Address: University of Kufa\College of Computer Science and Mathematics
Date: / /2023

Committee Member
Signature:
Name: **Mahdi Abed Salman**
Scientific Order: : **Assist Prof. Dr.**
Address: University of Babylon\College of Science for Women.
Date: / /2023

Committee Member (Supervisor)
Signature:
Name: **Ali Kadhum M. Al-Quraby**
Scientific Order **Assist Prof. Dr.**
Address: University of Babylon\College of Science for Women.
Date: / /2023

Date of Examination: / /2023
Deanship Authentication of college of Science for Women
Approved for the College Committee of graduate studies.
Signature:
Name: **Abeer Fauzi Al-Rubaye**
Scientific Order: **Prof. Dr.**
Address: **Dean of College of Science for Women**
Date: / /2023

Supervisors Certification

We certify that this thesis entitled **"*An improved method for energy conservation in Wireless Sensor Network*"** is done by **(*Wisal Basim Nedham Al-Marzoog*)** under our supervision.

Signature*:*
Name*: Asst. Prof. Dr. Ali Kadhum M. Al-Quraby*
Date*: / / 2023*
Address*: University of Babylon/ College of Science for Women*

Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled ***"An improved method for energy conservation in Wireless Sensor Network"*** for debate by the examining committee.

Signature*:*
Name*: Asst. Prof. Dr. Saif Mahmood Al-Alak*
Date*: / / 2023*
Address*: University of Babylon/College of Science for Women*

Acknowledgement

First of all, praise is to ***Allah***, the lord of the whole creation, for all the blessings, and for the help in achieving this research to its end.

I wish to express my thanks to my college (college of science for women), my supervisor, ***Asst.Prof.Dr. Ali Kadhum M. Al-Quraby*** for the generosity, patience and continuous guidance throughout the work. It has been my good fortune to have advice and guidance from him. My thanks to the academic and administrative staff at the department of computer science.

Wisal

Dedication

To ...

My father may God have mercy on him.

My dear Mother for her unlimited love, support, endurance and encouragement.

My supporters during this journey, ***my loyal friends***.

Wisal

ABSTRACT

In Wireless Sensor Network (WSN), the data collected and transmitted continuously over time. This presents an energy issue for such restricted battery powered devices (sensor nodes). In this thesis, exploit both spatial and temporal relations in data to address the energy consumption in WSN devices.

We proposed an energy-efficient protocol called a Data Reduction Based on a Dual Prediction Scheme (DRDPS). DRDPS suggests controlling the amount of data transmitted and avoiding the transmission of redundant data thus avoiding excessive energy consumption. DRDPS is implemented in three stages; network clustering stage, Dual Data Prediction (DDP) stage, and statistical-based data model in the cluster head stage. Three algorithms are employed to minimize energy consumption and prolong WSN lifetime, these algorithms are the K-means algorithm, Autoregressive Integrated Moving Average (ARIMA) algorithm, and Granger Causality (GC) algorithm.

K-means algorithm is a clustering algorithm employed to divide the sensors network into clusters, the optimal number of clusters is decided by using of Elbow method. ARIMA is a prediction method used dually to eliminate the redundant data transmissions between the sensor node and the cluster head. While the GC method is a statistical method used at the cluster head level to reduce the amount of data sets that are sent from the cluster head to the base station.

DRDPS performance was evaluated by computing network lifetime, network energy consumption, amount of data transmitted, and data prediction accuracy. Also, the DRDPS results were compared with P-DPA and ELR methods. The experiment results demonstrated that the suggested approach has significant improvement in terms of energy conservation. It has the ability to enhance the lifetime and keep acceptable data accuracy. DRDPS reduces the amount of energy expended by 75% when error prediction threshold is 0.5.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF PUBLICATIONS

ACCEPTED AND PUBLISHED ARTICLES	
No.	**Article Title**
1	An Improved Energy Efficient Clustering Protocol for Wireless Sensor Networks. 1st Al Hikma International Conference on Natural and Applied Sciences (IIICNAS2022).
2	A review of current prediction techniques for extending the lifetime of wireless sensor networks. International Journal of Computer Applications in Technology.
3	A comprehensive review of clustering approaches for energy efficiency in wireless sensor networks International Journal of Computer Applications in Technology.

LIST OF ABBREVIATIONS

Symbol	Description
ACF	Auto-Correlation Function
ADC	Analog-to-Digital Converter
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
BS	Base Station
CDS	Connected Dominant Set
CH	Cluster Head
DDP	Dual Data Prediction
DETS	Double Exponential Smoothing
DHSCA	Dual-Head Static Clustering Algorithm
DPS	Dual Prediction Scheme
DPSCAN	Density-Based Spatial Clustering of Applications with Noise
DRDPS	Data Reduction-Based Dual Prediction Scheme
ELR	Extended Version of Linear Regression
ECH	Enhanced Clustering Hierarchy
ETS	Exponential Smoothing
ETS separation	Error-Trend-Seasonality Separation
FND	First Node Died
GC	Granger Causality
GN	Gateway Node
GPS	Global Positioning System
GRNN	General Regression Neural Network
GSO	Glowworm Swarm Optimization
HND	Half Node Died
IS-K-means	Improved Soft-K-means
ITS	Intelligent Transportation Systems
KF	Kalman Filtering
LEACH	Low Energy Adaptive Clustering Hierarchy

LMS	Least Mean Square
LND	Last Node Died
LSE	Least Squares Estimation
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percent Error
MLE	Maximum Likelihood Estimation
MN	Member Node
MNMF	Multi-Node Multi-Feature Model
MSE	Mean Squer Error
NARX	Nonlinear Autoregressive Exogenous Model
NLT	Network Life Time
OK-means	Optimal K-means
O-LEACH	Orphan LEACH
PACF	Partial Auto-Correlation Function
PDCP	Pearson Data Correlation And Prediction
PDPA	Periodic Data Prediction Algorithm
PWSN	Periodic Wireless Sensor Network
QOS	Quality of Service
RMSE	Root Mean Square Error
REDM	Radio Energy Dissipation Model
SARIMA	Seasonal Autoregressive Integrated Moving Average
RSS	Residual Sum of Squares
SN	Sensor Node
SPS	Single Prediction Scheme
SSE	Sum of Squared Errors
TEG	Thermoelectric Generators
VAR	Vector Autoregression
WBAN	Wireless Body Area Network Technology
WSN	Wireless Sensor Network
WSS	Within-Clusters Sum of Squares

Chapter One

Introduction

CHAPTER ONE

INTRODUCTION

1.1 General Introduction

The wireless sensor network (WSN) is ordinarily made of a large number of homogenous/ heterogeneous microelectronic devices called sensor nodes (SNs). The SNs are low-cost devices made of low-energy batteries, limited storage, and restricted processing capabilities. The SNs distributed spatially to cover a widespread area and cooperatively track the physical or environmental phenomena of that region [1], [2]. The SNs operations comprise data sensing, make simple processing and short-range wireless communications. The collected data is sent to one or more collection points called sink node. From a technical perception, SN is a device that translates parameters or events in the physical world into measurable and analyzed signals [3].

The basic design of the WSN is always encountered with many limitations. The WSN lifetime is restricted due to obligate tied resources and accessibility of the actual sensors. The energy in WSN is the key issue that should be examined and analyzed because it is related to the lifetime survival of the network. Energizing or replacing these batteries of energy sources is very difficult. This implies managing the available energy efficiently [4].

In WSN, the energy is depleted in many ways (e.g. collecting and processing data, transmitting data). The data transmission is the most costly power exhausting, where the primary energy waster in the SN is the radio module.

The energy consumed for data processing is considered to be much fewer. Actually, the energy required to execute 3000 instructions may be equivalent to that required to transmit 1 bit of data over 100 meters distance [5]. Therefore, the critical issue of WSN related to extending its lifetime is well fulfilled the application demands.

To reduce energy depletion in WSN, many strategies are suggested to enhance the network performance and prolong the network lifetime [6], [7]. These strategies are to adjust the transmission power, improve energy-efficient routing protocols, decrease the amount of data transmissions (e.g. data prediction, data compression, data aggregation), and place unnecessary sensors in sleep mode (e.g. scheduling) [1].

At the topology level, WSNs can be constructed either flatly or hierarchically. Flat architecture is the traditional form of WSN. The flat network does not suppose any sub-networks, rather whole nodes are dealt as a single network. The state and capability of each node in classical flat WSN are equivalent, and each node serves as a data collector and gateway for its neighbor nodes. WSN built on this architecture has multiple issues, including sink dropout, energy-hole trouble, …etc. [8].

In hierarchical architecture, the network can take many forms like a chain, cluster, tree, or grid-based network. The most relevant method is clustering [9], [10]. The clustering technique is one of the approaches that divide the network into groups for the network maintenance. SNs work together to collect data from their environment. Due to the correlation of data collected from contiguous nodes, the cluster head (CH) aggregates data from every node in the cluster to lessen the amount of data that needs to be transmitted to the base station(BS). The fundamental idea of clustering is to localize message transmission inside clusters as well as between CHs and BS. This has various benefits, including maintaining bandwidth, avoiding redundancy, minimizing transmission overhead, and minimizing data travelling distance.

Based on application requirements in WSN, the collection of data can be classified into event-driven (for example oil and gas leaks, forest fires detection) and time-driven (for instance monitoring of ecosystem, weather temperature, humidity recording). The periodic data collection (time-driven model) is adopted in this thesis.

In applications where continuous monitoring is required, it has been noted that data changes occur slowly over extended periods. This can lead to significant duplication of data, as well as frequent communication between SNs and sink, which can quickly deplete the limited energy resources available. Consequently, reducing the number of transmitted data can help to conserve energy and increase the network's lifetime. As well as, it minimizes the network overhead at the same time. The amount of data produced by WSN prompted researchers to move towards approaches to obtain data without the need to transmit it over the network. This solution is suitable for applications that do not require real-time data.

Data reduction methods are employed for such purposes. Data prediction is considered to be a crucial criterion to reduce duplicated data of time series. It helps to minimize the number of repetitive transmissions and leads to energy conservation which results in enhanced network lifetime.

1.2 Problem Statement

The WSN composed of tiny battery powered sensors energy, so that the battery is the most important resource in the sensor node that influences the lifespan of WSN. As stated earlier, most of the energy of SN is dissipated as a result of the transmission and reception of data [11]. Since the SN's battery has a finite lifespan, replacing or recharging it might be challenging or impossible, particularly in harsh or remote environments [12]. So, to increase the lifetime of the network, the node's energy consumption must be appropriately managed. This thesis focuses on extending the lifetime of the network by eliminating redundant data transmissions which in turn reduces the amount of WSN energy expended.

However, The challenge is how to design an efficient protocol to minimize the number of transmitting data to achieve energy conservation and extend the lifetime of the WSN while maintaining overall efficiency.

1.3 Thesis Objectives

This thesis aims to achieve the following objectives:

1. Improve an energy-efficient technique based on data reduction will be proposed, which will operate at two levels: the sensor node level and the cluster head level.
2. Minimize the amount of data transmitted and increase the network lifetime.
3. Maintain acceptable data accuracy.
4. Achieving good performance (in terms of data reduction, prediction accuracy, and energy conservation) compared to some of the recent existing related works in the literature.

1.4 Main Contributions of Study

This study focuses on designing and implementing energy-efficient based on data reduction techniques to enhance the lifetime of WSN. The key contributions include:

1. We proposed a Data Reduction Based on a Dual Prediction Scheme (DRDPS) at the sensor node level. DRDPS effectively predicts WSN sensor data readings, conserving energy, reducing data transmission volume, and preserving the accuracy of data readings received at the cluster head, thus extending the WSN network lifespan.
2. We proposed a statistical analysis method at the cluster head level. It is used to identify correlations between the gathered data sets and send only the most representative data sets to the sink. This will minimize the need for transmissions and energy consumption at the CH.

1.5 Literature Review

The issue of limited energy in WSN has spurred significant research efforts aimed at addressing it resulting in a wealth of literature in this field [13]. However, for the purpose of illustrating the impact of our proposed solution, we limit our investigation to two main techniques that are used in this thesis (the clustering technique, and the data reduction technique).

1.5.1 Clustering Literature Review

A clustering process is a form of procedure that is used to save energy and expand the network's coverage. Different research is proposed and developed to prolong the network lifetime of the WSN. The K-means clustering strategy, which is widely used in various applications is a well-known approach for forming clusters. Some researches use K-means algorithm to increase network lifetime by improving it while other searches study the effect of using K-means on the efficiency of work. A brief description of some of this research is given below section:

Handy, M. J., et al. presented a clustering-based protocol called LEACH protocol. It employs a hierarchical approach to distribute the energy load among sensor nodes. LEACH rotates the clusters heads role among the sensor nodes using a probabilistic model at each round and taking in count the remaining energy of each sensor node. This approach ensure a certain nodes from depleting their energy before other nodes. LEACH also employs local data fusion techniques to minimize the communication overhead and conserve energy [14].

Panag T. and Dhillon J. suggested a dual-head static clustering algorithm (DHSCA). In this algorithm, the clusters are formed as static clusters. Each cluster contains two nodes elected to be cluster heads based on their position relative to the other nodes and the base station, and the remaining energy of them. One of these heads allocate to data gathering and the other transmit data to the BS [15].

Jain B., et al. suggested an Efficient K-Means Clustering algorithm called EKMT. This algorithm increases the efficiency of WSN by taking into account the shortest distance between the base station and the newly selected cluster head. The EKMT is based on the concept of locating the cluster head by reducing the sum of squared distances between the closest cluster centres and member nodes, as well as taking into account the shortest distance between cluster heads and base stations. When the remaining energy of a cluster head falls below a certain threshold, the suggested protocol can be implemented to reselect the cluster head by using the k-means technique that considers the shortest distance between the BS and the new CH. The CH selection and the communication between the CHs and BS are issues not covered in this research [16].

Yuan C. and Yang H. presented a comparative study of the four most common methods used to approximate the suitable number of clusters. The presented search included the implementation and analysis of Gap-statistic, Elbow algorithm, Silhouette coefficient, and Canopy algorithms. The Gap-statistic can be used with small data sets but it is not desired when we have large data sets because of its cost in terms of time complexity and space complexity. The Elbow method has the simplest implementation and best execution time, but it totally depends on extracting the inflection point which is difficult to notice when the curve is smooth. The silhouette coefficient also works well with a small data set. It is required to calculate the distance matrix; which is computationally expensive (O (n^2)), especially with a large dataset. The results explained that all prior methods can be used reasonably when dealing with a small amount of data. While Canopy is the most appropriate choice when the dataset is large [17].

El Alami, Hassan, and Abdellah Najid proposed a clustering algorithm based on hierarchical approach called an enhanced clustering hierarchy (ECH) protocol. It is aims to maximize the lifetime of WSNs. They minimize the redundant data produced by neighboring and overlapping nodes by employed sleeping waking mechanism. For more energy saving they proposed enhanced method to select the cluster head optimally [18].

Zhu B., et al. proposed an "improved soft- K-means (IS-K-means) clustering algorithm". It is working to achieve the balance in WSN energy exhaustion. The authors proposed perfection in the selection of initial centers by utilizing the method of "Clustering by FAST SEARCH and Find of Density Peaks (CFSFDP) and Kernel Density Estimation (KDE)". The probability of node membership is used to ensure that low density of each cluster edge to balance the node's number in each cluster. And to get more efficiency the idea of multi-CH within the cluster is used. When compared to various clustering techniques in the literature, the suggested algorithm may postpone FND, HND, and LND on average [19].

Khediri S., et al. introduced an algorithm called optimal K-means (OK-means) based on K-means algorithm. They employed the single-hop and multi-hop communication for the intra-communication and inter-communication respectively. In the initialization phase, the optimal K-means algorithm starts with a random number of clusters K. The optimal number of cluster heads is selected based on nodes density. The suggested approach can produce a uniform spatial distribution of CHs and balance the network's energy use to the greatest possible extent. OK-means results compared with O-LEACH, Bee-Cluster, and LEACH. The aggregation and security issues are not taken into account in this article [20].

Chowdhury A. and De D. explained that the distribution of sensor nodes is an essential issue in WSN to conserve energy and ensure monitoring quality. So that the coverage of the network considers one of the most parameters that should take into consideration to achieve these goals. The authors of this study suggested an approach to enhance the coverage area and expand the network lifetime by using of Voronoi cell structure, K-means algorithm, and Glowworm Swarm Optimization (GSO) algorithm [21].

1.5.2 Prediction Literature Review

Several research works have been done in the area of prediction in WSN, including energy consumption prediction, traffic prediction, node failure prediction, environmental parameter prediction, and anomaly detection. This thesis deals with environmental parameters prediction; this involves predicting parameters like temperature, humidity, etc., which is important for various applications such as agriculture, weather forecasting, etc. The next sections describe some prediction research.

Zhao J., et al. have presented the Periodic Data Prediction Algorithm (PDPA) which is based on a linear regression model. The algorithm utilises the potential periodicity pattern hidden in the data to improve the data prediction process, leading to improved prediction accuracy. The temporal periodic data correlation decreases the frequency of communications in WSN [22].

Shu T., et al. focused on reducing energy consumption in the environmental monitoring of WSN. The authors noted the energy consumed in data transmission operations between SNs and the sink accounted the major proportion of the energy consumed in a WSN. To address this issue, the authors proposed a solution that uses a "dual prediction scheme based on the least mean square (LMS) filter". The dual prediction scheme allows both the SNs and the sink to make data predictions at a simultaneous time. If the difference between the actual observations and the prediction beats a pre-determined thresh, the SNs will transmit the data to the gateway, and the filter coefficients will be updated. The results of this approach showed that it was effective in reducing the number of transmissions, leading to energy savings [23].

Cheng H., et al. used LSTM to develop a bidirectional prediction model, called the "multi-node multi-feature model (MNMF)". They showed that the long-term time series prediction can be improved by combining adaptive algorithms of feature selection, and the dynamic prediction model that can deal with a multi-dimensions time series [24].

Zhang C., et al. suggested an approach that makes use of the correlation between sensors in the same area to enhance prediction accuracy by combining observations from various SNs in a neural network that has multi-headed. The proposed method was compared to general regression neural network (GRNN), a nonlinear autoregressive exogenous model (NARX), "Elman neural network", and a typical local forward network.

Although the reduction of transmission was not included in the work, they found significant improvement in the accuracy of prediction and the conclusion that temporal & spatial correlation can make the performance of the model more efficient. This complicated scheme includes combining observations from various SNs into a single model and then transmitting observations from multiple SNs to the BS or CH to establish the model. The suggested method trains the machine learning by utilizing of moving sensor feature sliding window. Comprehensive experiments are carried out on two real datasets (Intel Berkeley research lab data, and Chicago park district weather and water data) [25].

Soleymani S., et al. utilized the symmetrical correlation of sensor data. They suggest a hybrid strategy based on clustering and prediction approaches. It combines Decision Tree (DT), Auto-Regressive Integrated Moving Average (ARIMA), and Kalman Filtering (KF) methods. the data prediction systems and data aggregation techniques are utilized to minimize redundant communications, and for satisfying sampling requirements in sensor nodes and to decrease processing overheads in cluster heads [26].

Salim C. and Mitton N. proposed Pearson Data Correlation and Prediction (PDCP) algorithm. They utilized the mathematical ability to retrieve data locally instead of sending/receiving it over the network; by using the correlations between data and combining it with the prediction techniques to construct a data reduction method. The objective of this method is to decrease the data transmitted to the sink by considering the level of correlation among various parameters. They applied equations of Pearson correlation coefficient between two nodes. This approach has reasonable accuracy because they implement the prediction on both sink and node levels [27].

Chreim B., et al. offered an autonomous prediction method for diverse applications in wireless sensor networks. The method is given in phases that may be customized to address specific challenges. Correlation analysis is used to predict time series, followed by a Multiple Linear Regression Model layer [28].

Jain K., et al. proposed an extended version of linear regression (ELR). It is an adaptive (i.e., it continuously updates the prediction model based on the latest sensor data) technique for reducing data in sensor networks by utilizing correlation-based data prediction. The basic idea is that predict the value of a sensor's data based on the values of other sensors. The paper addresses the problem of data overload in WSN where a large amount of data is generated [29].

1.6 Thesis Layout

The thesis structure includes the following chapters:

- **Chapter 2:** Presents the fundamental tools and basics that are needed to achieve the goals of the thesis.
- **Chapter 3:** Presents the proposed scheme that is used to maximize the conservation of energy and prolong the lifetime of the whole network.
- **Chapter 4:** Presents the implementation of the proposed model and comparison it with other models, and analysis of results.
- **Chapter 5:** Includes conclusions, and suggestions for future work.

Chapter Two

Theoretical Background

CHAPTER TWO

THEORETICAL BACKGROUND

2.1 Introduction

This chapter provides an outline of the clustering, prediction techniques in WSN, and energy consumption model. Focusing on the strategies needed for avoiding undue transmissions as well as on the approach that will be used to minimize wasted energy.

2.2 Wireless Sensors Network (WSN)

The WSN is a group of wirelessly connected devices called Sensor Nodes (SNs). These SNs are versatile design and are often equipped with a non-rechargeable and non-irreplaceable battery in an unattended access environment. Each SN comprises on four basic units: a sensing unit, processing unit, transceiver module and power source furthermore may contain other optional components for position recognition units such as a Global Positioning System (GPS) and a mobilizer. The sensing unit compromise on two major components; a sensor and an Analog-to-Digital Converter (ADC). The SNs make use their sensors to make collaboratively track the changes in the surrounding environment circumstances and to provide a global view of the target field. The sensed data are digitized to electrical signals through the ADC converter. Then; the transceiver unit transmits it wirelessly to the next station.

The deployment of sensor nodes is a crucial process in which SNs are positioned to enable their optimal functionality and operation in real-world scenarios, as well as in laboratory or simulated environments. The primary purpose of deploying sensor nodes is to gather data about their surroundings and relay it to the appropriate Cluster Heads (CHs) for further processing.

Depending on the nature of the network, various deployment strategies that can be employed for both mobile and static sensor networks. Sensors play several roles in their environment such as being a source node, cluster head, relay node or sink/BS node. Hence, any of the topologies (i.e. star, tree, mesh, star-mesh, … etc.) can be used to deploy SNs. To ensure implement an optimal deployment strategy, it is essential to satisfy the following points: 1) The application requirements should be defined clearly. 2) Striving for the best performance and maximizing the network lifetime. 3) Enable detection of network topology errors and failures.

2.3 WSN Applications

The applications of WSNs have become more widespread, especially with the recent development of sensors, as they have become cheaper, smarter, smaller, and more available. WSNs have evolved from being primarily used for military purposes, then spread to many various applications such as monitoring applications in environments (disaster monitoring/ management), health care, industrial (smart factories), and security applications. The most important attribute to make WSNs suitable for any application is real-time support, low power requirement, and low cost [30].

1. Environmental Applications

One of the interesting applications of WSNs is monitoring of the circumferential environment, involving the monitoring of disaster, monitoring the quality of air, water, and tracking animals/plants [31], analysis of the situation of pollution, debris flow prediction, forest fire prediction, and explosion prediction. For example, WSNs can be used to measure pollutants in the air and water, such as carbon monoxide, sulfur dioxide, and dissolved oxygen. The energy consumption for the environmental systems is generally caused by the radio connection when the data is transmitted/received, and the GPS when data is questioned. To improve the power consumption of such sensors, a power supply is designed for each unit, which is controlled separately. In WSN through using an adaptive sensing technique the behavior of an application is adjusted to check the energy budget of the SNs and ensure its energy does not run out and extend their battery life.

2. **Industrial Applications**

The industrial facilities use a connected network of sensors for ongoing administration and monitoring. WSNs are more efficient than conventional wired systems in terms of infrastructure costs, scalability, elasticity, lack of cabling, dynamic topology, and ease of deployment. These advantages are optimistic for the sector, where significant growth is anticipated over the next few years. In the industrial field, WSNs are used to monitor the production process and the status of the production equipment. For instance, oil producers or chemical plants may use sensors to monitor the condition of miles of pipes. These sensors are employed to alert users in the event of any errors [32].

3. **Healthcare Applications**

WSNs can be used as preventative healthcare networks for humans to track the body functions and to get the physiological characteristics of a person, analyze their conditions, and timely receive medical treatment. This can potentially improve that patient's health through ongoing monitoring at a relatively low cost of health monitoring. And the energy can be harvested from the biomechanical and biochemical sources of the human body and exploited to support the body monitoring network, where they found that the human body wastes an estimated 100 watts, which is energy that can be invested by using thermoelectric generators (TEG) to prolong the life of these networks [33], [34].

4. **Transportation Applications**

Transportation applications in WSNs involve using of SNs to track and monitor various aspects of transportation systems, such as vehicles, traffic, and infrastructure. This can include monitoring traffic flow, detecting accidents or congestion, measuring air quality, and monitoring the condition of roads and bridges. The recorded data can be used to improve traffic management, reduce accidents, and increase the efficiency of transportation systems. WSNs can also be used in intelligent transportation systems (ITS) that are utilized in the timely provision of real recent information to the drivers and other users, such as traffic conditions and the location of available parking spots [35].

5. **Military Applications**

Defense military is a significant application for WSNs. The role of the WSNs is represented by coordinating army movements and monitoring adversary movements. The sensor nodes are placed in the area to be tracked. The sensors work to scan the target area for signs of hostile activity, when distinct occurrences or suspicious activities the SNs will communicate with BS. The SNs will automatically trigger messages to the base station in such cases. The base station collects data from various sensors and implements necessary actions, such as reporting to the responsible authorities or sending notifications to nearby sensors. The base station should be placed in a secure location [36].

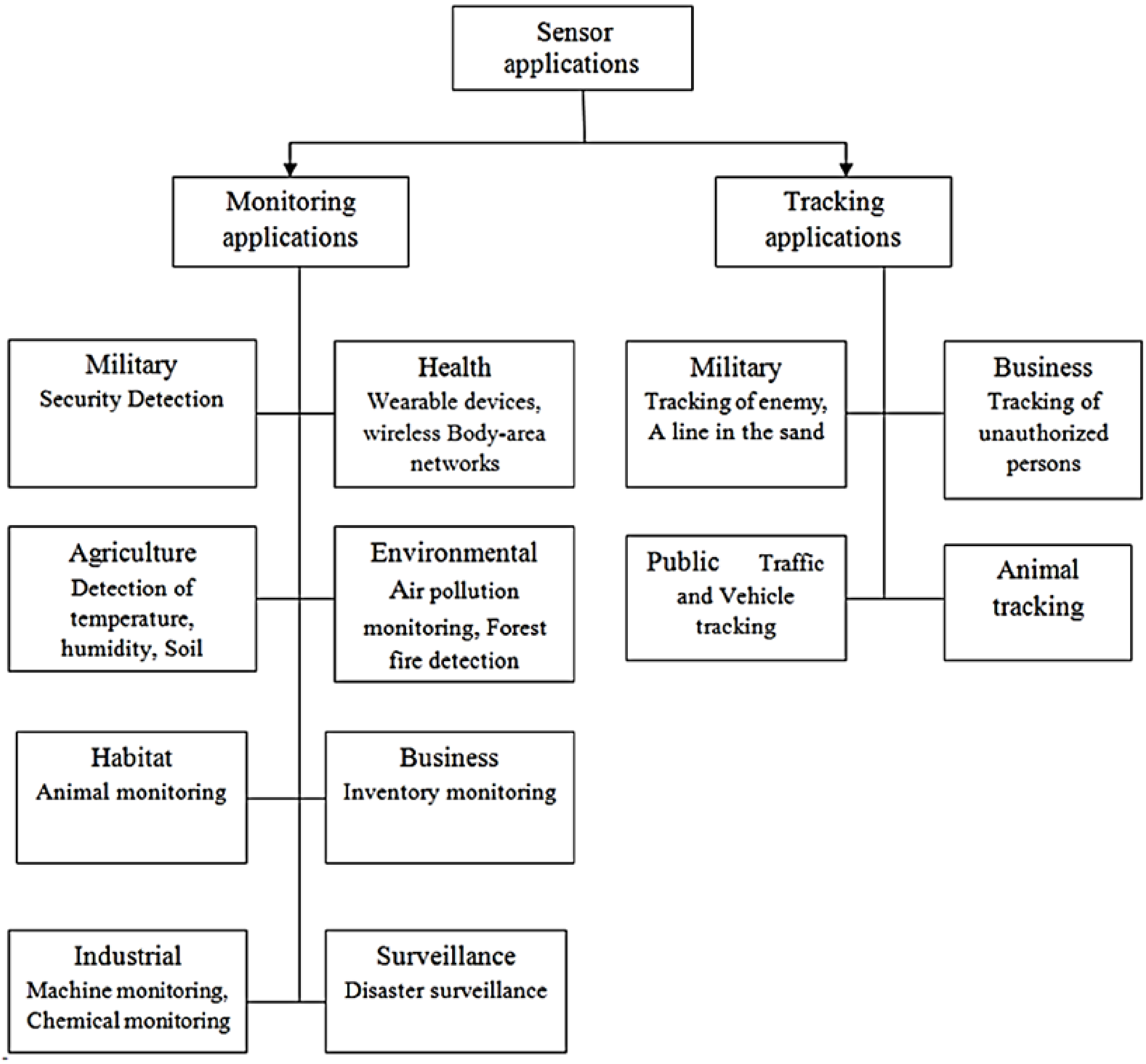


Figure 2.1 Wireless sensor network applications

2.4 WSN Challenges

Although the proliferation of WSNs, it is suffering from many issues that should be taken into account, such as restrictions on the hardware structure and physical attacks, lack of robust security mechanisms, etc. where the SNs can be faulty/ unreliable at any time according to the characteristics that distinguish such network to suit the harsh conditions environments and the huge amount of data. Some of these challenges explored and explained in the next sections . Figure 2.2 illustrated the main challenges.

1. Energy Consumption

The dramatic issue of WSN is the drain of energy over time. The primary goal of WSN is to collect data and send it over the network for additional processing. Moreover, because most of the sensor networks depend on batteries, they have limited energy, and the process of maintaining and recharging the battery is difficult, especially when these networks are in remote areas difficult to reach (e.g. observing natural phenomena such as earthquakes and volcanoes, observing forests or monitor underground areas... etc.). So that, energy conservation is one of most critical points that need to attend in WSN [37].

2. Privacy and Security

WSNs can collect sensitive data, for example, wireless body area network technology (WBAN), is increasingly popular for monitoring and collecting patient health data using wearable sensors. One of the factors that indicate the strength of the network is the ability to ensure that the information has arrived without being accessed, and viewed by an unauthorized entity. The fact that wireless communication is broadcasted means that the information being transmitted can be picked up by external parties who are listening to the radio signals, potentially revealing details from the signal's contextual [38]. This means that the network must be designed to protect the privacy of the individuals whose data is being collected to ensure that patients' records remain secure and protected. On the other hand, WSNs can be attacked via a variety of security risks, such as denial-of-service attacks, eavesdropping, and tampering. The integrity methods may be used to ensure that data are accessed as it is sent by the original source. To maintain security, the network must be designed to protect against these threats and secure the data transmitted over the network. Moreover, the security algorithms should be possible to execute on such memory and processor-restricted devices. The high levels of security and privacy are crucial for protecting these data, both when being used by healthcare professionals and when being stored [39].

3. Localization

WSNs are often deployed in large-scale environments randomly and require the ability to locate the position of SNs. It is a challenge that locates and manage the SNs without any supporting infrastructure, especially when dealing with large numbers of these nodes with limited resources. Localization refers to the process of defining the approximate physical location of sensor nodes in the network [40]. This can be accomplished through a variety of techniques, such as trilateration, multidimensional scaling, and particle filtering. The accuracy of localization in WSNs depends on many factors such as the density of the network, the quality of the sensor measurements, and the presence of interference or noise. Localization is an important aspect of WSNs because it allows for the efficient deployment of sensors and enables a wide range of applications, including monitoring and control of physical systems, navigation, and tracking. The localization algorithms should be scalable, reliable, precise, and distributed, in some cases, they should be able to handle mobile nodes [41].

4. Quality of Service (QoS)

Service quality is among the most interesting challenge in WSNs, particularly for surveillance systems. The quality metrics vary depending on the application requirements. QoS is defined as an evaluation of how the WSN satisfies the demands of a particular application. For example, the quality metric for the animal tracking application includes the accuracy of the approximate location of the target, and the rate of packets delivery for each sensor node QoS may include many evaluation metrics such as quality of measured data, delay, high throughput, low latency, bandwidth… etc. This can be challenging to achieve, especially when dealing with limited resources and a large number of SNs [31].

5. Mobility

In WSNs, mobility refers to the ability of sensor nodes to move within the network. It is a key feature that enables WSNs to adapt to changing environments and perform tasks such as target tracking, surveillance, and rescue operations. There are several protocols and algorithms that have been proposed to support mobility in WSNs (e.g. location-based routing protocols, handover schemes, and target tracking algorithms... etc.) [42].

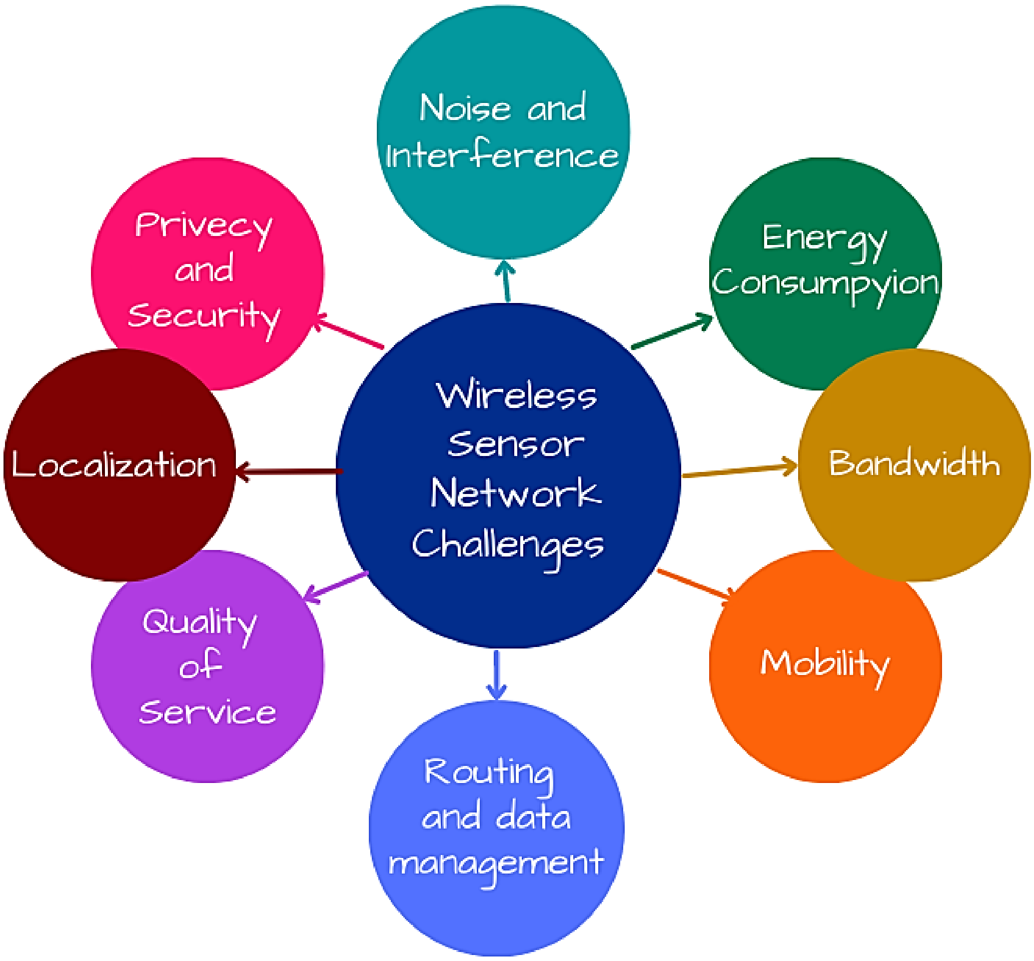


Figure 2.2 Wireless sensor network challenges.

2.5 Network Topology

Wireless sensor network applications are normally optimized by the given underlying network topology [43]. The network topology refers to the spatial arrangement or connectivity pattern of sensor nodes in the network. Most of topology schemes have proven to be able to provide a better network monitoring and communication performance with prolonged system lifetime. The network topology has effect on the communication distance, number of hops, routing paths between nodes. Also, the network topology plays a critical role in determining the scalability of a WSN and the degrees of coverage and so on [44]. WSN topology can be as following:

1. Point to point topology: In this topology, a central hub is not required. A sensor node can directly communicate with other nodes. This is a very popular topology and has a single channel. Every device can be used as a client and a server [45]
2. Star netwrok topology: Unlike point to point topology, a centralized communication hub is required in a star network. In this topology, there is no direct communication between the nodes; every communication is accomplished through the centralized hub [46].
3. Tree topology: This topology is a combination of point to point network and star network topologies. The central hub is known as the parent node. Data is transferred from leaf sensor node to parent sensor node .The main benefit of this topology is consuming less power as compared to other network topologies [47].

2.5.1 Clusters Based WSNs

Clustering is one of the most popular unsupervised energy-efficient techniques used to break down a large network into groups [48]. Clustering is considered the most commonly used approach for topology management in

WSN which can improve the efficiency of the network. Clustering architecture gets this attention because of its advantages over the flat WSN in terms of energy conservation, communication efficiency, topology management efficiency, delay minimization, and network lifetime extension [49], [50]. The principle of clustering is that partition the network into logical groups (called clusters) based on certain features (such as distances between nodes, nodes density…etc.) and the application requirements.

The basic idea of clustering implementation is to reduce communication distance between nodes and sinks, hence energy saving [51]. Clustering localizes message transmission and reduces the quantity of messages flowing through the network. In the clustering structure, each sensor node called cluster member will connect with a certain CH [52]. that's able SNs to avoid direct contact with sink node/BS, which may cause resource consumption inefficiency, energy depletion, and interference if it happens.

The clustering process can be centralized or distributed. Choosing which one is implemented depends on what is needed. If the network requires longest lifetime as possible then the centralized approach is more efficient since it utilizes BS global view of the network topology, and it is more suitable for high computations i.e. optimization algorithm, it offers reasonable quality and ensures balancing of CHs distribution through the network with respect the remains energy of each sensor node while distributed protocols have advantages such as fault tolerance, fast execution, and self-adaptation [52].

In a clustered network, every cluster incorporates three types of nodes. These nodes are the Member Node (MN), Cluster Head (CH), and Gateway Node (GN), figure 2.3 [53]. In fact, MNs cannot deliver data straightly to the BS, but only through the CH, which in turn forwards them to the BS [54]. It has been confirmed that a clusters heads all together with the encircling nodes consume more energy than the remaining nodes, and this is attributed to heavy cluster traffic of both types whether inter or intra, in addition to the process of collecting data.

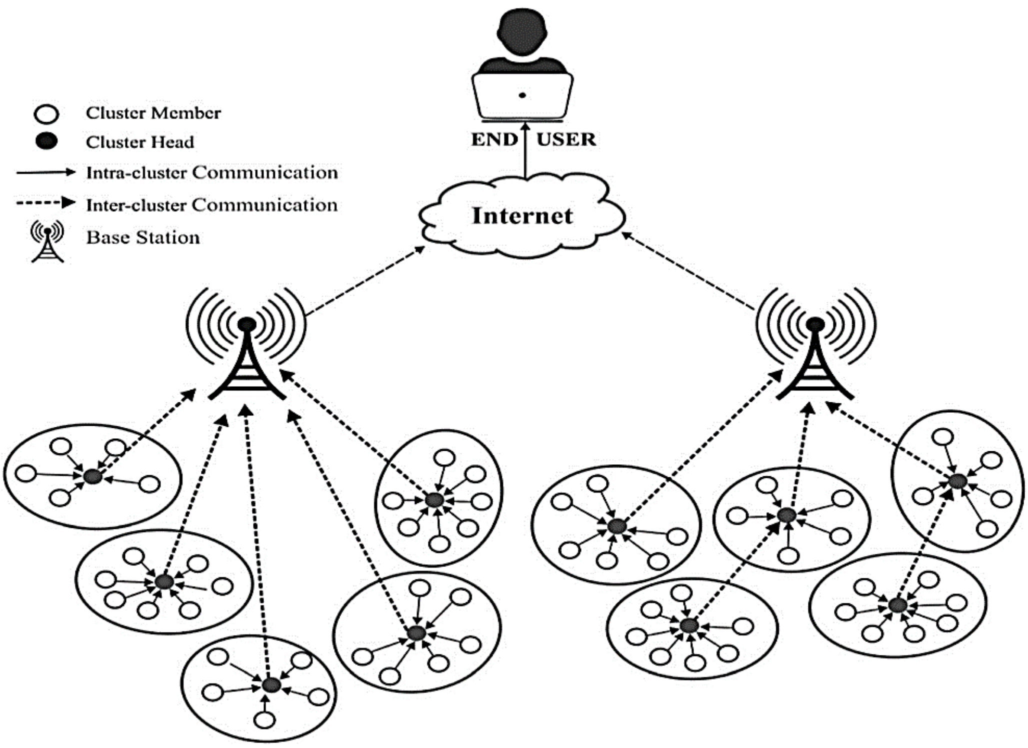


Figure 2.3 Network cluster component.

2.5.2 Characteristics of the Cluster

Clustering characteristics are associated with the internal structure of the cluster and include factors such as cluster size, clusters count, and cluster communication. The cluster size is classified into two classes: equal and unequal, with the same and differing sizes of clusters. Unequal clustering achieves uniform energy usage and avoids the energy-hole problem. Also, the cluster counts can be variable or fixed based on a variety of clustering algorithms. In clustering, communication between sensor nodes can take two forms: communication inside and across clusters. The communication that takes a place between SNs and the CH is intra-cluster communication. Inter-cluster communication, on the other hand, is the communication between the CHs and the BS or sink node [55].

2.5.3 Objectives of Clustering

Different objectives were explored in the literature based on the desired application. Following we'll go through some of the most important objectives of clustering [49]:

- **Scalability** refers to the ability of the network to handle an increasing number of nodes without significant degradation in performance [56].
- **Fault tolerance** is the ability of a system to continue functioning and reconfigure itself even in the presence of hardware or software failures [57].
- **Lifetime** refers to the duration for which the network can operate correctly[58] .
- **Data aggregation/ fusion** is a technique used to reduce the amount of data that needs to be transmitted to the BS, which relieves SNs of additional communication load [59].
- **Latency reduction**: The overall time taken to transmit messages from a source node to a destination node is called latency.
- **Secure data communication** refers to the ability of the network to resist security threats, such as eavesdropping, tampering, and denial-of-service attacks [60].
- **Quality of services efficiency (QoS):** WSN functions and network applications necessitate the requirement QoS. Effective QoS characteristics are often bandwidth, jitter, throughput, reliability, and end-to-end latency [61].

2.5.4 Clustering Algorithm

Clustering algorithms can be divided into partitioning and hierarchical techniques. The K-means algorithm is an unsupervised non-hierarchical popular clustering algorithm that is based on a partitioning (non-overlapping clusters) approach. The principle of this algorithm divides the entire network into K-clusters [20]. The cluster number is represented by K. The cluster formation is established based on Euclidian distance. The objective function of K-means algorithm is known as a sum of squared function, which aims to minify the sum distance between the cluster centroid and other points that belong to that cluster [62]. Each SN in the network belongs to only one cluster rather than being overlapped.

For the K-means algorithm, the cluster number is an important issue that needs to address, since it`s required to be entered by the user. Many suggestions are presented; the authors in [63] propose uses of DB-SCAN algorithm to estimate the number of clusters. Moreover, authors in [64] showed that Silhouette method and Elbow method, are the two recent most commonly used to choose the appropriate number of clusters. Gap-Statistic, Canopy is another way presented to initialize the K value [17]. The centroids locations selected by K-means convergence from the optimum solution [65]. A cluster's centroid is a point whose values are the average of all the points in the clusters. However, due to its simplicity of interpretation/ implementation, and speed, it is very helpful in a variety of scenarios. In addition, K-means can deal with different shapes and sizes of clusters.

The non-hierarchical partitioning algorithms, such as K-means provide rational results on large datasets and a short time in comparison with the hierarchical algorithms. K-means has a linear time complexity making it relatively superior on other clustering algorithms [66], [65], [67].

The steps of K-means algorithm are as follows: -

Inputs: D is a data points collection;

K is the desired of cluster number.

Step 1: Identify K locations at random positions from D to serve as centroids/ cluster centers. It is ideal to the initialization phase is that put the centroids as far as possible from each other.

Step 2: For each data point, find the closest centroid. Use Euclidean distance, to calculate the Euclidean distance between each centroid and data-points. Suppose C_i = (p_1, p_2, p_3... p_n) and P_j= (p_1, p_2, p_3... p_n), Euclidean distance is described as follows:

$$Distance(C_i,\ P_j) = \sqrt{\sum_{i,j=1}^{n} (C_i - P_j)^2} \qquad (2.1)$$

Step 3: Group points with minimum distances to form clusters.

Step 4: Using the average values of all data points in the corresponding cluster, determine the new cluster centre.

Step 5. Repetition of step 2 with the updated centroids; if the data points are assigned to different clusters, go back to step 3; otherwise, end the procedure.

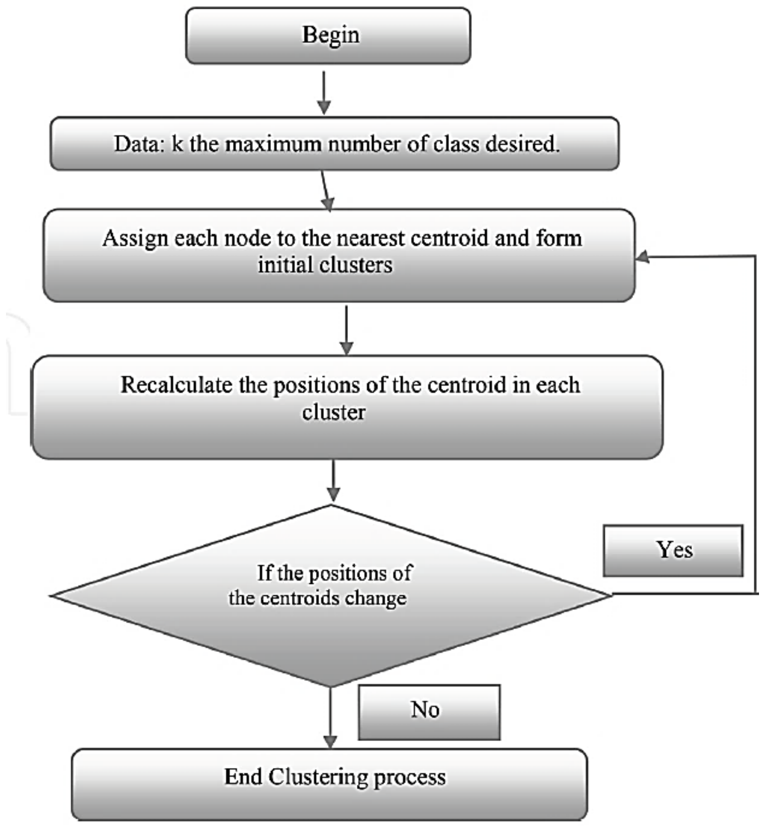


Figure 2.4 Procedure of K-means algorithm.

2.5.5 Elbow Method

The Elbow method is used in machine learning to calculate the suitable clusters number. The basic principle of Elbow is the degree of aggregation of each cluster would gradually grow with an increase in the clustering number K, leading to a naturally decreasing sum of squared errors (SSE). Reducing values of the SSE will drastically decline and flatten out with the growth in the K value [3], [68].WSS can be formulated as bellow.

$$WSS = \sum_{i=1}^{k} \sum_{p \in c_i} (p - \mu_i)^2 \qquad (2.2)$$

Where K and C_i are the number of clusters, p is each element in the cluster C_i, and μ_i is the C_i centroid (refers to the average of all points within a cluster). Initially start with K=2, increment K by one until reaches the maximum predefined (expected) ideal number of clusters, and then identify the prospective optimal cluster number K that corresponds to the plateau. The ideal number of clusters is determined by the point where the cost function rapidly decreases up to a certain cost peak value before K and then continues to increase with almost no change in the peak value after exceeding K. Elbow suppose that the point at which there is a considerable shift in the SSE value, resulting in the formation of an angle, is regarded to be the best location to determine the K value.

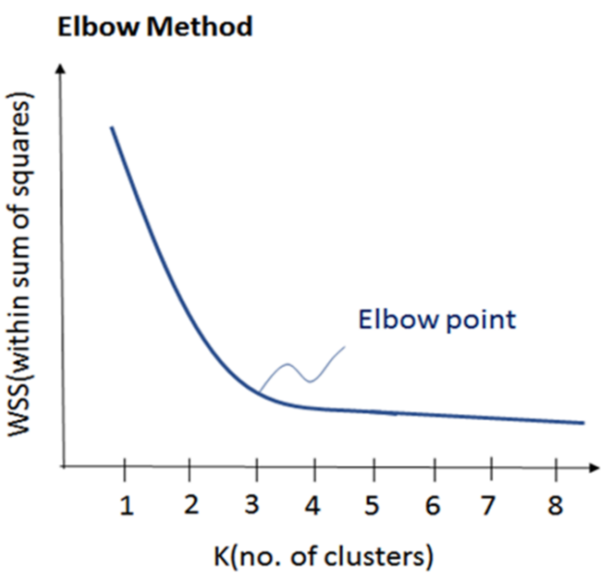


Figure 2.5 Elbow method.

2.6 Data Prediction

Data prediction is a popular data reduction method that aims to decrease the number of transmitted and sensed data as well. Prediction is a process of constructing a model to predict future values by using previously collected data, based on statistical computations and probability operations. The fundamental purpose of using prediction techniques is to create a regression model that represents the association between past, present, and future data values [69]. Many techniques were developed to solve the continuous communication of resource-constrained sensor nodes.

There are two types of prediction models; the first type is backward models that reconstruct the historical data and this takes superfluous time causing a delay. The other type is forward models that predict future values based on previous ones [70]. Data prediction approaches attempt to provide data locally without the need to brought it from a remote node. In the network, two model instances of prediction algorithms are formed: the first one, in the BS/sink side and another in the corresponding source node. The model in the BS answers all current queries without requiring connection with other nodes, thus minimizing communication costs and, as a result, energy usage saved. Each sensor node is preparing a prediction model by training on the current data and then equipping the sink with this model. After that, the node predicts/reconstructs detected data using the same model as the sink. A particular threshold is considered to adopt the predicted values and regard it as acceptable.

The data prediction method categorizes into algorithmic approaches, time series forecasting, and stochastic approaches [71]. The predictive model may create in the sink node as well, the same model is kept on the other nodes by utilizing the synchronization; this assumption will ensure both models produce same predictive values. In other cases, the prediction model generated at the sensor node then sends to the sink, this scheme requires continuous updating each time model get deteriorates when the model results don't match the actual data. This additional energy consumption to update the model can be reduced by putting the prediction model on the node side only; in such case, the node is decide when the reading data should be sent [72].

There are some areas where prediction plays a key role in WSNs. The specific prediction methods used in these areas can range from statistical methods like regression analysis to machine learning techniques.

2.6.1 Time Series Model

Time series is set time-based evenly spaced observations/data collected over specific series of periods (the time unit might be in minutes, hours, days, years, etc, but the time difference between any two consecutive samples will be the same). The key feature of such data is time-dependent, where time is an important vector in time series and this is what specialize it from regression methods, which are based on independent relations between variables hence, the regression method is not suitable for such data [73].

Time series analysis is a crucial statistical technique that enables us to understand the characteristics of changes in the values of a particular phenomenon over time, identify their causes and their effects, interpret the observed relationships between them, and predict future changes based on historical data and previous measurements of the phenomenon under study. The spatial patterns correlation can be used in analysis processes to able forecast future values [74].

Generally, time series forecasting categorizes into univariate time series forecasting (if the forecasting depends on past values only, predict the future time series values by using only its previous values), and multivariate time series forecasting (if the forecasting depends on external variables in addition to previous values. It is a chain of interrelated univariate time series, each one describing a specific variable (such as temperature, wind speed, humidity, etc.)

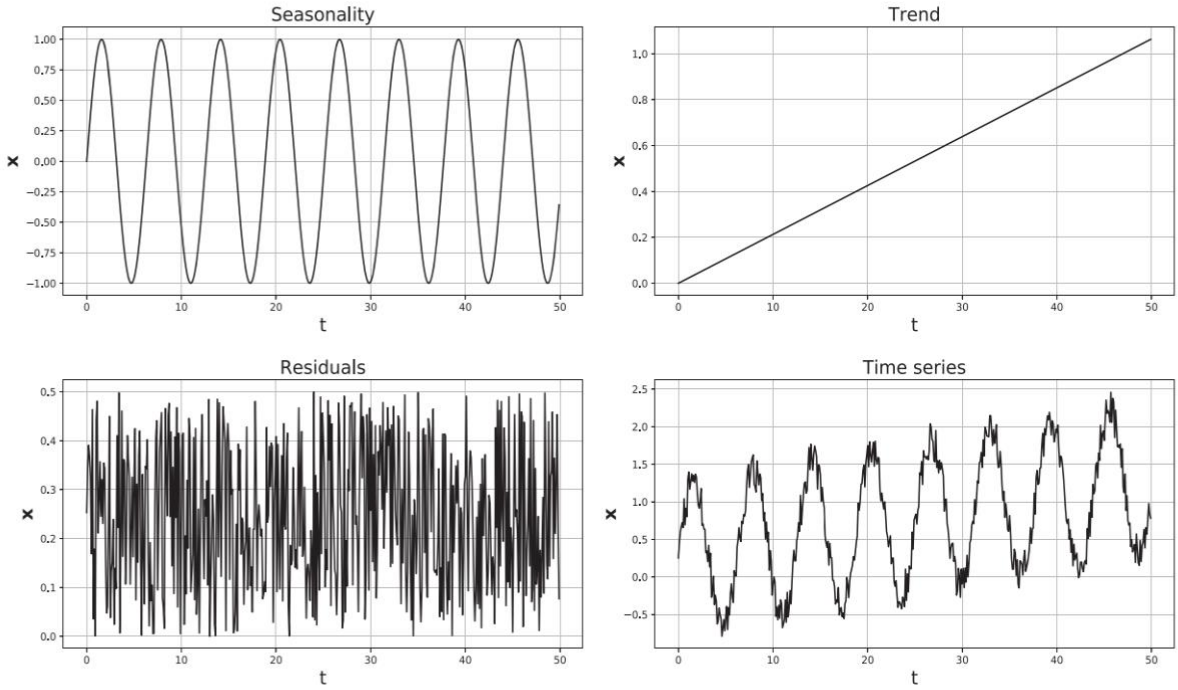
and exhibiting correlation with other variables in the series. Unfortunately, most methods can predict the value of a single variable only at a time [75]. ARIMA model is the most common linear algorithm example of univariate time series used in WSN [71], [76].

Time series data may be trendy, seasonal, recurrent or irregular [73], [77], [78]. In trend exploration, the analysis process tends to be growing, waney, or stay slack for a long period. Seasonal analysis reveals variations that recur at regular intervals within the observed time. The recurrent analysis describes patterns that may repeat in a single cycle in terms of average where each cycle might take a relatively long time. Irregular patterns are showing unexpected vicissitudes so the prediction is difficult using time series methods.

In time series, the stationary nature of data is really important so that can analyze this data using statistical methods where it is not possible to analyze data that does not show this feature. After removing the trend and cyclic fluctuations, there may still be residual values. Sometimes these values can be so high that they can hide both the seasonality and the trend. Robust statistics are typically employed to deal with these residuals, in this instance are referred to as outliers. the authors in [79] suggest methods to deal with such residuals.

The term "stationary" refers to data that has consistent changes/features over time. The series of stationary data has three traits: a settled covariance; settled mean; and settled variance. If a time series is stationary and exhibits a certain pattern across a given period of time, it is reasonable to think that it will exhibit the same pattern at a later time. Error-Trend-Seasonality separation (ETS separation) is the process of separating the error, trend and seasonal series of data into different components [74].

Time series mostly have serial correlation/autocorrelation relations where the current observation distribution depends on the previous one. This correlation makes time series not stationary so differencing is used to remove this autocorrelation. Autocorrelation and data stationary are checked by the autocorrelation function (ACF). The autocorrelation function plots the relationship between each observed data and its prior values at different lags. The lag refers to the correlation between an observation and its prior values with discarding the linear effects of the intermediate points between these observations (lag is the number of time points of the consecutive observations). The counterpart to the ACF is the partial autocorrelation function (PACF). It shows the association between an observation and its previous values that cannot be explained by the correlation at small lags. In the stationary series, the



autocorrelation in the ACF plot should rapidly decrease. In contrast, a non-stationary series will have a slowly decreasing ACF [80].

Figure 2.6 Illustration of seasonality, trend, residual, and time series.

2.6.2 Time Series Forecasting

A forecast is a prediction process of future values that is performed by using historical series data with taking the time dimension into account. A forecast can be one or several steps ahead. In the first situation, prediction is implemented to find the next value only, while in the second situation, the forecast is done for forecasting horizon which means forecasting several values in the same iteration [81].

Most of the forecasting algorithms statistical-based are put to work on static/ stationary time series. If data showed non-stationary, it should be transformed into stationary time series by eliminating the seasonal components and trends [75]. Statistical forecasting methods include Naive and sNaive, Moving Average, Exponential Smoothing (ETS), Double Exponential Smoothing (DETS), Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA). Figure 2.7 [82] shows the data reduction architecture based on the prediction methods of WSN, where categorized into a Dual Prediction Scheme (DPS), and Single Prediction Scheme (SPS).

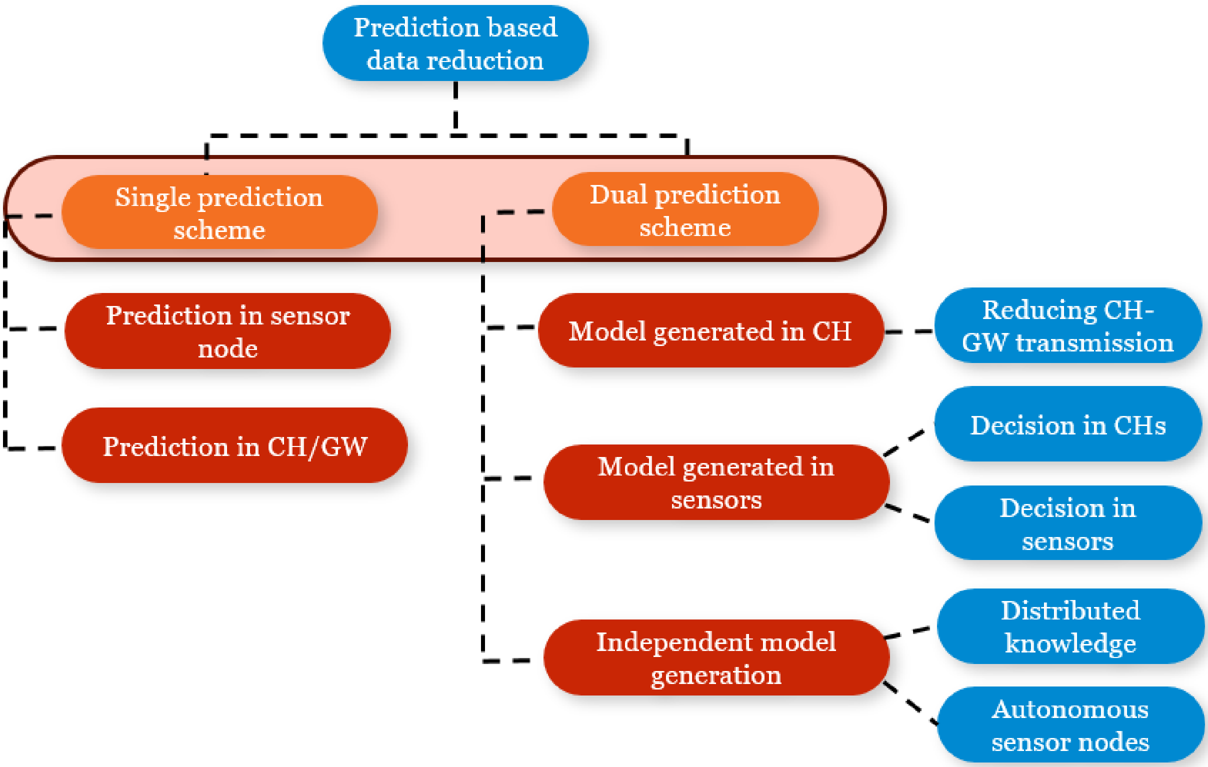


Figure 2.7 Data reduction architecture based on prediction methods of WSN.

1. **Single Prediction Scheme (SPS)**: the prediction method is implemented in a single node (could be either SN, CH, or BS). When the forecasting method is applied by BS or sink node, the accuracy of the current

prediction is taken into consideration to judge whether there is a need to make receiving data from other nodes and to determine the appropriate time points to make the next prediction [83]. This approach utilizes the spatial and temporal correlation feature in collected data that presents in close nodes and uses a certain value as a threshold, the prediction results are acceptable if it was in a certain range [84]. Alternatively, apply predictions in the SN if the prediction performs less cost than the data gaining.

2. **Dual Prediction Scheme (DPS)**: in a DPS, both SN and the sink node can make predictions. They are implementing the same prediction model to produce similar results. This has a benefit when the current period data and previous period data are approximately the same to avert redundant data transmission. The SN is continuously comparing the prediction result with current observed data in each iteration to ensure prediction accuracy, and if the difference (error) is sufficiently large(exceeds a pre-defined threshold), data transmission to the gateway is done to update the coefficients of the prediction model, and to use it in the next forecasting [23], [85]. In this manner, the entire process not only maintains updates of the most recent changes in the environment but also makes sure that forecasted data are pretty correct and transmissions are reduced, thereby reducing energy usage in long-term environmental monitoring applications. Figure 2.8 provides a detailed explanation of how the DPS is implemented.

When SN connects to a CH, n samples ($x_1, x_2, ..., x_n$) are first gathered by the SN and then sent to the CH. Both CH and SN initialize a prediction model using those samples. While DPS is being implemented, the sensor node and the CH both continuously predict data, which are represented by the symbol $\hat{x}$. At every iteration, the SN compares the recently sampled data with the recently predicted data, and if the differences are acceptable, the recently collected data is discarded. Otherwise, the SN will alert the CH by sending the collected data so that readjust the prediction model bilaterally if the difference exceeds a pre-defined threshold. However, the DPS implementation necessitates setting a strong network protocol, to ensure that every time a correct transmission certainly occurs, the subsequent receiver (such as SN, CH, or GW) will be able to successfully receive the data. The prediction model's coefficients may not be updated if the data are not synchronized correctly, rendering the DPS implementation ineffective [86]. In this thesis, the data transmission failure is not taken into consideration.

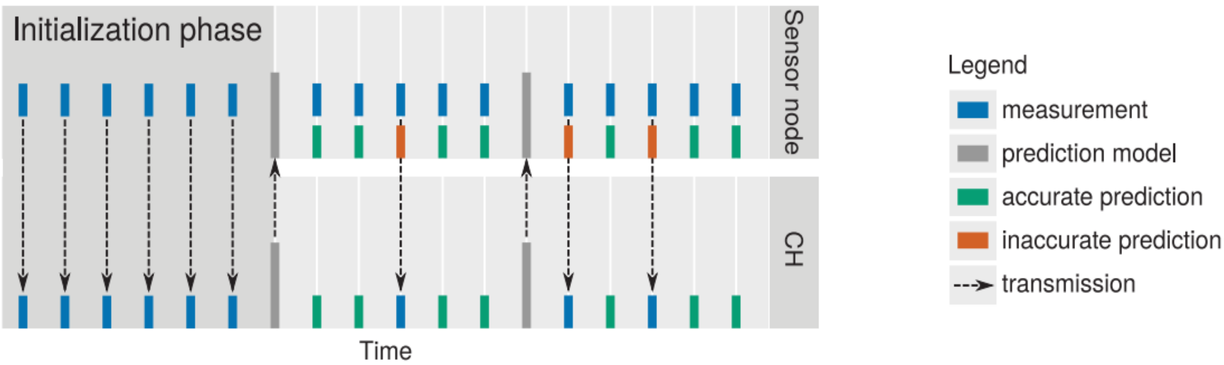


Figure 2.8 Dual prediction scheme in WSN.

2.6.3 ARIMA Model

The ARIMA model (also known as the Box-Jenkins model) is a linear regression model. It is one of the most recent prediction models for univariate time series used [87], [88]. The SNs and CHs reduce the amount of energy depletion within the network through utilize ARIMA to predict the data of the next periods based on the same quantity of observed data. There is no obvious size of the sample that is required to apply ARIMA model. In [89] Box and Jenkins suggest that the minimum number of time points is 50 but this does not have a practical basis. With short time series, ARIMA works well if the time points are enough to assess all parameters [73].

ARIMA method consists of four main processes: identification, estimation (model recognition), verifying (diagnostic checking) and predicting(forecasting) [90]. Before implementing these stages, it is necessary to ensure the stationary of the time series.

The user of this method can suppose a temporary model that is believed to be suitable for the data type and then test the predictive accuracy of the proposed model to determine whether the model makes the prediction correctly. With this iterative procedure, a predictive model with minimal errors can be obtained to represent the studied time series.
ARIMA model components include the following :

1. **Autoregressive (AR) model**: AR is the number of lags of y that can be used as predictors. It assumes that the present value y_t of the time series is dependent on and correlates with the past values (y_{t-1}, y_{t-2}, …, etc) of the series itself. The values of one or more lags are utilized to predict y_t. Equation 2.3 represents the AR model [91].

$$y_t = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t \qquad (2.3)$$

Where c is a constant value, t is the time $t = 1,2,3, \ldots, n$ where n is the length of the time series or the number of observations. p is the lagged values of y_t weighted by the corresponding coefficients α_1, α_2, …, α_p. These coefficients capture the influence of past values of y_t on the current value. ε_t represents the error [92].

2. **Moving average (MA) model**: While AR uses historical data, MA considers prior forecast errors. The MA model assumes that the current

value y_t is dependent on the error terms including the current error (ε_t, ε_{t-1}, ε_{t-2}, ..., ε_{t-q}). MA model is given in equation 2.4.

$$y_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \qquad (2.4)$$

The model includes q-lagged values of the error denoted by ε_{t-1}, ε_{t-2}, ..., ε_{t-q} that are weighted by the corresponding coefficients $\theta_1, \theta_2, ..., \theta_q$. These coefficients capture the influence of past values of the error term on the current value.

3. **Differencing (integration)**: To generate meaningful predictions for an ARIMA model, the time series being modelled must exhibit stationarity, which means that the observations should not be dependent on time. However, for time-dependent data, such as seasonal rainfall, this condition may not be met since different time intervals can result in different observation values. To make the data stationary, trend and seasonal patterns that negatively affect the regression model are removed by applying a degree of differencing. The "Integrated" component of ARIMA, denoted by the parameter d, represents this differencing process equation 2.5. By computing the differences between adjacent data points, differencing transforms the time series into a stationary one.

$$y'_t = y_t - y_{t-1} \qquad (2.5)$$

The AR and MA models are combined with the differencing to establish what is known as ARIMA model. ARIMA formula could be represented as:

$$\acute{y}_t = c + \alpha_1 \acute{y}_{t-1} + \cdots + \alpha_p \acute{y}_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (2.6)$$

Where α represents the autoregressive (AR), θ represents the moving average (MA) and ε represents the error . The order of the AR and MA is denoted by p and q, respectively.

The best values of AR and MA determined based on values of Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). These values are used to compare the models in terms of quality, which is achieved by striking a balance between a model's fit and its complexity. The formulas for determining the BIC and AIC are as follows:

$$AIC = -{}^{2l}/_{T} + {}^{2k}/_{T} \tag{2.7}$$

$$BIC = -{}^{2l}/_{T} + {}^{(k\log T)}/_{T} \tag{2.8}$$

In equations 2.7 and 2.8, l represents the log-likelihood, T represents the number of sensed data, k represents the number of regressors on the right-hand side, and $\hat{\varepsilon}'\hat{\varepsilon}$ in equation 2.9 represents the sum of squared residuals.

$$l = \frac{T}{2}\left(1 + \log(2\pi) + \log({}^{\hat{\varepsilon}'\hat{\varepsilon}}/_{T})\right) \tag{2.9}$$

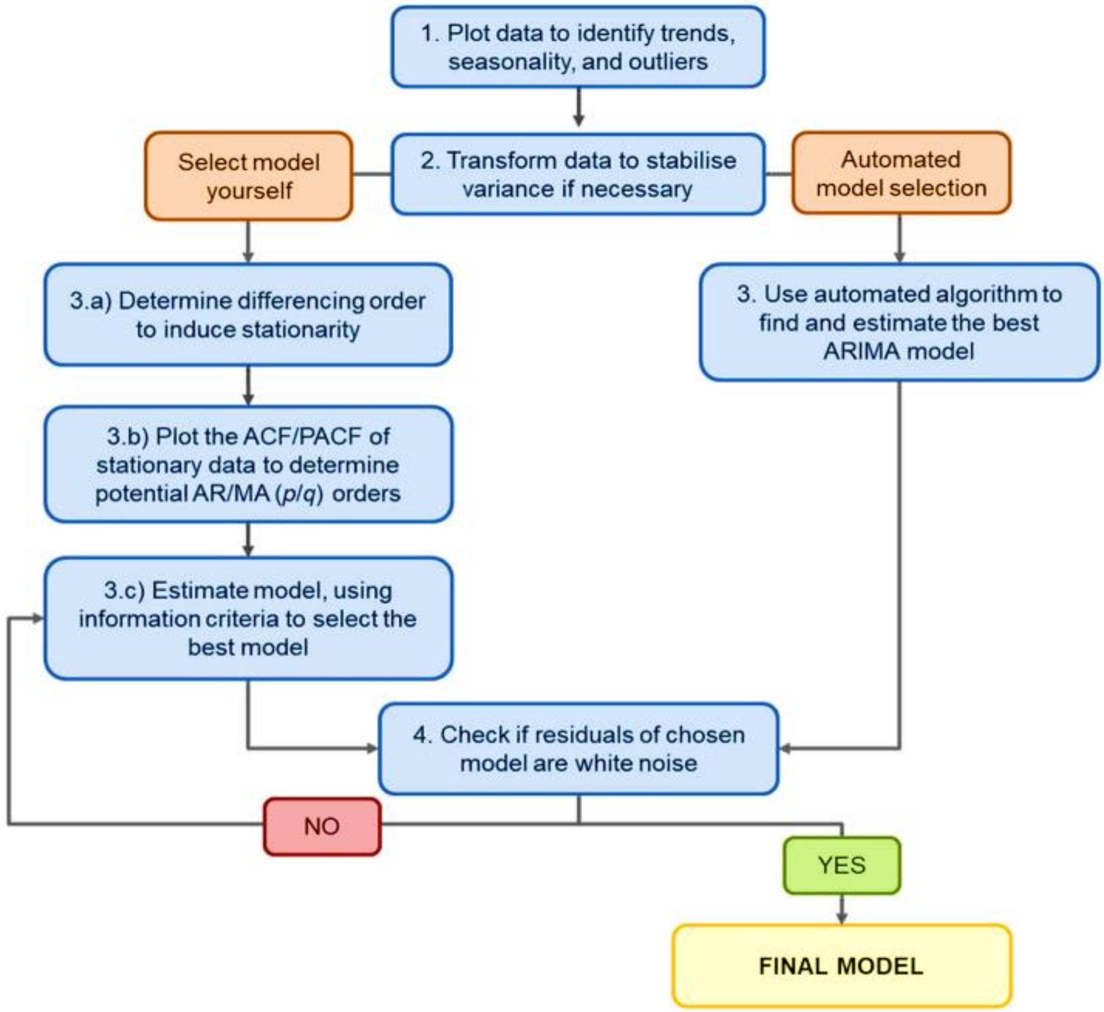


Figure 2.9 Building of ARIMA model [93].

2.6.4 Granger Causality

Forecasting methods are divided into systematic methods and in-systematic methods. The systematic methods are objective, giving the same results when using the same information to interpret any phenomenon. Systematic methods in turn are also divided into causal models and non-causal models. Causal models are based on the response of an element (dependent variable) to the influence of other phenomena (independent variables). For instance, during the past thirty years, researchers began analyzing fluctuations and changes in air temperature and studying them by determining the direction and amount of this change and trying to predict the weather for the next 24 hours in terms of different climatic elements. Predicting the air temperature with the maximum and minimum temperatures, which are considered more significant climatic conditions in the agricultural field. Non-causal models are based on deconstructing time series into their components. Which is considered the most common method if it is applied and achieved for statistical hypotheses. The most prominent of them are ARIMA models according to the Box Jenkins theory of time series.

Granger causality characterizes the dependence relations between bivariate time series, it talks about linear prediction, it is statistics widely used to examine whether it is possible to predict the value of a variable depending on lagged values of another variable (it assesses the temporal relationship between two events and facilitates its prediction) [94].

Definition: The concept of Granger causality involves the relationship between two stationary series, x_t and y_t. The x_t is Granger causal for y_t if x_t helps to predict y_t at some stage in the future where the regression result of y_t in terms of its past value is less accurate than the regression result of past values when use x_t and y_t together. The Granger model is defined as follows:

$$y_t = \alpha + \sum_1^k Y_k y_{t-k} + \sum_1^k \beta_k x_{t-k} + \varepsilon_t \qquad t = 1, \dots, n \qquad (2.10)$$

Where y_t is the variable needed to be examined, α is a constant, k represents the largest lag order, and ε is regression error. An F-test can be used to investigate this causality by testing the following null hypothesis:

Null hypothesis (H_0): Time series x does not Granger-cause time series y

$$H_0 : \beta_1 = \cdots = \beta_K = 0 \qquad (2.11)$$

Alternative hypothesis (H_A): Time series *x* is Granger causes time series y.

One may get the conclusion that there is causation from x to y if H_0 is rejected. It is feasible to see bidirectional causality by switching the x and y variables to test for causality in the other direction (in this instance, we are referring to a feedback system) [95].

Let $\{x_t\}_{t=1}^{T}$ represents the data sets at the cluster head CH_i and $\{y\}_{n=1}^{m}$ denote the representative data sets that will send from CH_i to the *BS*. Each set in y is "*Granger cause*" for one or more other sets in *x*. A linear regression model called Vector Auto-regression (VAR) is used to fitting two different regression models to the data: one model uses only the past values of the first time series to predict its current value, while the other model adds the past values of the second time series as additional predictors. The difference in the prediction error between these two models is then used to test the hypothesis of Granger causality. The residuals of the VAR model are computed and used to complete the statistical test (F-test) of GC. The F-test determine if the addition of lagged values of the first time series to the VAR model significantly improves the overall fit of the model when predicting the second time series.

The formula for the F-statistic for Granger causality is:

F = ((RSS_r - RSS_u)/q) / (RSS_u/(T - k –2 q)) (2.12)

RSS_r is the residual sum of squares from the restricted model.

RSS_u is the residual sum of squares from the unrestricted model.

q is the number of lags used in the VAR model.

k is the number of variables in the VAR model.

T is the number of observations in the time series data.

The F-test result of both models is compared to a critical value (e.g. 0.05). If it is lower than the critical value, we reject the null hypothesis of no GC and conclude that the potentially causal variable is a significant predictor of the other variable [86].

2.7 Evaluation Parameters

There are several parameters to evaluate the efficiency of the suggested method some of them are described below [96]:

1. Energy Consumption

Energy consumption is a crucial factor in the development of WNSs [97]. First order radio method is employed to evaluate the energy of the presented method. We just take into account the communication unit energy usage. The communication energy includes the transmission and the reception energy. The same energy consumption model indicated in [98] is used to evaluate the proposed methods' energy usage. The energy consumption model for the wireless channel is presented in equation 2.13 which represents the energy consumption to transmit *K-bits* data from sender E_{tr}. Equation 2.14 represents energy consumption for the received *K-bits* data in E_{rc}.

$$E_{tr}(K,d) = \begin{cases} E_{elec}\,K + K\,\varepsilon_{amp} d^2 & d < d_0 \\ E_{elec}\,K + K\,\varepsilon_{amp}\,d^4 & d > d_0 \end{cases} \tag{2.13}$$

$$E_{tr}(K) = E_{elec}\,K \tag{2.14}$$

The energy required to transmit *K*-bits is proportional with the distance. The E_{tr} is the energy consumed per sensor. E_{elec} refers to the energy that is expended to operate the transceiver. E_{elec}=50 nJ/bit. $\mathcal{E}_{amp}$ is the energy dissipated of transmit amplifier with different distance levels $\mathcal{E}_{amp}$=100 $pJ/bit/m^2$. K is the length of the message, d is the transmission distance and d_0 is the threshold distance given by equation 2.15 [99], [100].

$$d_0 = \sqrt{\frac{\mathcal{E}_{\text{amp}\, free\ space}}{\mathcal{E}_{amp\ multi-path}}} \tag{2.15}$$

If distance value d is smaller than d_0, the free-space propagation model is deployed otherwise; it is a multi-path fading channel model.

At the sensor node level, and after the end of every period, every sensor node will have several data readings. The length of the message forwarded is computed as follows:

$$message_{\ length} = Number\ of\ readings\ in\ the\ data\ set \times Number\ of\ bits \quad (2.16)$$

2. **Network Lifetime (NLT)**

It is one of the important metrics which reflects the efficiency of a WSN. The application specifications dictate when to consider the network to be nonfunctional, and numerous metrics are employed to assess the network's functionality [101]. The first node died (FND), half node died (HND), and the last node died (LND) providing important information about the energy consumption patterns in the network. FND is round at which the first node exhausts its energy. HND is round at which half of the nodes exhaust their energy. LND is the round at which the last node in the network exhausted its energy. The node is considered alive if its residual energy exceeds a certain threshold. Conversely, if its energy falls below that threshold, the node is declared dead. When the number of live nodes drops below a predetermined threshold, the network is considered to have collapsed since dead nodes can no longer transmit or receive data.

3. **Prediction Accuracy**

Prediction accuracy refers to the degree of correctness of a prediction made by a model, algorithm, or system. The concept of data accuracy is determined by the requirements of the particular application for which the network is being created. The data accuracy is represented by the rate of measure loss. It is the disparity between the original sensed data and predicted data. The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are the two most often used accuracy metrics that depend on scale. The most used scale-independent measurement is Mean Absolute Percent Error (MAPE) [102]. In this study, RMSE is used. The accuracy increases with decreasing in RMSE value. The definition of RMSE as the formula:

$$RMSE = \sqrt{\frac{1}{n-m} \sum_{t=m+1}^{n} (y_t - \hat{y}_t)^2} \qquad (2.17)$$

4. Amount of Transmitted Data

It refers to the amount of data that is transmitted through the network at each period.

Summary

This chapter provides an overview of the theoretical concepts. The most common applications of WSN are presented; including environmental monitoring, industrial monitoring, healthcare applications, business applications, and so on. The main challenges of WSN are also explained. The chapter also dealt with some methods of energy conservation in WSNs, which included the use of network clustering and data reduction that is based on the prediction methods to maximize energy conservation in WSNs. At last, the common evaluation metrics that are used in this thesis are introduced and explained.

Chapter Three

The Proposed Protocol

CHAPTER THREE

THE PROPOSED PROTOCOL

3.1 Introduction

The purpose of this chapter is to provide an overview of the proposed protocol utilized in this study. The chapter gives a detailed illustration of the proposed clustering and data prediction techniques. The general architecture and all the algorithms used to build the protocol should help to provide a better understanding of the proposed protocol.

3.2 Network Design and Preliminaries

This section describes the network model, the problem description, the assumptions of the network, and the energy depletion model.

3.2.1 Network Design

In this thesis, we emphasize a scenario where a set of sensor nodes has been organized into a WSN with the goal of collecting data and transmitting this data to a sink. One promising strategy for lowering the network's overall energy usage involves breaking it up into smaller sections or "clusters", each of which elects its own leader. This specific leader node, known as a Cluster Head (CH), is a special kind of node responsible for collecting information from the other nodes in the cluster and transmitting it to the sink node. Choosing a cluster-based network architecture has several advantages, including the fact that it guarantees the scalability of network, shortens the distance between the sink node and the source nodes, allows for data fusion and aggregation. The cluster-based WSN architecture used in this thesis is shown in figure 3.1.

Separating the network into distinct clusters is a complex and difficult operation. Therefore, there are many studies in the literature that focus on cluster network-related difficulties, such as choosing a cluster leader, determining the optimal cluster size, and facilitating two-way communication between sensors and CHs and between CHs and the sink. However, our focus in this thesis is not on the cluster formation process, but on the variability in the data acquired by the sensors. Because of this, we propose a geographical clustering system in which nearby sensors are automatically placed in the same cluster.

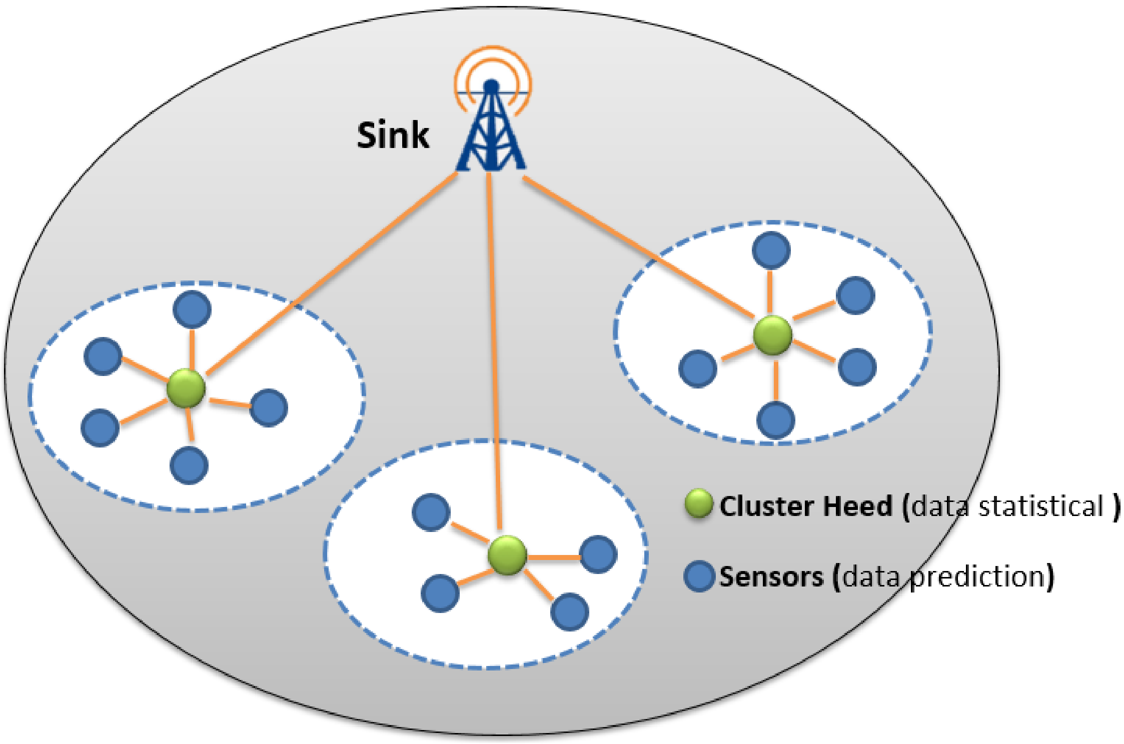


Figure 3.1 The architecture of cluster-based network

3.2.2 Problem Description and Notations

In this scenario, as can be shown in figure 3.1, we take into account a three-tiered design for a cluster-based WSN architecture. Wireless sensor network is demonstrated as a connected graph $G = (SN, E)$, where $SN =$

$\{SN_1, SN_2, \ldots, SN_n\}$ is a collection of n sensor nodes, which represent the first layer, deployed in the region of interest at locations $L_{SN} = \{l_{SN_1}, l_{SN_2}, \ldots, l_{SN_n}\}$, and E is a set of edges. We are assuming a convergent pattern of traffic here (i.e., all traffic flows from the sensors toward the next layer CH or sink). After sensing its environment for some time period (T_p), a sensor node sends its collected data to the next layer of nodes in the network's organizational structure (e.g. CH). CHs represent the second layer of the architecture and are deployed at locations $L_{CH} = \{l_{CH_1}, l_{CH_2}, \ldots, l_{CH_K}\}$. The third layer comprises the sink placed at a distinct location L_{Sink}. Periodic Wireless Sensor Networks (PWSNs) are the name given to the kinds of sensor networks that may enable these kinds of applications. During a sensor node SN_i's lifetime, each time period T_p is separated into discrete intervals called "time slots". Specifically: $T_p = [s_1, s_2, .., s_T]$. Each sensor node SN_i collects a new data value v_{ip_j} at each slot s_j, and this information is combined during the time period T_p to generate a vector of sensed data as follows: $V_{ip} = [v_{ip_1}, v_{ip_2}, .., v_{ip_T}]$.

As an example of data collection and as shown in figure 3.2, each SN_i collects data by taking five measurements at regular intervals T_p ($p \in [1, 5]$) and sending the resulting vector of data $V_{ip} = [v_{ip_1}, v_{ip_2}, v_{ip_3}, v_{ip_4}, v_{ip_5}]$ to the next tier of nodes.

In this thesis, the proposed protocol introduces an ARIMA-based predictive model to iteratively predict the data of SN_i sensor at a time period $(T_p + 1)$, where $T_p + 1 = [s_1, s_2, .., s_T]$ in which T is the times of iterative predictions. The proposed model predicts the sensory vector of data $V'_{ip}(T_p + 1)$ by utilizing the vector of data $V_{ip}(T_p)$. After getting $V'_{ip}(T_p + 1)$, the protocol can iteratively use the predictive model to get $V'_{ip}(T_p + 2)$, $V'_{ip}(T_p + 3)$ and so on.

The pace of change in the environment or the observed phenomenon greatly influences the information gathered by sensor nodes. When the changes that happen in the observed state are slow or the taken slot is small, there is a higher correlation and redundancy in the sensed data. If the monitored situation changes slowly, if the sensing frequency is high, or if the time windows are short, then the resulting data vector V_{ip} formed by the data of node SN_i may include duplicated data (or similar data).

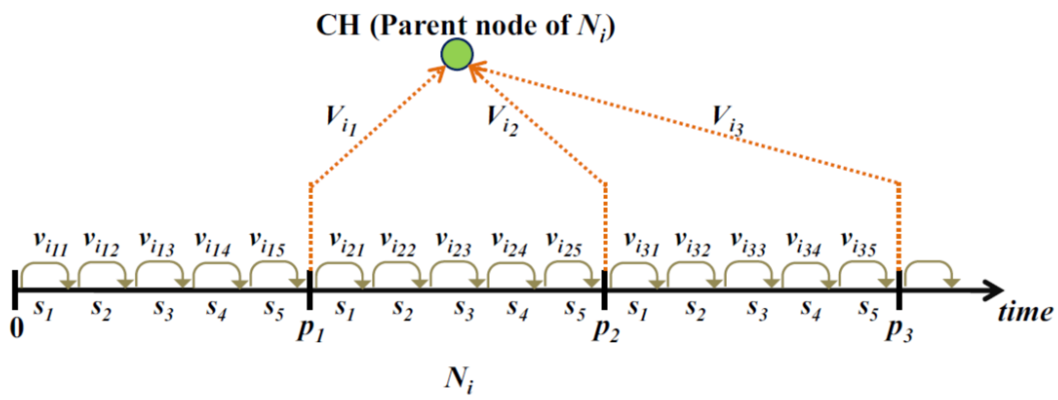


Figure 3.2 Periodic data collection at the cluster head.

3.2.3 Network Assumptions

The proposed protocol makes the following assumptions since the performance of each WSN is affected by unique characteristics unique to that network, such as the detected phenomenon being monitored and the kind of node (mobile or static).

- The topology of the network remains static throughout the network operation.
- Sensor nodes are deployed systematically in a uniform pattern.
- All of the sensor nodes are homogeneous.
- All sensor nodes have limited energy and initialized with an equal amount of energy.
- Unlike the sensor nodes, the cluster head is a super node that has more energy, memory, and processing.

- The sink is designed to be devoid of limitations in computation, energy, and coverage of network.
- Radio interference and physical obstructions that cause signal attenuation are not considered.
- Aggregation of data occurs at CHs.
- Sensing nodes transfer the data to their CHs directly (i.e., single hop).
- Each node senses the environment at a fixed rate and always has data to send to the CH.
- The proposed protocol assumes error-free and reliable communication between the nodes.

3.2.4 Energy Consumption Model

The nodes' energy consumption is calculated as we mentioned previously by using the first-order radio model [103] [104]. Equations 2.13 and 2.14 explain how to compute the consumed energy to send and receive m-bits of data across a network over a d-distance.

The cumulative energy consumed by the CH during data gathering from all cluster members is calculated as follows:

$$E_{Total} = E_{intra\_cluster} + E_{inter\_cluster} \tag{3.1}$$

Where

$$E_{intra\_cluster} = E_{Rx} + E_{DA} + E_{SN} \tag{3.2}$$

The energy expended at the moment of data reception is denoted by E_{Rx}. The energy expended at the moment of aggregation is referred to as E_{DA}. The energy expended by a specific cluster's member nodes is denoted as E_{SN}.

$$E_{SN} = \sum_{j=1}^{K} \sum_{i=1}^{|C_j|} E_{Tx}(x_i, CH_j) \tag{3.3}$$

Where $E_{Tx}(x_i, CH_j)$ indicates the amount of energy used in transmitting data from node x_i to its CH in the j^{th} cluster, $|C_j|$ indicates the cluster C_j's nodes count and $j \in [1,2,\dots,K]$ denotes the number of the clusters.

$$E_{DA} = m \times |C_j| \times E_{sdb} \tag{3.4}$$

where E_{sdb} is the energy of aggregation for a single data bit and m is the bits number.

$$E_{inter\_cluster} = \sum_{i=1}^{K} E_{Tx}(CH_i, BS) \tag{3.5}$$

Furthermore, if CH serves as a members' gateway node and will not create any data by itself, then

$$E_{CH} = (m \times E_{elec} + m \times E_{DA})\left(\frac{N}{K} - 1\right) + (m \times E_{elec} + m \times \varepsilon_{mp} \times d^4) \tag{3.6}$$

N is the nodes count.

In contrast, if the CH participates in sensing and generates data,

$$E_{CH} = \left(m \times E_{elec}\left(\frac{N}{K} - 1\right)\right) + \left(m \times E_{DA}\left(\frac{N}{K}\right)\right) + (m \times E_{elec} + m \times \varepsilon_{mp} \times d^4) \tag{3.7}$$

3.3 The DRDPS Proposed Protocol

In many situations, sending all of the sensed data might be harmful and useless. In order to address problems specific to WSNs, such as transmission collision, energy usage, and redundant data elimination, data reduction is essential. To resolve these problems, a Data Reduction Based on a Dual Prediction Scheme (DRDPS) for data-driven in WSN is presented. In this thesis, we focus on presenting a good data prediction and statistic mechanism that comprises on three main stages to minimize energy consumption and maintain the network lifetime by minimizing excessive data transmission while preserving the accuracy of sensor readings for clustered WSNs.

The first stage (clustering stage) diminishes energy consumption through clustering the network. The clustering stage is further broken into two phases. The first phase is to decide on the best number of clusters to use. The K-Means method is in charge of distributing sensor nodes to the closely relevant clusters depending on the geographic positions of these sensors during the second phase of clustering. The sink performs this step just once at the start of the protocol, and it stays unaltered across the lifetime of the network. To put it another way, once the network topology has been created, no sensor can be relocated from one cluster to the next.

The second stage is a dual data prediction (DDP) scheme for WSN. It is implemented at the SNs and the CHs. At this stage, the SNs are gathering and transmit large amounts of data packets at regular times to a pre-defined node called CH. To conserve energy and decrease the volume of transmitted data, we implemented a DDP model between the SNs and CHs. This model will predict the future data readings of sensors, so that, the SNs will transmit data that differs from the predicted values by a pre-determined threshold, rather than transmitting all collected data.

In the third stage, the CH leverages a statistical-based data model to identify the adjacent SNs that produce similar data periodically. Such statistical representation is used to eliminate redundancy and minimize the need for data transmission, resulting in a significant reduction in energy consumption at the CH node. The proposed statistical model is mainly based on the Granger causality test

3.3.1 The First Stage (Clustering Stage)

In this stage, the proposed K-Means clustering approach for the WSNs' energy conservation issue is presented. The protocol is centralized in which the clustering algorithm in the sink establishes the architecture of a specific

WSN. The sink will deploy the network's CHs in the first round after the sensor nodes have been deployed and each node communicates its position and energy information to the sink. There are three phases in this stage, namely, choosing the optimal number of clusters, cluster formation, and the CHs selection.

A. The Optimal Number of Clusters

Since the small value of K (i.e. number of cluster) means an increase in the intra-cluster communications which impose additional overhead on the cluster head. Determining most suitable number of clusters on which the data can be grouped is a critical stage in any unsupervised method. The K-Means algorithm's performance is determined by selecting the network's optimal number of clusters. One of the really prominent approaches for determining the optimal K value would be the Elbow technique. The optimal cluster's number is discovered using the Elbow technique by computing the within-cluster sum of squared ($WCSS$) distances from the assigned centroids for the given set of nodes as mentioned in equation 2.2.

B. Cluster Formation

A centralized clustering algorithm based on the K-Means approach is suggested in this thesis to divide the network of n sensor nodes, $SN = \{SN_1, SN_2, \dots, SN_n\}$, distributed over an area of $M \times N$ meters, into a stable $K -$fixed optimal number of clusters. Stable clustering aids in lessening the transmission of cluster update messages, which in turn reduces energy usage. Suppose that sink node already has awareness about each sensor node location. The sink computes the cluster centers based sensor nodes locations and assigns sensor nodes SN to the clusters K using the K-Means method.

The sink node connects all CHs and separates the sensor nodes into K clusters as shown in figure 3.1.

The initial division of K clusters, $C = \{C_1, C_2, \dots, C_K\}$, which are denoted by respective centroids, $\mu = \{\mu_1, \mu_2, \dots, \mu_K\}$, is set at random by the K-Means method. The unsupervised method is then used with the objective of minimizing the distance between every sensor and the closest centre.

$$J_{min} = \sum_{j=1}^{K} \sum_{x_i \in C_j} (x_i - \mu_j)^2 \quad (3.8)$$

where x_i indicates the i^{th} node in the cluster C_j and μ_j denotes the physical centre (centroid) of the sensor nodes in the cluster C_j, $j \in [1, 2, \dots, K]$, $|C_j|$ indicates the cluster C_j's nodes count and is provided by:

$$\mu_j = \left(\frac{1}{|C_j|} \sum_{x_i \in C_j} x_i, \frac{1}{|C_j|} \sum_{y_i \in C_j} y_i\right) \quad (3.9)$$

The sizes of clusters in this study are chosen to guarantee that communication between every node and the CH needs just a single-hop. This strategy ensures that every cluster has a greater density and, therefore, all cluster members are near one another. This proximity allows for credible information acquisition in the cluster-monitored region even after the demise of numerous nodes in the responsibility of a different zone. Algorithm 3.1 illustrates the proposed K-Means clustering method.

Algorithm 3.1: *K*-Means Clustering

Input: n: *Sensor nodes No.*
Output: C: *Number of clusters created*

```
for i ← 1 to n do                          /* Network Initialization */
    Deployment sensor nodes
    SN_id ← i                              /* Set sensor (id, initial energy, position) */
    SN_en ← initial energy
    SN_po ← position (x_i, y_i)
```

$\boldsymbol{end\ for}$

$\boldsymbol{Sink_{buffer}} \leftarrow \{\emptyset\}$ **/*** SNs sending their id, energy, and position to the sink ***/**

$\boldsymbol{for}\ i \leftarrow 1\ to\ n\ \boldsymbol{do}$

$\quad \boldsymbol{Sink_{buffer}}[\boldsymbol{i}] \leftarrow \boldsymbol{SN_i}[\boldsymbol{id}, \boldsymbol{en}, \boldsymbol{po}]$

$\boldsymbol{end\ for}$

$\boldsymbol{K} \leftarrow Calculating\ optimal\ No.of\ clusters\ using\ Eq.(2.2)$

$\boldsymbol{t} = 0$

$\quad \boldsymbol{\mu} \leftarrow reandomly\ initialize\ \boldsymbol{K}\ centroids\ from\ \boldsymbol{Sink_{buffer}}: \{\boldsymbol{\mu_1^t}, \dots, \boldsymbol{\mu_K^t}\}$

$\quad C_j \leftarrow \emptyset$ for all $j = 1, \dots, K$

$\boldsymbol{repeat}$

$\quad \boldsymbol{foreach}\ (x_i \in \boldsymbol{Sink_{buffer}})\ \boldsymbol{do}$

$\quad\quad \boldsymbol{foreach}\ (\mu_j^t \in \boldsymbol{\mu})\ \boldsymbol{do}$

$\quad\quad\quad j \leftarrow argmin\ \left\|x_i - \mu_j^t\right\|^2$ /* assign x_i to closest centroid */

$\quad\quad\quad C_j \leftarrow C_j \cup \{x_i\}$

$\quad\quad \boldsymbol{end\ for}$

$\quad \boldsymbol{end\ for}$

$\quad t \leftarrow t + 1$

$\quad \boldsymbol{for}\ \ j \leftarrow 1\ to\ K\ \boldsymbol{do}$

$\quad\quad \mu_j^t \leftarrow \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$ /* centroid update (average) */

$\quad \boldsymbol{end\ for}$

$\boldsymbol{until}\ \left\| \mu_j^t - \mu_j^{t-1} \right\|^{\boldsymbol{2}} = \boldsymbol{0}$ /* centroid do not change */

$return\ \boldsymbol{C}$

C. Cluster Head Selection

The CH location is chosen as the centroid of the cluster. The process of changing the CH periodically is outside the scope of this study. Also, the CH is not chosen from among the sensing nodes, but rather is a node whose specifications are higher than the specifications of the nodes used in the process of sensing the surrounding environment.

The sink sends the clustering findings straight to every individual node. In reality, due to the kinds of nodes, if the node is a member node, the sink provides the CH's id of the cluster i to which the node belongs directly. Alternatively, in the case when the node is a CH node, the sink notification will be sent directly to the CH. After that, the cluster members transmit to the CH a "*Membership request*" to join this cluster. figure 3.2 depicts the centralized procedure of clustering the network.

3.3.2 The Second Stage (Dual Prediction Stage)

Sensors in WSN applications may gather information on environmental conditions (such as noise pollution, temperature, pressure, relative humidity, etc.). Because the data produced by sensor nodes during periods of continuous monitoring are often highly correlated in time, This suggests that the next sequence contains unnecessary (redundant) data, which is causing additional data transmissions. A major obstacle, however, is the massive amount of data that these sensors generate. In addition, transmitting data in a WSN uses up a lot of resources, including sensor energy.

Here we describe the second stage of the proposed protocol, which uses the ARIMA model to predict data for the subsequent few times at sensor nodes and CHs depending on the same last historical data. Based on the prediction model, this stage seeks to avoid the transmission of identical or similar data points observed by each sensor at each time period T_p. Through coordinated efforts between sensor nodes and CHs, the network is able to send fewer data packets.

1. The ARIMA Model

In order to predict future data values, time series analysis builds a model from the available historic or previous data. When making predictions about univariate time series, the ARIMA model is often used. As was discussed in section 2.6.3, the ARIMA model consists of three sub-processes: "auto-regressive (AR), moving average (MA), and one-step differencing". figure 3.3 depicts the numerous steps used by the suggested protocol in this thesis to generate the automatic ARIMA model. Here are the four steps that make up this method:

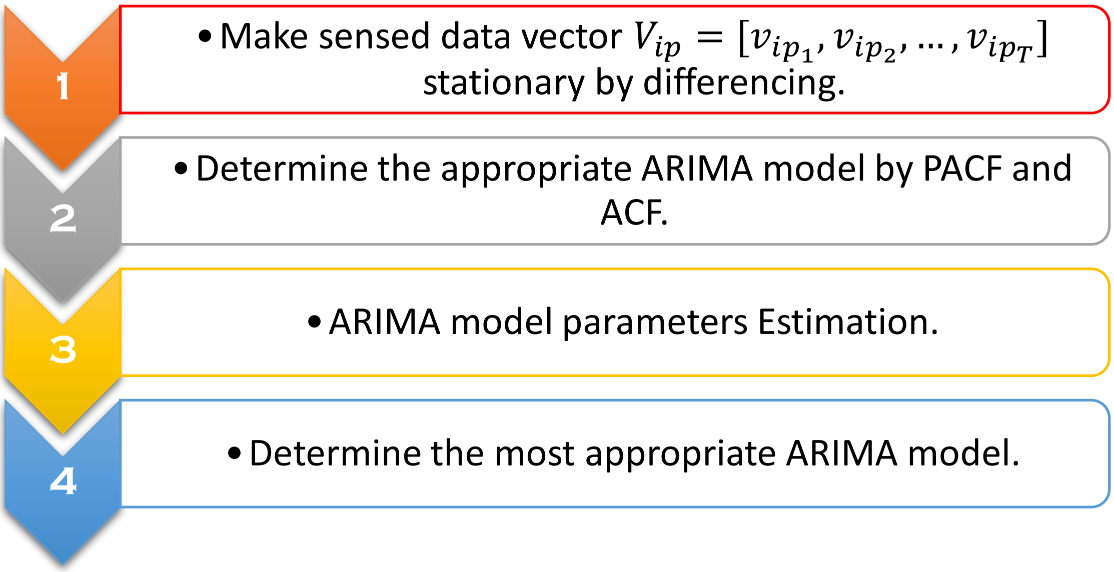


Figure 3.3 The automatic ARIMA model construction steps.

1. Step 1

For this analysis to be valid, the sensed data vector V_{ip} must be stationary. Data transformation performed by differentiate data source. This is required when the sensed data vector variance is non-stationary and stationary results are desired. In some cases we need to perform many differences on the series until the sensed data vector becomes stationary if it displays a trend over time, seasonality, or any other non-stationary pattern.

2. **Step 2**

Once the sensed data vector V_{ip} has become stationary, potential ARIMA models is determined. Autocorrelation and partial autocorrelation functions (ACF and PACF) allow for the identification of different ARIMA models that provide a good fit to the sensed data vector V_{ip}.

3. **Step 3**

After settling on an ARIMA model to fit the sensed data vector, we conduct a time series analysis and parameter estimation. The protocol might think about including one or more AR terms in the model if the PACF of the differenced sensed data vector exhibits an abrupt cutoff and the lag-1 autocorrelation is positive. The PACF ends after the stated number of AR terms, which is the maximum delay before it begins to operate. Incorporate an MA term into the model if the ACF of the difference-sensed data vector exhibits an abrupt cutoff and the lag-1 autocorrelation is negative. It is the given number of MA periods that represents the maximum delay before the ACF stops collecting data.

4. **Step 4**

The best ARIMA model for analysis is the one with the least indicator value for either the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) equation 2.7 and 2.8.

The sensor node must gather recently sensed data vector $V_{ip} = [v_{ip_1}, v_{ip_2}, \dots, v_{ip_T}]$ in order to construct the ARIMA prediction model. If the sensed data vector V_{ip} isn't stationary, we need to apply the differencing adjustment until the variance with the succeeding data vector is less than the stationary threshold ε provided by the application. Later, fit an ARIMA prediction model to the differentiated data vector $[v'_{ip_1}, v'_{ip_2}, \dots, v'_{ip_T}]$ using least square method. As illustrated in figure 3.4, the ARIMA model fitting procedure uses a box search path for its iterations. In a limited number of iterations, it may discover a model that fits the data well enough to be used. In ARIMA models, the fitting process will terminate when the BIC indicator is less than the BIC threshold set by the application. An adequate ARIMA model for making forecasts has thus been developed. Because of its greater consistency and harsher punishment of free parameters, the BIC indicator is preferred in this thesis over the alternative AIC indicator.

Algorithm 3.2 illustrates how the ARIMA modeling algorithm is executed on the sensor node and CH to automatically construct an ARIMA prediction model.

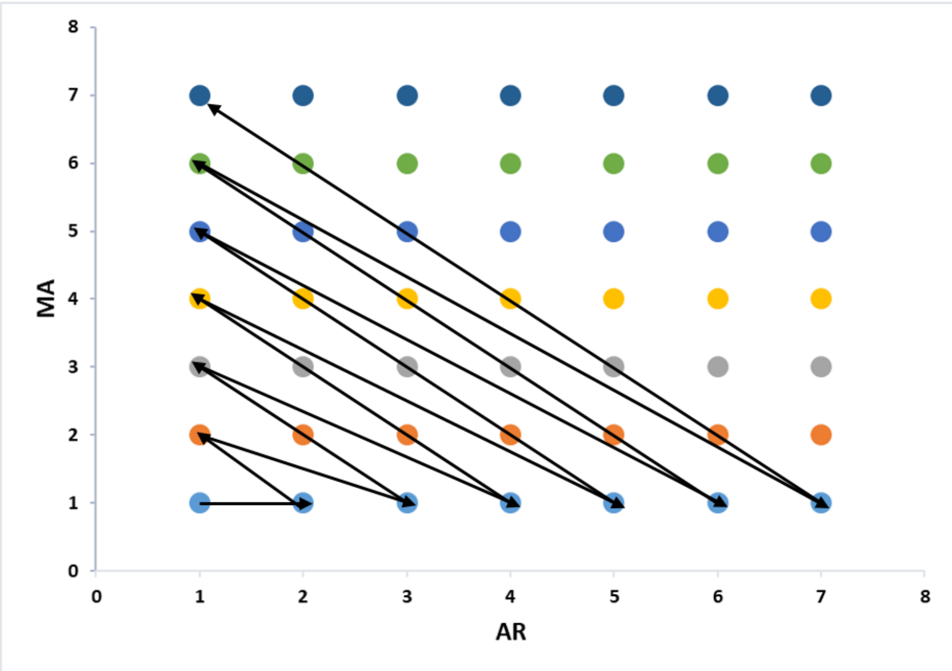


Figure 3.4 Box search path.

Algorithm 3.2: $\boldsymbol{Automatic\ ARIMA\ modeling}$

Input: $Sensed\ data\ vector\ V_{ip} = [v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}]$,
ε: $application\ defined\ stationary\ threshold$,
δ: $Application\ defined\ BIC\ indicator\ threshold$

Output: $ARIMA\ model\ Parameters\ (ARIMA)$

1 $\boldsymbol{for}\ i \leftarrow 1\ to\ T\ \boldsymbol{do}$

$Collect\ recently\ sensed\ data\ vector\quad V_{ip} = [v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}]$
$\quad I \leftarrow 0$
$\boldsymbol{end\ for}$
$\boldsymbol{while}\ \left|variance\left(diff\left([v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}], I\right)\right) - variance\left(diff\left([v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}], I+1\right)\right)\right| > \varepsilon\ \boldsymbol{do}$
$\quad I \leftarrow I + 1$
$\boldsymbol{end\ while}$
Make $[v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}]$ stationary by I order differencing and get $[v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}]$
$\boldsymbol{for}\ AR \leftarrow 1\ to\ MaxAR\ \boldsymbol{do}$
$\quad \boldsymbol{for}\ MA \leftarrow 1\ to\ AR\ \boldsymbol{do}$
$\quad\quad fit\ ARIMA(AR, 0, MA)\ model\ using\ least\ square\ method$
$\quad\quad claculate\ BIC\ indicator\ for\ ARIMA$
$\quad\quad \boldsymbol{if}\ (BIC < \delta)\ \boldsymbol{then}$
$\quad\quad\quad \boldsymbol{break}\ ARIMA\ modeling$
$\quad\quad \boldsymbol{end\ if}$
$\quad \boldsymbol{end\ for}$
$\boldsymbol{end\ for}$
$\boldsymbol{return\ AR, MA}$

5. Data Reduction Based on Dual Prediction Scheme

The goal of the data reduction based on a dual prediction scheme (DRDPS) is to reduce the amount of data transfers between sensor nodes and CH/sink nodes. As previously stated, each sensor node collects T data readings in a specific time period T_p before sending the constituted sensed data vector, $V_{ip} = [v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}]$, to the CH. There is a storage buffer for previously sensed data vector at each sensor node. CH also maintains 'c' buffers, one for each cluster member. When the first period is finished, the CH saves the following matrix $buffer_{CH}$ in its memory.

$$buffer_{CH} = \begin{bmatrix} v_{1p_1}, v_{1p_2}, \cdots, & v_{1p_T} \\ v_{2p_1}, v_{2p_2}, \cdots, & v_{2p_T} \\ \vdots, \quad \vdots, \quad \ddots, & \vdots \\ v_{cp_1}, v_{np_2}, \cdots, & v_{cp_T} \end{bmatrix}$$

Where c represents the number of SN in a cluster and v_{ip_T} represents the sensed data vector of i^{th} sensor node.

Meanwhile, the data vector V_{ip} from the sensor is used to start the same prediction model in both the sensor node and its CH. The first step in developing an accurate ARIMA prediction model is for the sensor node and CH to execute an automated ARIMA modeling algorithm. During the phase of DPS implementation, data prediction is carried out simultaneously in the CH and sensor node. The predicted data made by the sensor node is the same as that made by the CH. However, the sensor node can check the accuracy of the predictions by comparing the data vector V_{ip} they have collected with the data vector V'_{ip} they have predicted. The sensor node will add the predicted data vector V'_{ip} to the buffered data queue if the difference between the two is smaller than the prespecified threshold. If that isn't the case, it will add the sensed data vector V_{ip} to the buffered data queue and transmit it to the CH simultaneously. The behaviour of the sensor node is illustrated in equation 3.10.

$$SN_i = \begin{cases} Transmitted\left(V_{ip}\right) & if\ \left|V_{ip} - V'_{ip}\right| > \varepsilon \\ Discarded\left(V_{ip}\right) & Otherwise \end{cases} \tag{3.10}$$

If the difference between the predicted and sensed data is greater than the fault-tolerable range, the AIRMA model will be reconstructed, and the relevant ARIMA model parameters for the CH will be updated.

Consequently, sensor nodes may save energy by not sending sensed data to the CHs until the forecasts are not precise enough.

In order to collect data from each sensor node, the CH must monitor its wireless link. If the CH hasn't received any data from a sensor node after a certain amount of time, then the difference between the sensed data and the predicted data is within the allowable error range. The ARIMA model will then be used to predict the data, and the CH will use past data to make this prediction. Unless the condition is fulfilled, the sensed data is added to the buffered data queue, and preparations are made to update the ARIMA model parameters.

Detailed interaction of DRDPS is seen in figure 3.5. The quantity of data sent between the sensor node and the CH is reduced owing to their coordinated efforts.

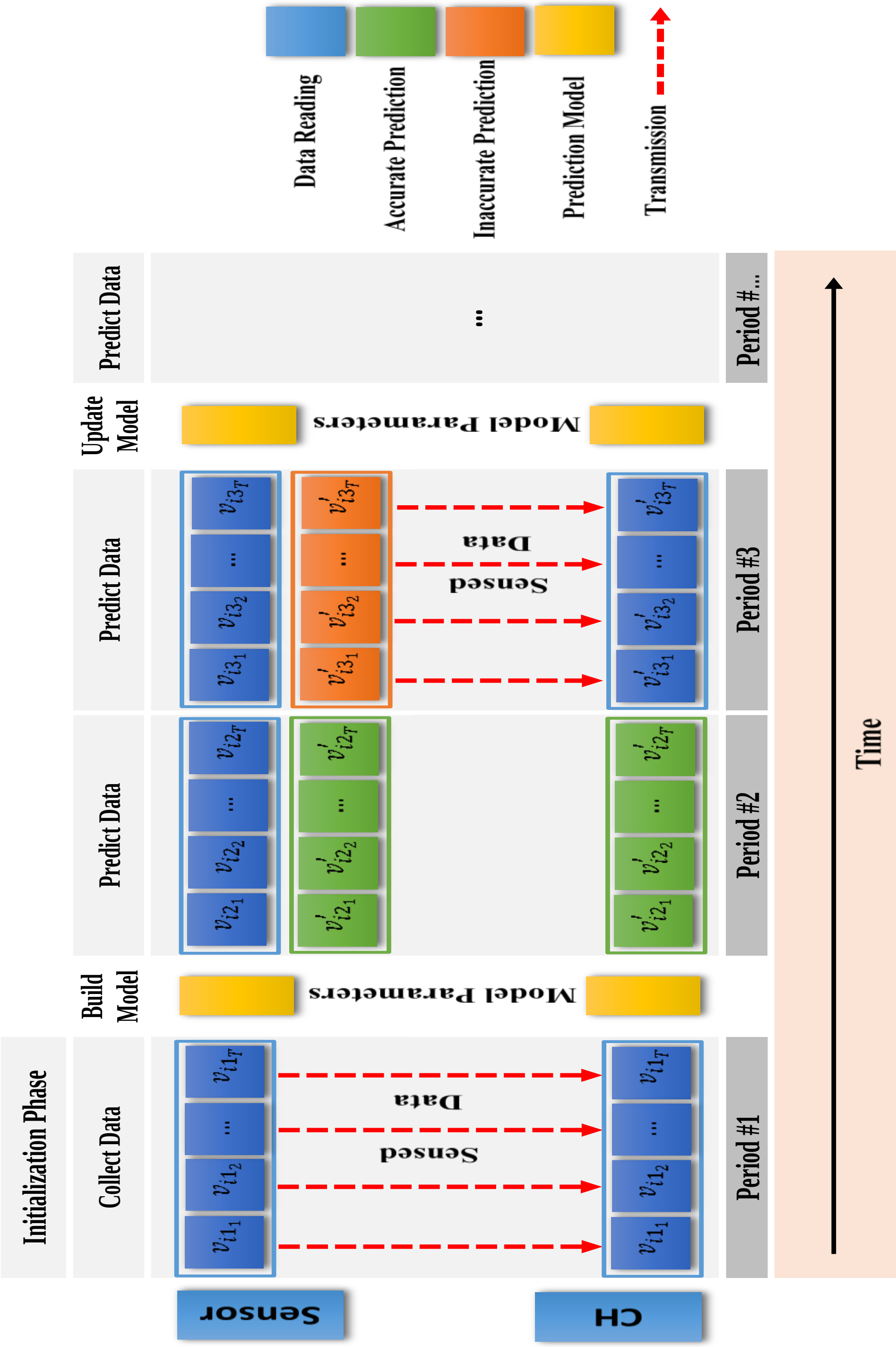


Figure 3.5 Process of implementing DPS between a sensor node and CH.

Algorithm 3.3: $\boldsymbol{The\ Algorithm\ at\ the\ SN\ Level}$

Input: $Sensed\ data\ vector\ V_{ip} = \left[v_{ip_1}, v_{ip_2}, \dots,\ v_{ip_T}\right]$,
ε: $application\ defined\ prediction\ threshold$

Output: Collect recently sensed data vector V_{ip}

$buffer = \emptyset$ // buffer data queue

$V_{ip} = V'_{ip} = \emptyset$ // Sensed and predicted data vector

$\boldsymbol{foreach}\ period\ \boldsymbol{do}$

    $\boldsymbol{for}\ j \leftarrow 1\ to\ T\ \boldsymbol{do}$

        $V_{ip} \leftarrow v_{ip_j}$ // Collect recently sensed data vector

    $\boldsymbol{end\ for}$

    $\boldsymbol{if}\ (period = 1)\ \boldsymbol{then}$

        $buffer \leftarrow V_{ip}$

        $ARIMA_{model} \leftarrow Run\ automatic\ ARIMA\ using\ \boldsymbol{buffer}$

        $CH \leftarrow Transmitted\ (V_{ip})$ //Send sensed data vector to CH

    $\boldsymbol{else}$

        $V'_{ip} \leftarrow ARIMA_{model}(buffer)$ //Predict data vector using ARIMA

        $\boldsymbol{if}\ \left(RMSE(V_{ip}, V'_{ip}) < \varepsilon\right)\ \boldsymbol{then}$

            $buffer \leftarrow V'_{ip}$

            $Discarded\ (V_{ip})$

        $\boldsymbol{else}$

            $buffer \leftarrow V_{ip}$

            $CH \leftarrow Transmitted\ (V_{ip})$ //Send sensed data vector to CH

            $ARIMA_{model} \leftarrow Run\ automatic\ ARIMA\ using\ \boldsymbol{buffer}$

        $\boldsymbol{end\ if}$

    $\boldsymbol{end\ if}$

$\boldsymbol{end\ for}$

$\boldsymbol{return}\ V_{ip}$

Algorithm 3.4: Cluster head behavior (dual prediction)

Input: $Sensed\ data\ vector\ V_{ip},$
$\varepsilon: application\ defined\ prediction\ threshold$

Output: $CH\ buffer\ as\ a\ vector\ of\ vectors$

$buffer_{CH} \leftarrow [[\emptyset]]$ // CH buffer as matrix

$CHV'_{ip} = \emptyset$ // Predicted sensed data vector

$\boldsymbol{foreach}\ period\ \boldsymbol{do}$

$wait\ periodical\ data\ collection\ time$

$\quad \boldsymbol{if}\ (received\ \boldsymbol{V_{ip}}\ from\ sensor\ members)\ \boldsymbol{then}$

$\quad\quad buffer_{CH}[\boldsymbol{i}] \leftarrow V_{ip}$

$\quad\quad ARIMA_{model} \leftarrow Run\ automatic\ ARIMA\ using\ \boldsymbol{buffer_{CH}}$

$\quad \boldsymbol{else}$

$\quad\quad CHV'_{ip} \leftarrow ARIMA_{model}(buffer_{CH})$ //Predict data vector

$\quad\quad buffer_{CH}[\boldsymbol{i}] \leftarrow CHV'_{ip}$

$\quad \boldsymbol{end\ if}$

$\boldsymbol{end\ for}$

3.3.3 The Third Stage (CH Data Reduction)

Our goal at this stage is to lessen the load on the CH and BS by allowing CH to send fewer data sets. The CH perform long-distance communication with the BS. Furthermore, it is responsible of receiving and aggregating data from all SNs in the cluster and this operation would add more overload on the CH energy source. Therefore, it is important to reduce the amount of communication between CHs and BS.

The statistical analysis technique is an essential foundation with the goal of discovering the correlation between data sets to lessen the need for both over-transmission and energy usage in the CH. Statistical analysis is used because it allows the CH to compile data sets from several SNs and send just the most relevant sets of data to the BS. A simple and efficient statistical method is proposed to implemented here called Granger Causality (GC). Each CH works by clustering data, to merge the sensor nodes' data sets for each cluster into representative data sets. GC is used to identify which SNs are driving the similarity between the data sets in each cluster. GC test was performed between each pair of data sets in the cluster and using the results to build a causal array. The causal array can then be used to identify the SNs that are driving the similarity between the data points in each cluster.

	S1	**S2**	**.**	**.**	**Sn**
S1	*0*	*GC(D1, D2)*	.	.	*GC(/D1, Dn)*
S2	*GC(D2, D1)*	*0*	.	.	*GC(D2, Dn)*
.	.	.	*0*		.
.	.	.	.	*0*	.
Sn	*GC(Dn, D1)*	*GC(Dn, D2)*	.	.	*0*

Figure 3.6 The causality matrix

Figure 3.6 illustrate the causality matrix, where $GC(S_i,S_j)$ represents the causality relation between data sets D_i,D_j of two sensor nodes S_i and S_j.

Algorithm 3.5: Cluster head behavior (statistical test)
Input: *Data sets received by CH from cluster members*

Output: *Data sets that CH should send to BS*

```
Sending set = Data set [0]
for  i = 1 to length of Data set
      Flag = False
      j = 0
      While (Flag = False and  j  <  length of Sending set) Do
            P-value = GC ( Data set [i], Sending set [j]
            if  p-value < threshold
                  Flag = True
            end if
      end while
      if  Flag = False
            Sending  set . append ( Data set [i] )
      end if
end for
send Sending set to BS
```

Summary

In this thesis, the WSN network is divided into clusters. The proposed clusters-based WSN architecture comprises three layers: the SN layer, the CH layer, and the sink layer. The first-order radio model is used to compute the energy consumption of the proposed method. The data reduction based on dual prediction scheme (DRDPS) is proposed to achieve efficient energy consumption in such networks. The proposed protocol includes three stages. In the first stage, a centralized clustering algorithm such as K-means is employed to make a geographical clustering process based on the position of each sensor node. This algorithm is applied by the sink node. The Elbow method is used to define the optimal number of clusters. The clusters centers are set to place the CH of each cluster. The second stage includes the dual data prediction scheme DDS where both sensor nodes and the cluster heads make a simultaneous prediction to produce the future values based ARIMA method. ARIMA model consists of three sub-processes: AR, MA, and I. The third stage includes employing a statistically based data model called Granger causality. It is used to identify similar data sets and extract the most relevant data sets to send them to the base station. Each cluster head groups the data sets that have similar values and select representative data sets sent to the base station node.

Chapter Four

The Experimental Results and Discussions

CHAPTER FOUR

THE EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter presents the simulation results of the proposed energy conservation method of WSNs and compared the results with common methods. The results are represented graphically in this chapter. The energy consumption, prediction accuracy, and size of data readings are the main effective metrics of the performance of the purposed protocol.

4.2 Simulation Setup

In this thesis, the network is composed of 54 Mica2Dot sensors that use a tiny DB in-network query processing system. It was deployed in the Intel Berkeley research lab. Various SNs deployed systematically and the sink node is placed at the center of the sensing area. The network is set up according to the diagram illustrated in figure 4.1. The weather data comprise four different parameters namely temperature, humidity, light intensity and voltage. The sensor nodes use a log file in our simulations that contains 2.3 million readings. In this thesis, we will deal and simulate the temperature values for 47 sensors only because of incomplete data of seven sensor nodes that mark in yellow color in figure 4.1. The proposed method is implemented in every sensor node. Python 3.7 is used to perform the calculations.

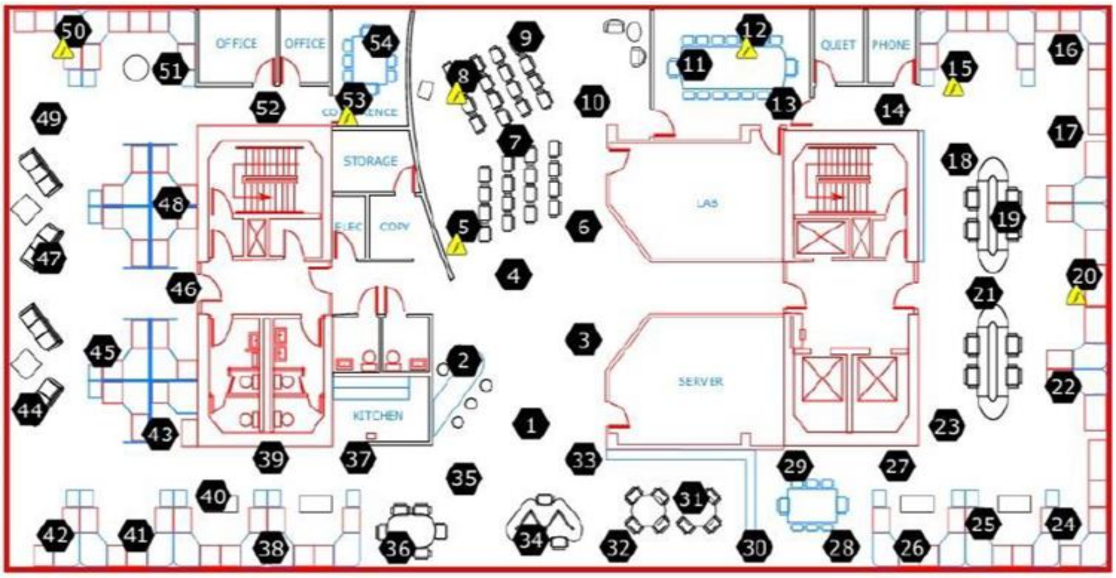


Figure 4.1 Deployment of Sensors in Intel Berkeley Lab [25].

The sensors collect the temperature in periodic intervals called "time slots"; where each slot is 31 seconds. The period length depends on the application specification. In this study, each period can be (20, 50, or 100 slots). The simulation is running with the parameters declared in table 4.1.

Table 4.1. Network parameters

Parameter	**Value**
Number of nodes (network size)	47
Base station location	$Center$
Area size	$45\ m \times 35\ m$
Number of clusters	4
Initial energy	$0.1\ Joule$
Readings size	$20, 50, 100$
$\boldsymbol{Energy_{elec}}$	$50\ nJ/\ bit$

$\varepsilon_{free-space}$	$10\ pJ/\ bit/\ m^2$
Data readings	$1\ reading\ /31\ sec$
Base station location	$(22.5, 17.5)$
Prediction error threshold	$0.1 - 0.9$

4.3 Performance Metrics

The performance of the proposed techniques in the experimental simulations is usually judged upon the following aspects: *Energy consumption, network lifetime, prediction accuracy, amount of data transmission, and data reduction rate.* These metrics are used in the first level while in the second level, the following metrics are used: *Ratio of data sets sent to the sink, and energy consumption at CH.* Secondly, we compared the proposed method with some competitive methods belonging to the same field.

4.4 System Evaluation

This section shows the result analysis of the proposed system. The results of this study will be at three stages, network configuration, the sensor node level, and the cluster head level.

4.4.1 Optimal Number of Clusters

The most significant requirement of the K-means algorithm is a predetermined number of clusters. To determine the ideal clusters number, The Elbow method was used. Elbow method works to make with-in cluster sum square (WSS) values as minimum as possible. Figure 4.2 shows that the increment in *K* leads to a respective decrement in *WSS* value. By looking at the curve, the angle is formed at the point equal to 4. It is the point at which the WSS starts to level off. Therefore, the network can be divided perfectly into four clusters. Figure 4.3 shows the network after clustering.

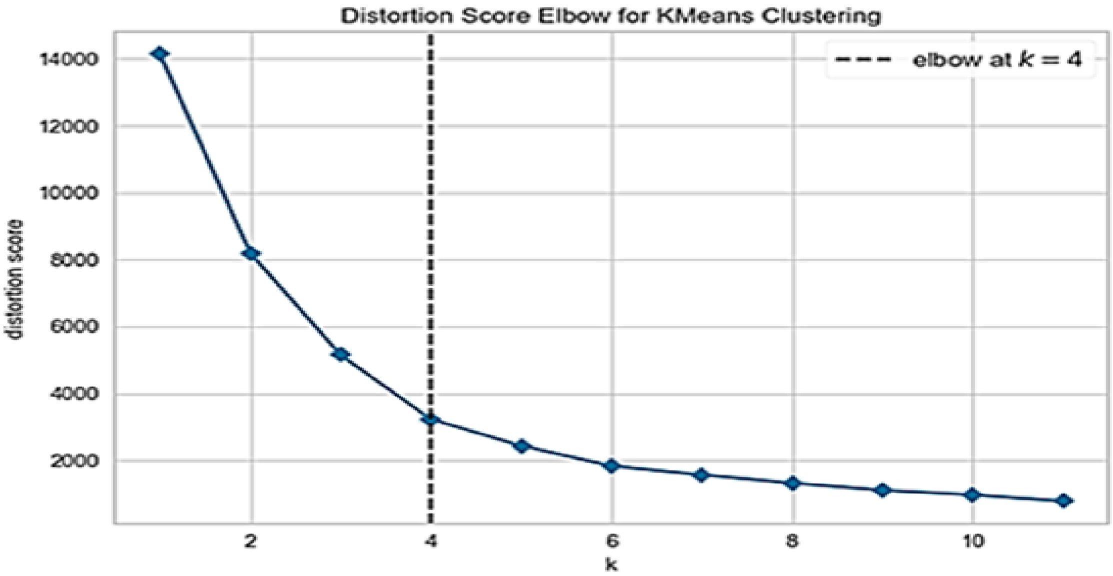


Figure 4.2 The optimal number of clusters by using Elbow method.

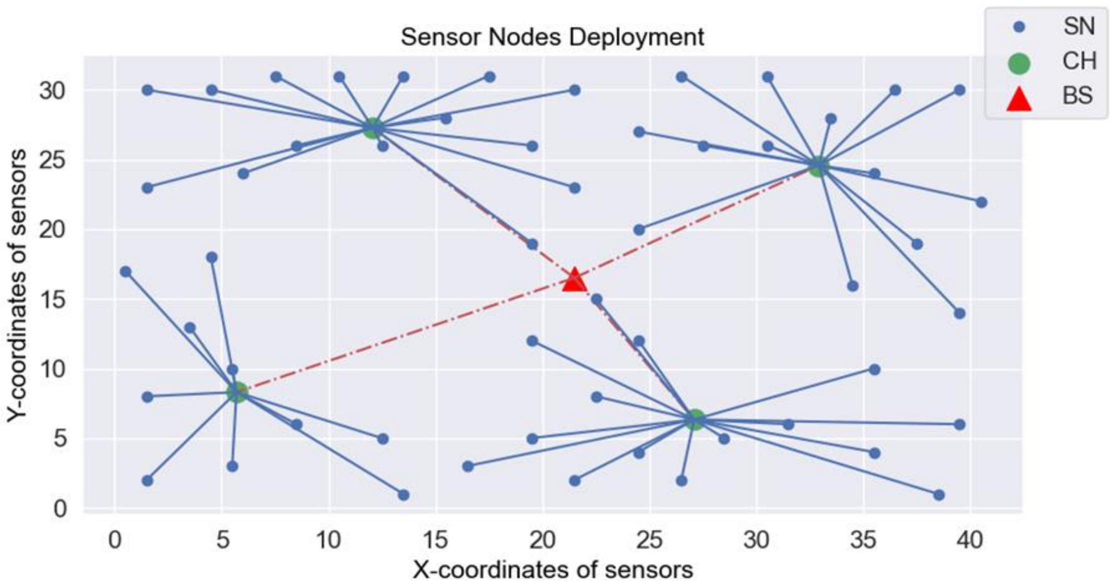


Figure 4.3 The construction of the network after clustering.

The number of nodes in each cluster is different for the different clusters. K-means algorithm joins the nodes into one or different clusters depending on the distance between the nodes. Figure 4.4 shows a comparison of the number of sensor nodes in each cluster depending on different algorithms (LEACH [14], [18], ECH[18], K-means).

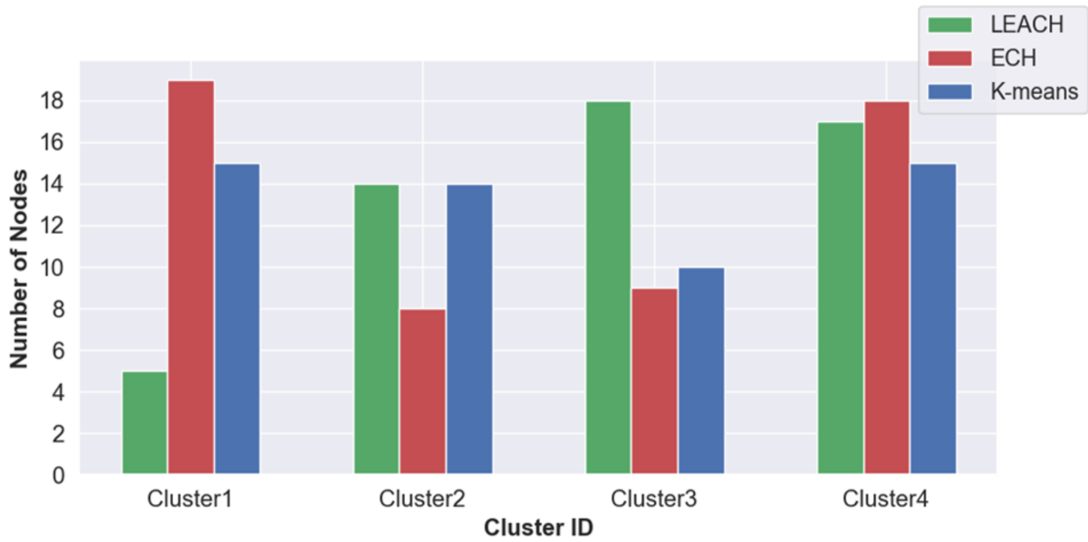


Figure 4.4 Number of nodes in clusters by LEACH, ECH, and K-means algorithms.

For the Intel network distribution, in the LEACH approach, the number of nodes varies from 5 to 18 and for the ECH approach ranging from 8 to 19. While for our proposed approach, the number of sensor nodes varies from low as 10 to high as 15 and is also very near to the average of sensor nodes. The k-means algorithm approximately achieves a fair distribution of nodes between the clusters, this provides a more accurate representation of the data since each cluster is made up of approximately similar data readings. Especially when figuring out the representative data sets at the CH.

4.4.2 Performance of Sensor Node Level

This section shows of the analysis and discussion of the simulation results when the prediction algorithm applies at the sensor node level. The main objective of these experiments is to demonstrate whether our methods successfully conserve the energy of WSN at this level.

A. Energy Consumption

One of the important metrics to evaluate the efficiency of WSN is energy consumption. Figure 4.5 shows a comparison in terms of energy consumption between the flat architecture and the clustered architecture using different readings sizes. This experiment illustrates the effect of the clustering process on the energy consumption of each sensor node.

Figure 4.5 Energy consumption of flat vs. clustered network.

There are many parameters that have a significant effect on energy consumption. The first parameter is the distance over which data are transmitted. From the results obtained, notice that the energy consumption for each sensor node in the clustered network decreased compared with the energy spent in the flat network, we notice that the energy expenditure is 40-50% less than in the clustered network, due to the fact that distance travelled between the SNs and the CHs in the clustered network is much shorter than the distance in the flat network between the SNs and the BS.

The second parameter is the amount of data readings that send from the SN to the CH/BS. It has a clear effect on battery energy depletion. Figure

4.5 shows this effect on the sensor's node energy for both the flat network and the clustered network.

The amount of energy consumed by each SN is proportional to the amount of data it transmits. In other words, more data transmits, means more energy consumed. We notice that the energy consumption of each sensor node in the network decreases by 20% - 50%, according to the amount of data that is transmitted over the network. Where energy spent when sending data with a size of 100 readings is 50% higher than sending it with a size of 20 readings.

When the prediction method (DRDPS) is used the energy consumption of the network will be as illustrated in figure 4.6.

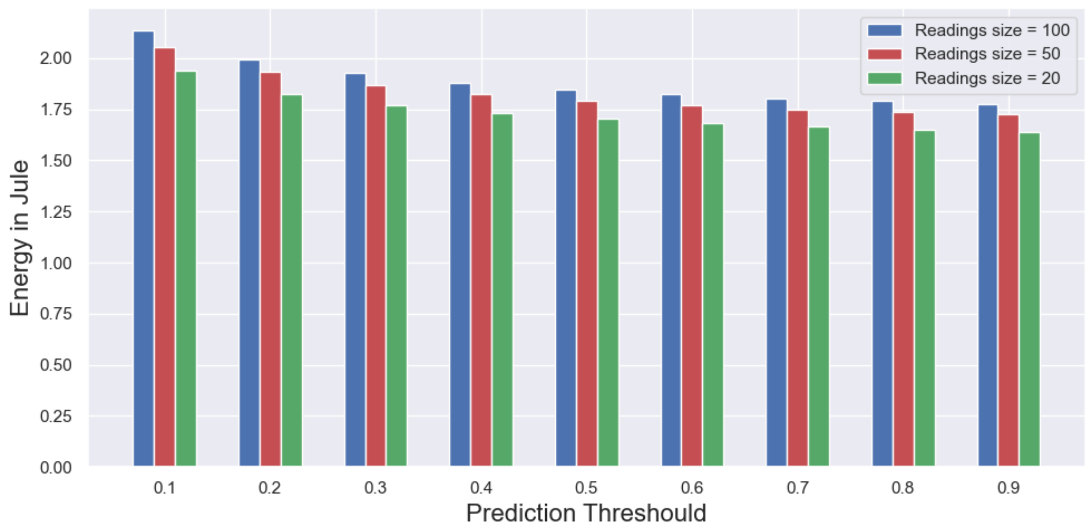


Figure 4.6 Energy consumption rate of the network with DRDPS.

There is an extreme relationship between energy consumption and prediction threshold. From figure 4.6, it is obvious when the prediction threshold is equal to 0.9, the energy saving increase to 13% in comparison with 0.1. that’s because a low prediction threshold means more data transmission between the CH and the SN thus more energy consumption. Therefore, it is possible to increase the error rate within the acceptable range to conserve more energy. On the other hand, the energy expenditure of the network when the data readings size is 100 increases by 10% in comparison with the case in which data readings are 20.

Figure 4.7 displays the energy consumption of the network in different scenarios. In the mentioned figure, the "blue bar" and the "red bar" represent the energy consumed to send the data readings continuously over the flat network and clustered network respectively. While the "green" bar represents the energy consumption of the network when DRDPS is employed, where data readings send in necessary cases only when the prediction model fails to predict data correctly.

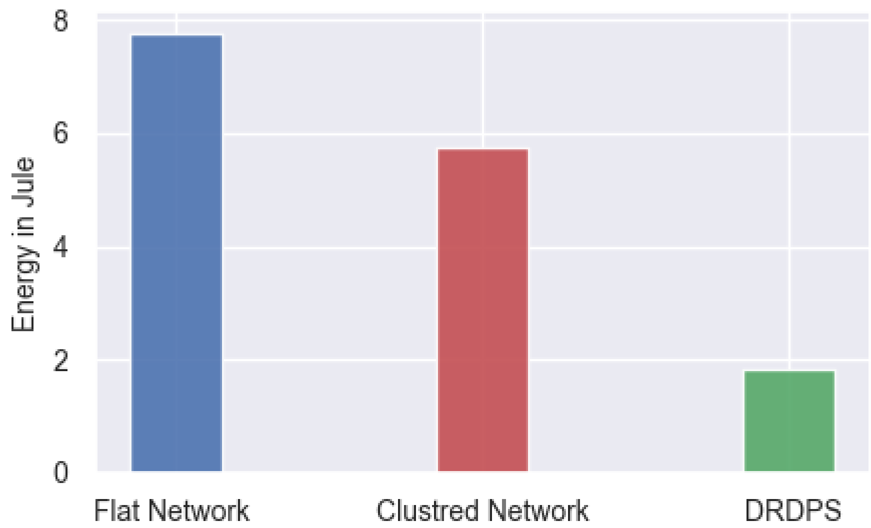


Figure 4.7 Energy consumption of different scenarios

DRDPS reduces the amount of energy expended by 75% compared with the energy consumption in the flat network.

B. Rate of Transmitted Data from SN to CH

In this experiment, the amount of data transmission after employing DRDPS is investigated. Figure 4.8 show the transmission ratio of the network.

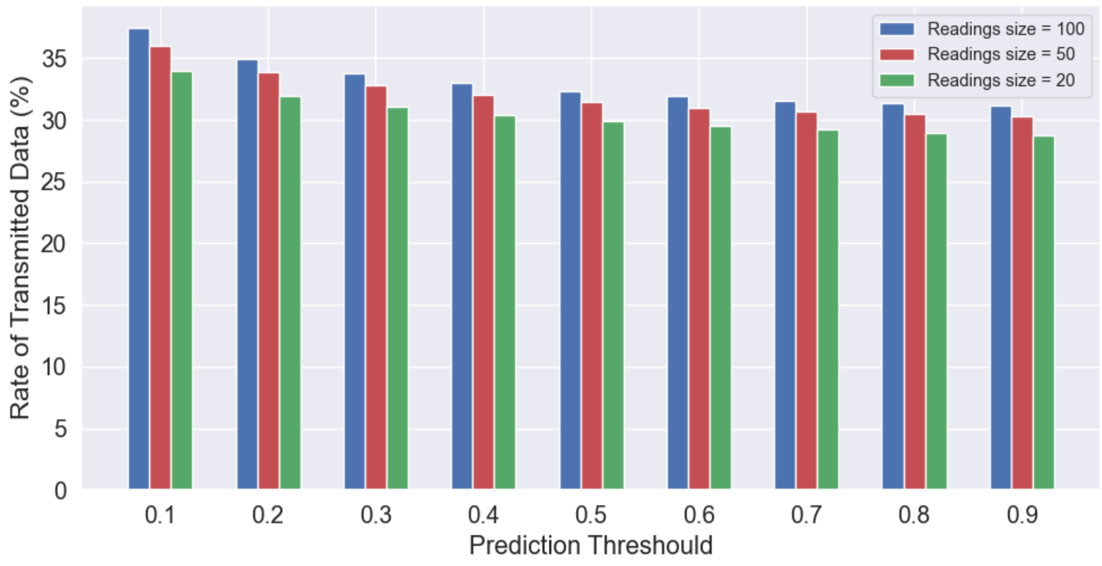


Figure 4.8 Rate of transmitted data.

In figure 4.8, for each value of prediction error in the range 0.1- 0.9, three data readings size are represented. The obtained results explain the relationship between the amount of transmitted data and the prediction error. When the prediction error has a low value this imposes more transmissions because of the sensitivity to the error in the prediction.

Moreover, we can note that there is a relationship between the data readings size and the number of transmissions, as the data readings size increases, the data transmission rate increases. That's due to the increase in the length of the interval in which data are collected, which leads to more variance in data, thus the prediction model is unable to make correct predictions. When the size of the data readings is 100,50,20, the amount of transmitted data decreased by 69%, 70%, and 72% respectively at a prediction threshold of 0.9.

The prediction algorithm can work more accurately with the low-variance data collected at close intervals. Generally, The rate of transmitted data is less for the high threshold of prediction and small size of data readings.

C. Rate of Data Transmissions Reduction

For each SN in a particular pe\riod, the collected data is either send or not. If the predicted readings are within the prediction threshold with the actual data then it is considered a correct prediction. The percentage of cases in which no need to transmit data to the CH is calculated. Figure 4.9 shows the percentage of data transmission reduction of the proposed method. The low prediction threshold value means an increase in transmissions. When the data readings size are equal to 100, 50, and 20 and the prediction threshold is 0.1, the rate of transmitted data is 37%, 36%, and 35%, while when the prediction threshold is 0.9 the rate of transmitted data decreased into 31%, 30%, 27% respectively. The increase in the prediction threshold means more of an inaccurate value is considered at the CH.

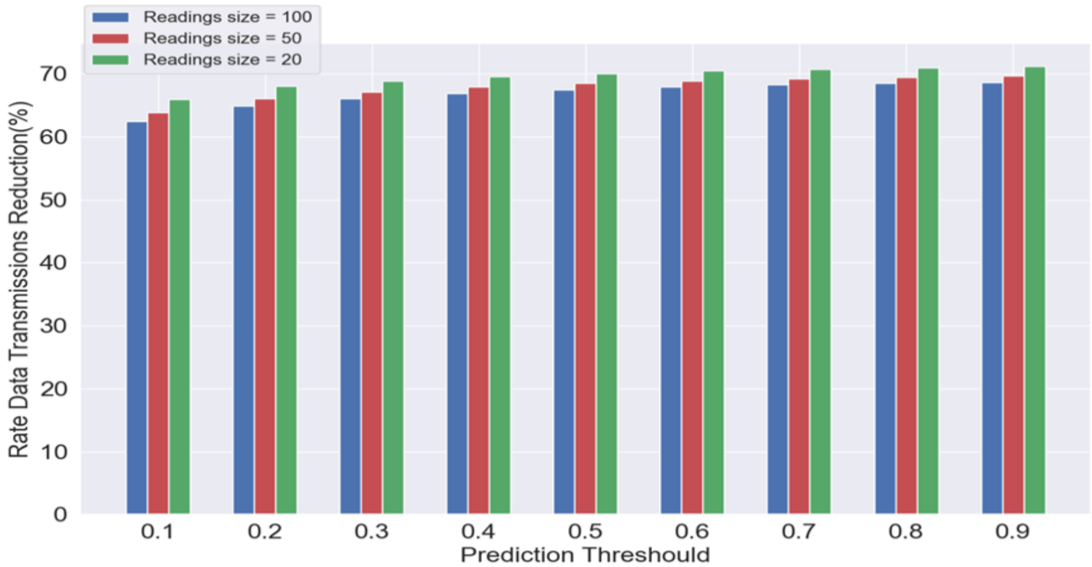


Figure 4.9 The rate of data transmission reduction.

D. Network Lifetime

A crucial factor that can accurately depict network performance is the number of alive nodes at the end of each period. From figure 4.10 it is easy to observe the relation between the period number and the amount of alive nodes. The gained results indicate that clustering can increase the node's lifetime by 40%. Most of nodes in a clustered network die after having lived approximately twice in comparison to the network was flat. This is subject to the condition of the distance between the node and the cluster head, as some nodes close to the sink node will not benefit from the clustering process and will not improve their lives, and thus die in the same period, whether the network is clustered or not. This results when the network is not used any prediction method, obviously when the prediction method is used then the network sends (e.g. half of data) thus, the network lifetime increases figure 4.11. It is clear that our suggested method has a substantially longer network lifespan than the flat network.

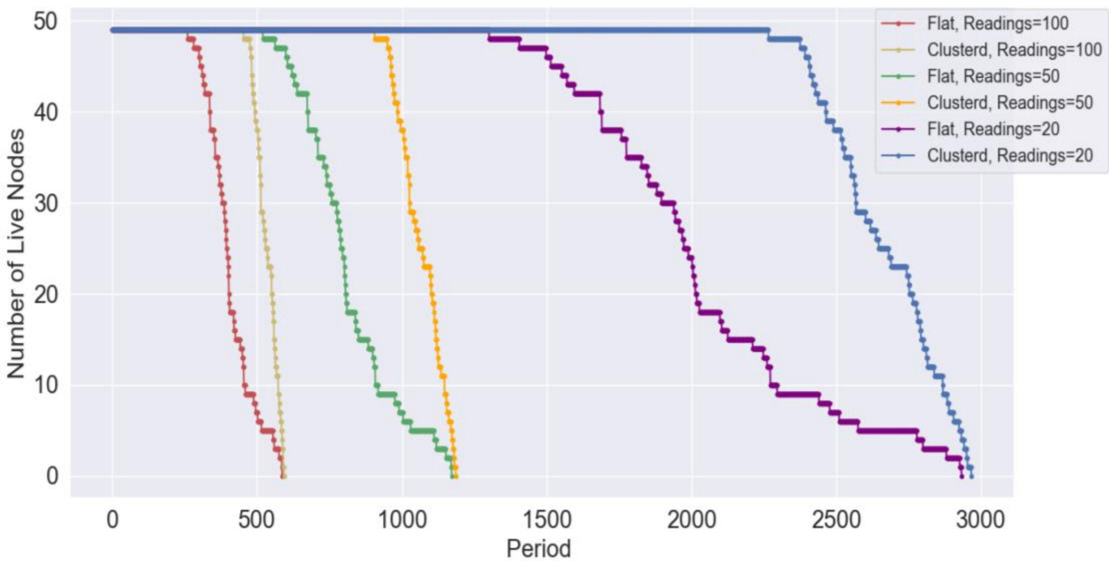


Figure 4.10 Number of nodes live along rounds without prediction.

Figures 4.11 and 4.12 show that DRDPS increase the network stability of FND because the prediction process saves more energy. Figure 4.12 displays the round in which FND, HND, and LND occur using the initial energy 0.1 Joule. When the DRDPS is employed, most of the sensor nodes keep alive because they don't need to make data transmission instead the CH get the data from the prediction.

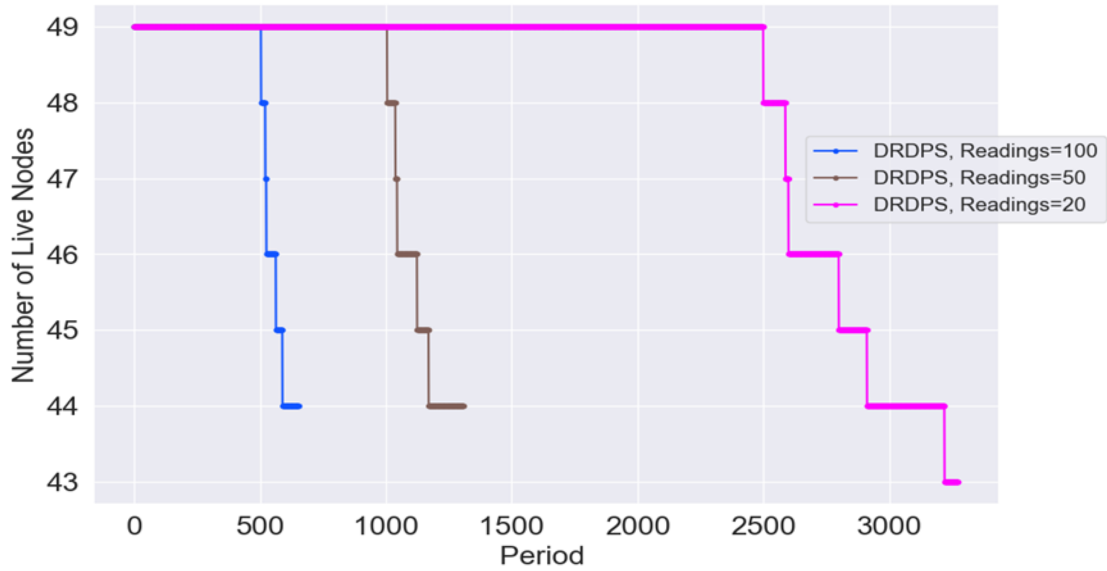


Figure 4.11 Number of nodes live along rounds with DRDPS.

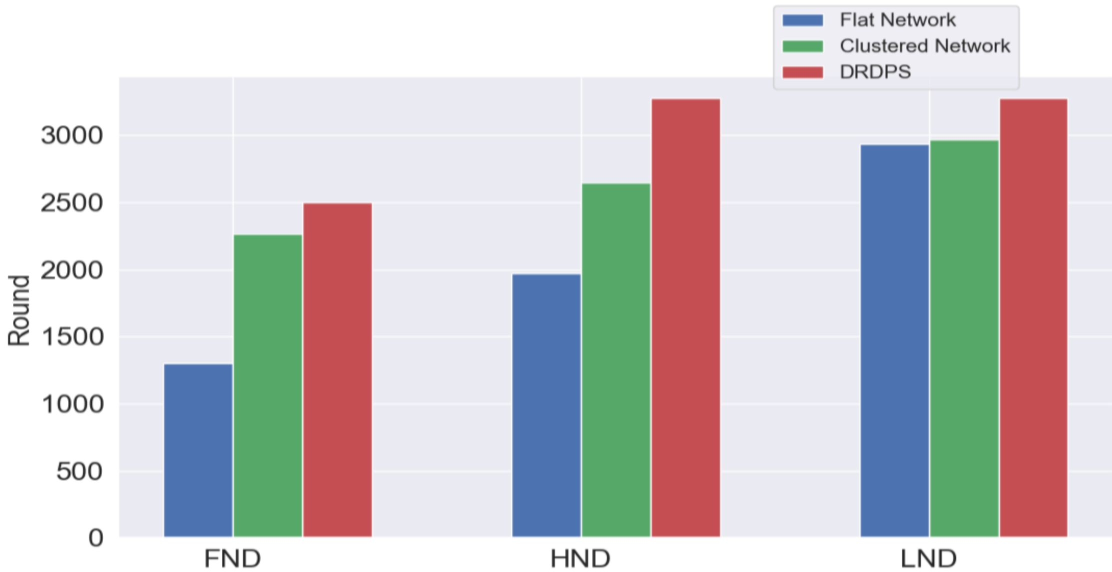


Figure 4.12 Network lifetime in terms of FND, HND, and LND.

E. Prediction Accuracy

There are two factors that have a significant effect on the accuracy of predicted data. They include the prediction error tolerance (pre-defined threshold) and the amount of transmitted data. Obviously, as the error tolerance increase, the amount of transmitted data is decrease as well as the prediction accuracy decrease. There is a trade-off between the amount of data that transmitted, the energy consumption and accuracy. The decrease in transmitted data means less accuracy of predicted data. Figure 4.13 shows the proportion between the error rate and the prediction threshold. To achieve high accuracy, there must be use low prediction threshold which means, high energy consumption and high data transmissions.

The reason for the potential errors is that the prediction method is sensitive to the accumulation of errors within the forecasting horizon. There is an inverse relationship between prediction error and accuracy. The high prediction error value means a low accuracy rate.

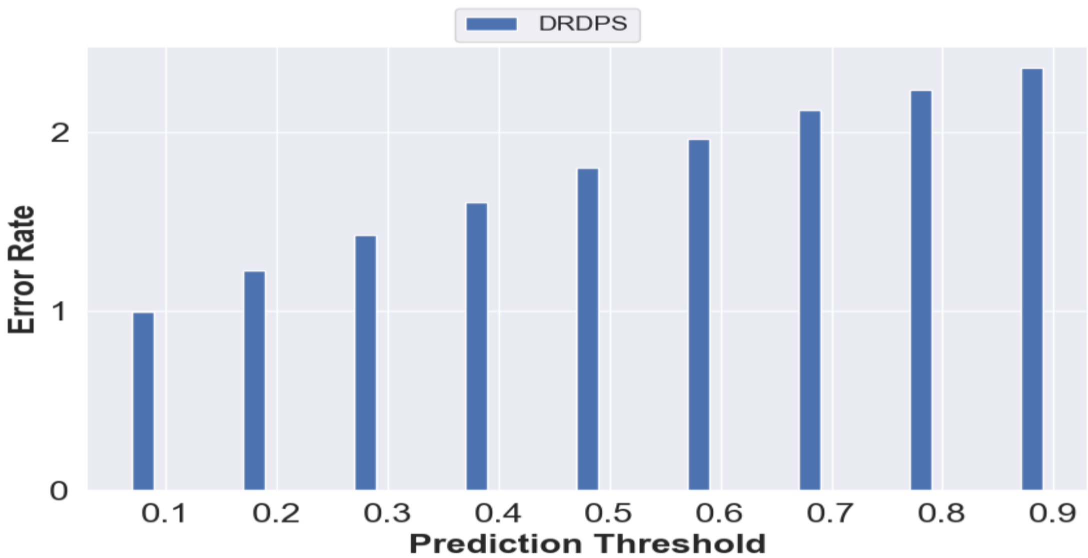


Figure 4.13 RMSE of different error rates of the network

4.4.2 Performance of Cluster Head Level

To illustrate the advantages of the proposed method on the cluster head level, this section presents experiments for the analysis and discussion of the simulation results at this level.

A. Number of Data Sets Transmitted

At the CH level, the amount of data sets transmitted to the BS is an important pointer to the effectiveness of the proposed method. Figure 4.14 shows the number of data sets transmitted from CHs to the BS while various sizes of data readings are tested with/without the Granger causality (GC) algorithm. From the result, it is obvious, the proposed GC make a considerable reduction in the amount of transmission by 54% when the data readings size is 20, which is mean it saves more energy for the network.

Figure 4.14 Amount of data sets transmitted from the CHs to the BS.

B. Energy Consumption at CH Level

Another important factor is the amount of energy consumed by the CH to transmit data to the BS. Figure 4.15 illustrates the energy reduced by half when GC is used which imposes another way to save more energy in the network.

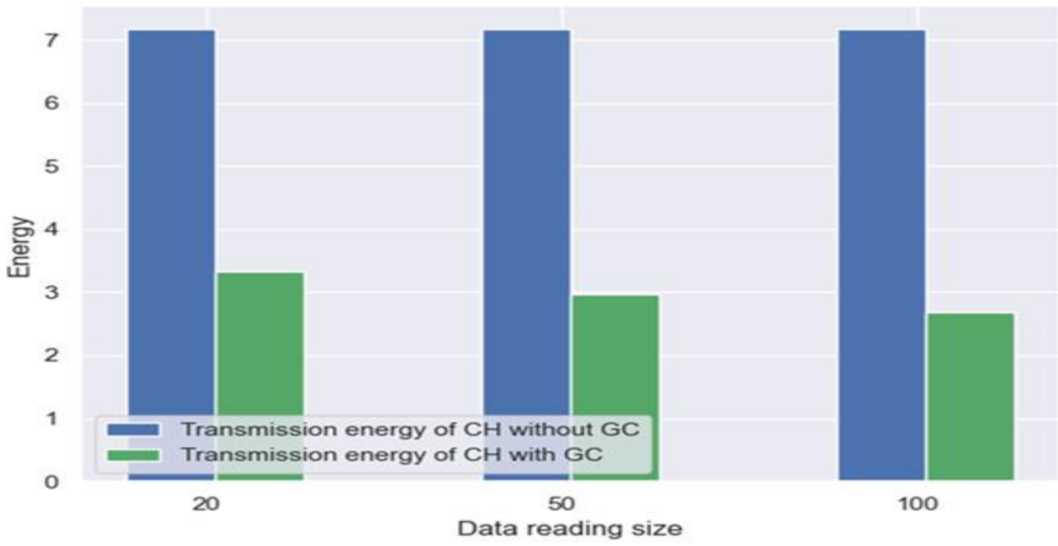


Figure 4.15 Energy consumption of cluster heads level.

4.4.3 Comparison with Other Algorithms

To prove the validity of our proposed algorithm, we have compared the results of the proposed DRDPS algorithm with ELR and P-DPA at varying experiments [23]. Ten data gathering cycles with a small batch of 48 data values of temperature parameters are used to inspect the effect of defining different error thresholds for transmission. The process starts using the historical data and uses the observations only to update the points transmitted.

A. Rate of Data Transmissions

The percentage of data readings transmitted from the SN to the CH can be seen in figure 4.16 with the thresholds in the x-axis.

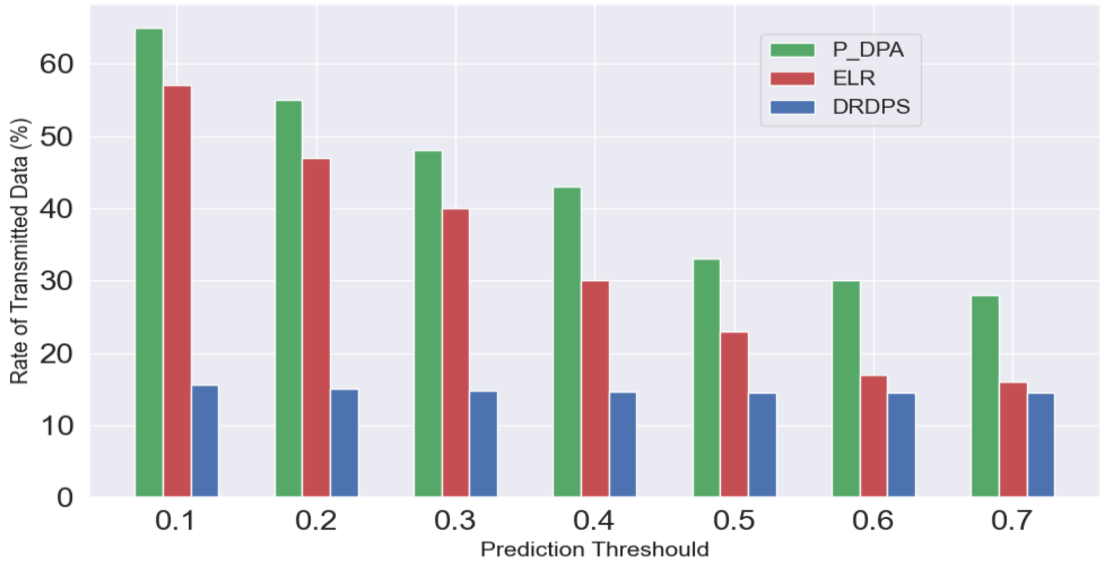


Figure 4.16 Rate of transmitted data comparison of DRDPS, ELR, and P-DPA algorithms.

In the figure above for lower thresholds, the number of readings transmitted is higher because the data have to be more precise; therefore, the SN has to transmit more data to the CH. For our presented method, the data was collected at very convergent intervals, and for a short time. Thus, the data readings are very slight variations, and therefore the transmission rate is saved by 80% because the data readings can be predicted with a very acceptable error value.

B. Rate of Successful Prediction

The proposed method is compared with other algorithms in terms of successful prediction. Figure 4.17 shows this comparison.

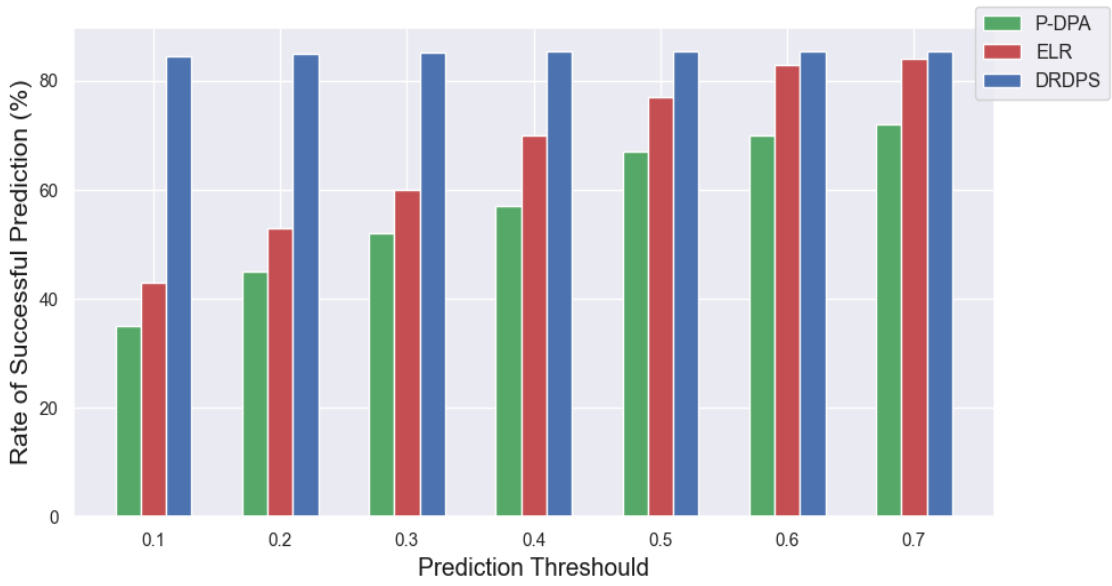


Figure 4.17 Rate of successful prediction comparing DRDPS, ELR, and P-DPA algorithms**.**

In comparison with ELR, and P-DPA methods, it can be seen that the attention model performs much better than the rest especially when the prediction threshold is small. Because of the high correlation of the samples taken by this experiment. The prediction achieves high performance. The figure above shows there is a reduction in transmission of around 85%.

C. Amount of Energy

Another metric is considered to prove the effectiveness of the proposed method. Conservation of energy is a real pointer to evaluate any network performance. Figure 4.18 shows the compression between our method and others.

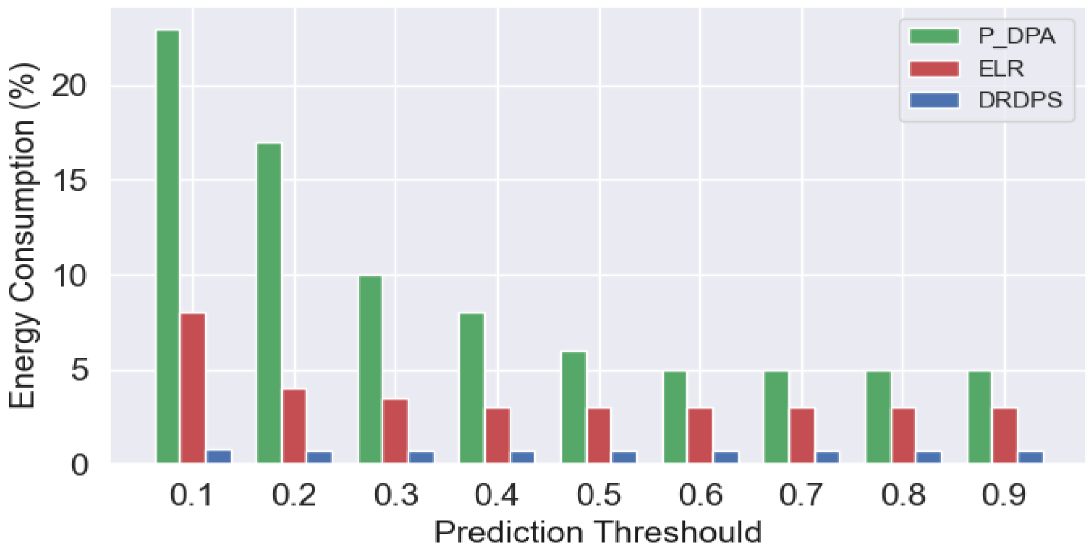
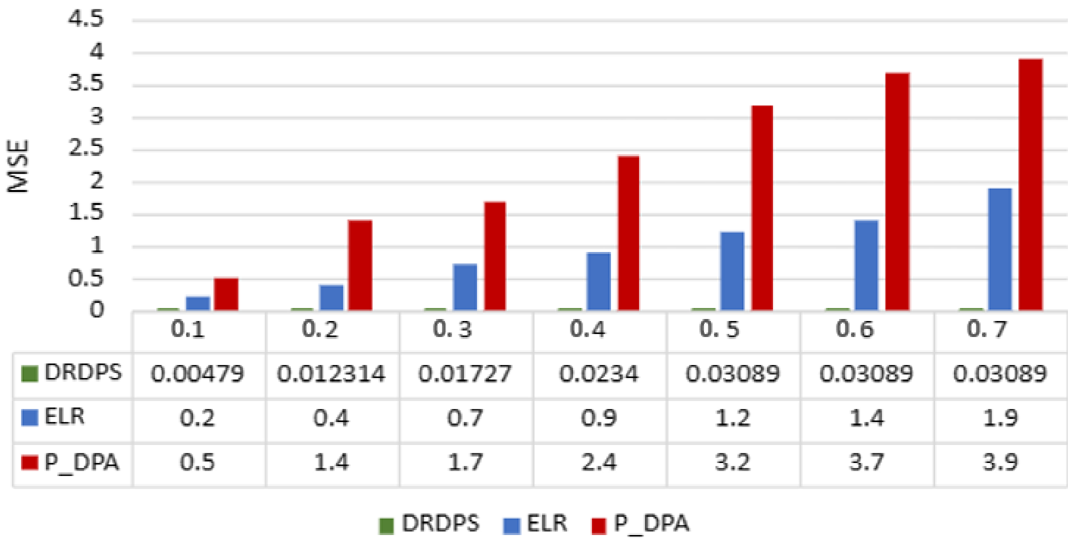


Figure 4.18 Energy conservation comparison of P-DPA, ELR, and DRDPS algorithms.

DRDPS algorithm shows less energy while producing significantly accurate predictions. The results explain that DRDPS provides better energy savage than P-DPA and ELR. As mentioned previously the high value of the prediction threshold means fewer transmissions thus, less energy consumption. Figure 4.18 shows that DRDPS achieve 95% energy saving compared with P-DPA and ELR respectively.

D. Mean Square Error

Mean square error evaluates the error of our proposed prediction algorithm. Figure 4.19 presents the comparison in terms of MSE.



	0.1	0.2	0. 3	0.4	0. 5	0.6	0. 7
DRDPS	0.00479	0.012314	0.01727	0.0234	0.03089	0.03089	0.03089
ELR	0.2	0.4	0.7	0.9	1.2	1.4	1.9
P_DPA	0.5	1.4	1.7	2.4	3.2	3.7	3.9

Figure 4.19 RME comparison for DRDPS with ELR, P-DPA.

From figure 4.19, we notice that the MSE is almost small due to the strong spatial and temporal correlation in the data. The data sample on which the experiment was conducted was collected at a close time, which allows the prediction algorithm to produce accurate results. We note that the MSE of the proposed method is much lower than other methods, where the MSE is 0.005, and this proves the effectiveness of the proposed algorithm when there is a high redundancy in the data.

Chapter Five

Conclusions and Future Works

CHAPTER FIVE

CONCLUSIONS AND FUTURE WORKS

5.1. Conclusions

From the work of this thesis, the following conclusions can be drawn:

1. We propose an energy conservation approach based on two main techniques; network clustering and data reduction-based prediction method called DRDPS.
2. The clustering technique achieves satisfied energy conservation (50%) in comparison with traditional flat networks due to the fact that localize the data transmission distance in short-range communication.
3. In WSN, there is a direct relationship between the amount of transmitted data and energy consumption. More data transmission means more energy consumed by both SNs and CHs. This is because transmitting data requires the use of a node's radio transceiver, which consumes a significant amount of energy.
4. We notice that the energy spent when sending data with a size of 100 readings is 50% higher than sending it with a size of 20 readings
5. DRDPS achieves significant energy conservation (69%, 70%, and 72%) for readings size (100, 50, and 20) respectively. Because it reduces the amount of data transmitted.
6. Using of DRDPS conserves the energy of the whole network by 75% because it works to minimize both distance and amount of data.
7. We noticed a relationship between data accuracy and energy consumption, an increase in the desired data accuracy will increase energy consumption because it imposes more data transmitted to the CH node.
8. The network lifetime increases significantly by applying DRDPS, where network stability (FND), HND, and LND of the network

increased by 100% rounds in comparison with the traditional flat network.

9. The energy consumption in the CH level is minimized by 50% because using such a statistical analysis method able CH sends representative data sets only instead of sending whole data sets thus, minimize the amount of transmitted data which conserves more energy in WSN.

10. The DRDPS outperform some existing approaches (P-DPA, ELR) in terms of several performance metrics like percentage of data sent to CH, energy consumption, prediction accuracy and network lifetime

5.2. Suggestions for Future Works

This section describes plans and suggestions for future work that might follow the work described in this thesis.

1- Using a method to ensure that the data arrives properly when it is sent (Reliability achievement).

2- Use a mobile sink to conserve more energy.

3- Scheduling algorithms can be employed for network lifetime extension.

4- Implementing different strategy to address the cluster head selection issue.

5- Using machine learning as a data reduction method.

References

REFERENCES

[1] M. Song and M. Zheng, “Energy Efficiency Optimization For Wireless Powered Sensor Networks With Nonorthogonal Multiple Access,” *IEEE Sens Lett*, vol. 2, no. 1, pp. 1–4, Jan. 2018, doi: 10.1109/LSENS.2018.2792454.

[2] K. P. Sinaga and M. S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[3] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, “The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News,” *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018*, pp. 533–538, Nov. 2018, doi: 10.1109/ISEMANTIC.2018.8549751.

[4] N. F. Ali, A. M. Said, K. Nisar, and I. A. Aziz, “A survey on software defined network approaches for achieving energy efficiency in wireless sensor network,” *2017 IEEE Conference on Wireless Sensors, ICWiSe 2017*, vol. 2018-January, pp. 28–33, Jul. 2017, doi: 10.1109/ICWISE.2017.8267157.

[5] J. Zhang, K. Yang, L. Xiang, Y. Luo, B. Xiong, and Q. Tang, “A self-adaptive regression-based multivariate data compression scheme with error bound in wireless sensor networks,” *Int J Distrib Sens Netw*, vol. 9, no. 3, p. 913497, 2013.

[6] G. Anastasi, M. Conti, M. di Francesco, and A. Passarella, “Energy conservation in wireless sensor networks: A survey,” *Ad Hoc Networks*, vol. 7, no. 3, pp. 537–568, May 2009, doi: 10.1016/J.ADHOC.2008.06.003.

[7] S. Abbasian Dehkordi, K. Farajzadeh, J. Rezazadeh, R. Farahbakhsh, K. Sandrasegaran, and M. Abbasian Dehkordi, “A survey on data aggregation techniques in IoT sensor networks,” *Wireless Networks*, vol. 26, no. 2, pp. 1243–1263, Feb. 2020, doi: 10.1007/S11276-019-02142-Z/METRICS.

[8] S. Katiyar and D. Prasad, “A Comprehensive Survey of Data Processing Approaches,” 2015.

[9] R. M. Curry and J. C. Smith, “A survey of optimization algorithms for wireless sensor network lifetime maximization,” *Comput Ind Eng*, vol. 101, pp. 145–166, Nov. 2016, doi: 10.1016/J.CIE.2016.08.028.

[10] M. Ezhilarasi and V. Krishnaveni, “A survey on energy and lifetime in wireless sensor networks,” *Comput. Rev. J*, vol. 1, no. 1, pp. 30–51, 2018.

[11] J. Singh, R. Kaur, and D. Singh, “Energy harvesting in wireless sensor networks: A taxonomic survey,” *Int J Energy Res*, vol. 45, no. 1, pp. 118–140, Jan. 2021, doi: 10.1002/ER.5816.

[12] D. Salman and Q. Yas, “Challenges and Issues for Wireless Sensor Networks: A Survey A NEW METHODOLOGY FOR EVALUATION ANDBENCHMARKING OF SKIN DETECTOR BASED ON AI MODEL USING MULTI CRITERIA ANALYSIS View project Plant Pathology View project”, Accessed: Jul. 09, 2023. [Online]. Available: https://www.researchgate.net/publication/349738262

[13] A. K. Idrees, “Energy conservation approach of wireless sensor networks for IoT applications,” *Modern Science*, vol. 7, no. 4, p. 9, 2021.

[14] M. J. Handy, M. Haase, and D. Timmermann, “Low energy adaptive clustering hierarchy with deterministic cluster-head selection,” *2002 4th International Workshop on Mobile and Wireless Communications Network, MWCN 2002*, pp. 368–372, 2002, doi: 10.1109/MWCN.2002.1045790.

[15] T. S. Panag and J. S. Dhillon, “Dual head static clustering algorithm for wireless sensor networks,” *AEU - International Journal of Electronics and Communications*, vol. 88, pp. 148–156, May 2018, doi: 10.1016/J.AEUE.2018.03.019.

[16] B. Jain, G. Brar, and J. Malhotra, “EKMT-k-Means Clustering Algorithmic Solution for Low Energy Consumption for Wireless Sensor Networks Based on Minimum Mean Distance from Base Station,” *Lecture Notes on Data Engineering and Communications Technologies*, vol. 3, pp. 113–123, 2018, doi: 10.1007/978-981-10-4585-1_10.

[17] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” *J (Basel)*, vol. 2, no. 2, pp. 226–235, Jun. 2019, doi: 10.3390/J2020016.

[18] H. El Alami and A. Najid, “ECH: An Enhanced Clustering Hierarchy Approach to Maximize Lifetime of Wireless Sensor Networks,” *IEEE Access*, vol. 7, pp. 107142–107153, 2019, doi: 10.1109/ACCESS.2019.2933052.

[19] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, and J. Henry, “Improved soft-k-means clustering algorithm for balancing energy consumption in wireless sensor networks,” *IEEE Internet Things J*, vol. 8, no. 6, pp. 4868–4881, 2020.

[20] S. el Khediri, W. Fakhet, T. Moulahi, R. Khan, A. Thaljaoui, and A. Kachouri, “Improved node localization using K-means clustering for Wireless Sensor Networks,” *Comput Sci Rev*, vol. 37, Aug. 2020, doi: 10.1016/J.COSREV.2020.100284.

[21] A. Chowdhury and D. De, “Energy-efficient coverage optimization in wireless sensor networks based on Voronoi-Glowworm Swarm Optimization-K-means algorithm,” *Ad Hoc Networks*, vol. 122, p. 102660, Nov. 2021, doi: 10.1016/J.ADHOC.2021.102660.

[22] J. Zhao, H. Liu, Z. Li, and W. Li, “Periodic data prediction algorithm in wireless sensor networks,” in *Advances in Wireless Sensor Networks: 6th China Conference, CWSN 2012, Huangshan, China, October 25-27, 2012, Revised Selected Papers 6*, Springer, 2013, pp. 695–701.

[23] T. Shu, J. Chen, V. K. Bhargava, and C. W. De Silva, “An Energy-Efficient Dual Prediction Scheme Using LMS Filter and LSTM in Wireless Sensor Networks for Environment Monitoring,” *IEEE Internet Things J*, vol. 6, no. 4, pp. 6736–6747, Aug. 2019, doi: 10.1109/JIOT.2019.2911295.

[24] H. Cheng, Z. Xie, L. Wu, Z. Yu, and R. Li, “Data prediction model in wireless sensor networks based on bidirectional LSTM,” *EURASIP J Wirel Commun Netw*, vol. 2019, no. 1, pp. 1–12, Dec. 2019, doi: 10.1186/S13638-019-1511-4/FIGURES/12.

[25] C. Zhang, Y. Liu, F. Wu, W. Fan, J. Tang, and H. Liu, “Multi-Dimensional Joint Prediction Model for IoT Sensor Data Search,” *IEEE Access*, vol. 7, pp. 90863–90873, 2019, doi: 10.1109/ACCESS.2019.2927239.

[26] S. A. Soleymani *et al.*, “A Hybrid Prediction Model for Energy-Efficient Data Collection in Wireless Sensor Networks,” *Symmetry 2020, Vol. 12, Page 2024*, vol. 12, no. 12, p. 2024, Dec. 2020, doi: 10.3390/SYM12122024.

[27] C. Salim and N. Mitton, “K-predictions based data reduction approach in WSN for smart agriculture,” *Computing*, vol. 103, no. 3, pp. 509–532, Mar. 2021, doi: 10.1007/S00607-020-00864-Z.

[28] B. Chreim, J. Nassar, and C. Habib, “Regression-based data reduction algorithm for smart grids,” *2021 IEEE 18th Annual Consumer Communications and Networking Conference, CCNC 2021*, Jan. 2021, doi: 10.1109/CCNC49032.2021.9369555.

[29] K. Jain, A. Agarwal, and A. Kumar, “A Novel Data Prediction Technique Based on Correlation for Data Reduction in Sensor Networks,” *Advances in Intelligent Systems and Computing*, vol. 1164, pp. 595–606, 2021, doi: 10.1007/978-981-15-4992-2_56/COVER.

[30] S. Gowrishankar, T. G. Basavaraju, D. H. Manjaiah, and S. K. Sarkar, “Issues in wireless sensor networks,” in *Proceedings of the world congress on engineering*, 2008, pp. 978–988.

[31] K. S. Adu-Manu, N. Adam, C. Tapparello, H. Ayatollahi, and W. Heinzelman, “Energy-harvesting wireless sensor networks (EH-WSNs): A review,” *ACM Trans Sens Netw*, vol. 14, no. 2, Mar. 2018, doi: 10.1145/3183338.

[32] J. Aponte-Luis, J. A. Gómez-Galán, F. Gómez-Bravo, M. Sánchez-Raya, J. Alcina-Espigado, and P. M. Teixido-Rovira, “An efficient wireless sensor network for industrial monitoring and control,” *Sensors*, vol. 18, no. 1, p. 182, 2018.

[33] F. Akhtar, M. R.-I. Professional, and undefined 2017, “Energy harvesting for self-sustainable wireless body area networks,” *ieeexplore.ieee.org*, Accessed: Jan. 27, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7914600/

[34] M. Saida, G. Zaibi, M. Samet, and A. Kachouri, “Improvement of energy harvested from the heat of the human body,” in *2016 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, IEEE, 2016, pp. 132–137.

[35] X. Hu, W. Xiong, W. Li, and L. Ke, “Application scenarios of wireless sensor networks for urban transportation: A survey,” *Chinese Control Conference, CCC*, vol. 2015-September, pp. 7688–7691, Sep. 2015, doi: 10.1109/CHICC.2015.7260860.

[36] S. Pragadeswaran, S. Madhumitha, and S. Gopinath Assistant Professor, “Certain Investigations on Military Applications of Wireless Sensor Networks,” *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 3, no. 1, 2021, doi: 10.48175/IJARSCT-819.

[37] T. Rault, A. Bouabdallah, and Y. Challal, “Energy efficiency in wireless sensor networks: A top-down survey,” *Computer Networks*, vol. 67, pp. 104–122, Jul. 2014, doi: 10.1016/J.COMNET.2014.03.027.

[38] M. Anwar, A. H. Abdullah, K. N. Qureshi, and A. H. Majid, “Wireless Body Area Networks for Healthcare Applications: An Overview,” *TELKOMNIKA (Telecommunication*

Computing Electronics and Control), vol. 15, no. 3, pp. 1088–1095, Sep. 2017, doi: 10.12928/TELKOMNIKA.V15I3.5793.

[39] N. Hrovatin, A. Tošić, and J. Vičič, “PPWSim: Privacy preserving wireless sensor network simulator,” *SoftwareX*, vol. 18, p. 101067, Jun. 2022, doi: 10.1016/j.softx.2022.101067.

[40] L. Cheng, C. Wu, Y. Zhang, H. Wu, M. Li, and C. Maple, “A Survey of Localization in Wireless Sensor Network,” *Int J Distrib Sens Netw*, vol. 2012, 2012, doi: 10.1155/2012/962523.

[41] A. K. Paul and T. Sato, “Localization in Wireless Sensor Networks: A Survey on Algorithms, Measurement Techniques, Applications and Challenges,” *Journal of Sensor and Actuator Networks 2017, Vol. 6, Page 24*, vol. 6, no. 4, p. 24, Oct. 2017, doi: 10.3390/JSAN6040024.

[42] J. Wen and W. Dargie, “A Mobility Management Protocol for Wireless Sensor Networks,” *2018 IEEE Symposium on Computers and Communications (ISCC)*, vol. 2018-June, pp. 390–396, Jun. 2018, doi: 10.1109/ISCC.2018.8538482.

[43] M. Li and B. Yang, “A Survey on Topology issues in Wireless Sensor Network.,” in *ICWN*, Citeseer, 2006, p. 503.

[44] K. Kishore, S. Sharma, and D. Kumar, “Wireless Sensor Networks-A Review on Topologies and Node Architecture,” *International Journal of Computer Science International Journal of Computer Science International Journal of Computer Science International Journal of Computer*, no. 2, pp. 19–25, 2013, Accessed: Jun. 24, 2023. [Online]. Available: www.ijcaonline.org

[45] P. Garg, K. Saroha, and R. Lochab, “Review of wireless sensor networks-architecture and applications,” *IJCSMS International Journal of Computer Science & Management Studies*, vol. 11, no. 01, pp. 2231–5268, 2011.

[46] L. Borges, F. Velez, … A. L.-C. S. &, and undefined 2014, “Survey on the characterization and classification of wireless sensor network applications,” *ieeexplore.ieee.org*, Accessed: Jun. 24, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6805127/

[47] M. K. Singh, S. I. Amin, S. A. Imam, V. K. Sachan, and A. Choudhary, “A Survey of Wireless Sensor Network and its types,” *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, pp. 326–330, Oct. 2018, doi: 10.1109/ICACCCN.2018.8748710.

[48] S. Messaoud, A. Bradai, S. H. R. Bukhari, P. T. A. Quang, O. ben Ahmed, and M. Atri, “A survey on machine learning in Internet of Things: Algorithms, strategies, and applications,” *Internet of Things (Netherlands)*, vol. 12. Elsevier B.V., Dec. 01, 2020. doi: 10.1016/j.iot.2020.100314.

[49] A. Zeb *et al.*, “Clustering Analysis in Wireless Sensor Networks: The Ambit of Performance Metrics and Schemes Taxonomy:,” *http://dx.doi.org/10.1177/155014774979142*, vol. 12, no. 7, Jul. 2016, doi: 10.1177/155014774979142.

[50] F. Engmann, K. S. Adu-Manu, J.-D. Abdulai, and F. A. Katsriku, "Applications of prediction approaches in wireless sensor networks," in *Wireless Sensor Networks-Design, Deployment and Applications*, IntechOpen, 2021.

[51] V. Chauhan and S. Soni, "Mobile sink-based energy efficient cluster head selection strategy for wireless sensor networks," *J Ambient Intell Humaniz Comput*, vol. 11, no. 11, pp. 4453–4466, Nov. 2020, doi: 10.1007/s12652-019-01509-6.

[52] N. Merabtine, D. Djenouri, and D. E. Zegour, "Towards Energy Efficient Clustering in Wireless Sensor Networks: A Comprehensive Review," *IEEE Access*, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 92688–92705, 2021. doi: 10.1109/ACCESS.2021.3092509.

[53] B. Jiang, G. Huang, T. Wang, J. Gui, and X. Zhu, "Trust based energy efficient data collection with unmanned aerial vehicle in edge network," *Transactions on Emerging Telecommunications Technologies*, 2020, doi: 10.1002/ett.3942.

[54] S. Arjunan and S. Pothula, "A survey on unequal clustering protocols in Wireless Sensor Networks," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 3, pp. 304–317, Jul. 2019, doi: 10.1016/J.JKSUCI.2017.03.006.

[55] M. M. Afsar and M. H. Tayarani-N, "Clustering in sensor networks: A literature survey," *Journal of Network and Computer Applications*, vol. 46, pp. 198–226, Nov. 2014, doi: 10.1016/J.JNCA.2014.09.005.

[56] A. Shahraki, A. Taherkordi, Ø. Haugen, and F. Eliassen, "Clustering objectives in wireless sensor networks: A survey and research direction analysis," *Computer Networks*, vol. 180, p. 107376, Oct. 2020, doi: 10.1016/J.COMNET.2020.107376.

[57] G. H. Adday, S. K. Subramaniam, Z. A. Zukarnain, and N. Samian, "Fault Tolerance Structures in Wireless Sensor Networks (WSNs): Survey, Classification, and Future Directions," *Sensors 2022, Vol. 22, Page 6041*, vol. 22, no. 16, p. 6041, Aug. 2022, doi: 10.3390/S22166041.

[58] C. Nakas, D. Kandris, and G. Visvardis, "Energy Efficient Routing in Wireless Sensor Networks: A Comprehensive Survey," *Algorithms 2020, Vol. 13, Page 72*, vol. 13, no. 3, p. 72, Mar. 2020, doi: 10.3390/A13030072.

[59] S. Ozdemir and Y. Xiao, "Secure data aggregation in wireless sensor networks: A comprehensive overview," *Computer Networks*, vol. 53, no. 12, pp. 2022–2037, Aug. 2009, doi: 10.1016/J.COMNET.2009.02.023.

[60] Z. Huanan, X. Suping, and W. Jiannan, "Security and application of wireless sensor network," *Procedia Comput Sci*, vol. 183, pp. 486–492, Jan. 2021, doi: 10.1016/J.PROCS.2021.02.088.

[61] P. V. M. Movva and R. R. Chintala, "A Brief Survey on Enhanced Quality of Service Mechanisms in Wireless Sensor Network for Secure Data Transmission," pp. 303–314, 2023, doi: 10.1007/978-981-99-1745-7_22.

[62] L. Wu, Y. Peng, J. Fan, Y. Wang, and G. Huang, "A novel kernel extreme learning machine model coupled with K-means clustering and firefly algorithm for estimating

monthly reference evapotranspiration in parallel computation,” *Agric Water Manag*, vol. 245, Feb. 2021, doi: 10.1016/J.AGWAT.2020.106624.

[63] Y. Hussein and S. A. Jalil, “Proposed KDBSCAN Algorithm for Clustering,” *Iraqi Journal of Science*, pp. 173–178, 2018.

[64] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A Quantitative Discriminant Method of Elbow Point for the Optimal Number of Clusters in Clustering Algorithm”.

[65] A. Chowdhury and D. De, “Energy-efficient coverage optimization in wireless sensor networks based on Voronoi-Glowworm Swarm Optimization-K-means algorithm,” *Ad Hoc Networks*, vol. 122, p. 102660, 2021.

[66] A. Alam, M. Muqeem, and S. Ahmad, “Comprehensive review on Clustering Techniques and its application on High Dimensional Data,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 21, no. 6, 2021, doi: 10.22937/IJCSNS.2021.21.6.31.

[67] P. Garikapati, K. Balamurugan, T. P. Latchoumi, and R. Malkapuram, “A Cluster-Profile Comparative Study on Machining AlSi7/63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means,” *Silicon*, vol. 13, no. 4, pp. 961–972, Apr. 2021, doi: 10.1007/S12633-020-00447-9.

[68] R. Gustriansyah, N. Suhandi, and F. Antony, “Clustering optimization in RFM analysis based on k-means,” *Indones. J. Electr. Eng. Comput. Sci*, vol. 18, no. 1, pp. 470–477, 2020.

[69] A. Agarwal and A. Dev, “A normalized cosine distance based regression model for data prediction in WSN,” in *Communications in Computer and Information Science*, Springer Verlag, 2018, pp. 298–308. doi: 10.1007/978-981-13-1813-9_30.

[70] X. Xu and G. Zhang, “A Hybrid Model for Data Prediction in Real-World Wireless Sensor Networks,” *IEEE Communications Letters*, vol. 25, no. 5, pp. 1712–1715, May 2021, doi: 10.1109/LCOMM.2017.2706258.

[71] S. Khriji, D. el Houssaini, I. Kammoun, and O. Kanoun, “Energy-efficient techniques in wireless sensor networks,” in *Energy Harvesting for Wireless Sensor Networks: Technology, Components and System Design*, De Gruyter, 2018, pp. 287–303. doi: 10.1515/9783110445053-017.

[72] F. Engmann, F. A. Katsriku, J.-D. Abdulai, K. S. Adu-Manu, and F. K. Banaseka, “Prolonging the lifetime of wireless sensor networks: a review of current techniques,” *Wirel Commun Mob Comput*, vol. 2018, 2018.

[73] A. L. Schaffer, T. A. Dobbins, and S.-A. Pearson, “Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions,” *BMC Med Res Methodol*, vol. 21, no. 1, pp. 1–12, 2021.

[74] “Learn Time Series Analysis in Python- A Step by Step Guide using the ARIMA Model | Ai Online Course.” https://www.aionlinecourse.com/blog/time-series-analysis-in-python (accessed Jan. 24, 2022).

[75] C. R. Morales, F. R. de Sousa, V. Brusamarello, and N. C. Fernandes, “Multivariate Data Prediction in a Wireless Sensor Network based on Sequence to Sequence Models,”

Conference Record - IEEE Instrumentation and Measurement Technology Conference, vol. 2021-May, May 2021, doi: 10.1109/I2MTC50364.2021.9459957.

[76] O. J. Odeyinka, O. A. Ajibola, M. C. Ndinechi, O. C. Nosiri, and N. C. Onuekwusi, “A Review on Conservation of Energy in Wireless Sensor Networks,” *International Journal of Smart Sensor Technologies and Applications*, vol. 1, no. 2, pp. 1–16, Jul. 2021, doi: 10.4018/ijssta.20200401.oa1.

[77] A. M. Kadhum Al-Qurabat, A. Kadhum Idrees, A. Makhoul, and C. Abou Jaoude, “A Bi-Level Data Lowering Method to Minimize Transferring Big Data in the Sensors of IoT Applications,” *kijoms.uokerbala.edu.iq*, vol. 8, no. 2, 2022, doi: 10.33640/2405-609X.3228.

[78] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, “Deep Learning for Time Series Forecasting: A Survey,” *Big Data*, vol. 9, no. 1, pp. 3–21, Feb. 2021, doi: 10.1089/BIG.2020.0159.

[79] Y. Wang and L. Cheng, “A combination of residual and long–short-term memory networks for bearing fault diagnosis based on time-series model analysis,” *Meas Sci Technol*, vol. 32, no. 1, p. 015904, Oct. 2020, doi: 10.1088/1361-6501/ABAA1E.

[80] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[81] G. A. M. Jawad, A. K. M. Al-Qurabat, and A. K. Idrees, “Maximizing the underwater wireless sensor networks’ lifespan using BTC and MNP5 compression techniques,” *Annales des Telecommunications/Annals of Telecommunications*, 2022, doi: 10.1007/S12243-021-00903-6.

[82] C. R. Morales, F. R. de Sousa, V. Brusamarello, and N. C. Fernandes, “Evaluation of deep learning methods in a dual prediction scheme to reduce transmission data in a WSN,” *Sensors*, vol. 21, no. 21, Nov. 2021, doi: 10.3390/s21217375.

[83] H. Liazid, M. Lehsaini, and A. Liazid, “An improved adaptive dual prediction scheme for reducing data transmission in wireless sensor networks,” *Wireless Networks*, vol. 25, no. 6, pp. 3545–3555, Aug. 2019, doi: 10.1007/s11276-019-01950-7.

[84] G. M. Dias, B. Bellalta, and S. Oechsner, “The impact of dual prediction schemes on the reduction of the number of transmissions in sensor networks,” *Comput Commun*, vol. 112, pp. 58–72, Nov. 2017, doi: 10.1016/J.COMCOM.2017.08.002.

[85] J. C. López-Ardao, R. F. Rodríguez-Rubio, A. Suárez-González, M. Rodríguez-Pérez, and M. E. Sousa-Vieira, “Current Trends on Green Wireless Sensor Networks,” *Sensors 2021, Vol. 21, Page 4281*, vol. 21, no. 13, p. 4281, Jun. 2021, doi: 10.3390/S21134281.

[86] G. M. Dias, B. Bellalta, and S. Oechsner, “A Survey About Prediction-Based Data Reduction in Wireless Sensor Networks,” *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, Nov. 2016, doi: 10.1145/2996356.

[87] X. Wang, Y. Kang, R. J. Hyndman, and F. Li, “Distributed ARIMA models for ultra-long time series,” *Int J Forecast*, 2022.

[88] J. Fattah, L. Ezzine, Z. Aman, H. el Moussami, and A. Lachhab, “Forecasting of demand using ARIMA model,” *International Journal of Engineering Business Management*, vol. 10, Oct. 2018, doi: 10.1177/1847979018808673.

[89] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, “Time series analysis: Forecasting and control: Fourth edition,” *Time Series Analysis: Forecasting and Control: Fourth Edition*, pp. 1–746, May 2013, doi: 10.1002/9781118619193.

[90] S. A. Soleymani *et al.*, “A Hybrid Prediction Model for Energy-Efficient Data Collection in Wireless Sensor Networks,” *Symmetry 2020, Vol. 12, Page 2024*, vol. 12, no. 12, p. 2024, Dec. 2020, doi: 10.3390/SYM12122024.

[91] C. B. Aditya Satrio, W. Darmawan, B. U. Nadia, and N. Hanafiah, “Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET,” *Procedia Comput Sci*, vol. 179, pp. 524–532, 2021, doi: 10.1016/J.PROCS.2021.01.036.

[92] J. Fattah, L. Ezzine, Z. Aman, H. el Moussami, and A. Lachhab, “Forecasting of demand using ARIMA model,” *International Journal of Engineering Business Management*, vol. 10, p. 1847979018808673, 2018.

[93] A. L. Schaffer, T. A. Dobbins, and S.-A. Pearson, “Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions,” *BMC Medical Research Methodology 2021 21:1*, vol. 21, no. 1, pp. 1–12, Mar. 2021, doi: 10.1186/S12874-021-01235-8.

[94] Ștefan C. Gherghina, D. Ștefan Armeanu, and C. C. Joldeș, “Stock market reactions to COVID-19 pandemic outbreak: Quantitative evidence from ARDL bounds tests and granger causality analysis,” *Int J Environ Res Public Health*, vol. 17, no. 18, pp. 1–35, Sep. 2020, doi: 10.3390/IJERPH17186729.

[95] L. Lopez and S. Weber, “Testing for Granger causality in panel data,” *Stata Journal*, vol. 17, no. 4, pp. 972–984, 2017, doi: 10.1177/1536867X1801700412.

[96] M. de Liu, L. Ding, and Y. L. Bai, “Application of hybrid model based on empirical mode decomposition, novel recurrent neural networks and the ARIMA to wind speed prediction,” *Energy Convers Manag*, vol. 233, Apr. 2021, doi: 10.1016/J.ENCONMAN.2021.113917.

[97] J. Li, H. Y. Zhou, D.-C. Zuo, K. M. Hou, H. P. Xie, and P. Zhou, “Energy consumption evaluation for wireless sensor network nodes based on queuing Petri net,” *Int J Distrib Sens Netw*, vol. 10, no. 4, p. 262848, 2014.

[98] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” in *Proceedings of the 33rd annual Hawaii international conference on system sciences*, IEEE, 2000, pp. 10-pp.

[99] R. D. Gawade and S. L. Nalbalwar, “A Centralized Energy Efficient Distance Based Routing Protocol for Wireless Sensor Networks,” 2016, doi: 10.1155/2016/8313986.

[100] S. A. Abdulzahra, A. K. M. Al-Qurabat, and A. K. Idrees, “Compression-based data reduction technique for IoT sensor networks,” *Baghdad Sci. J*, vol. 18, no. 1, pp. 184–198, 2021.

[101] F. Engmann, F. A. Katsriku, J.-D. Abdulai, K. S. Adu-Manu, and F. K. Banaseka, “Prolonging the lifetime of wireless sensor networks: a review of current techniques,” *Wirel Commun Mob Comput*, vol. 2018, 2018.

[102] P. Ramos, N. Santos, and R. Rebelo, “Performance of state space and ARIMA models for consumer retail sales forecasting,” *Robot Comput Integr Manuf*, vol. 34, pp. 151–163, Aug. 2015, doi: 10.1016/J.RCIM.2014.12.015.

[103] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” in *Proceedings of the Hawaii International Conference on System Sciences*, 2000. doi: 10.1109/hicss.2000.926982.

[104] A. K. M. Al-Qurabat, A. K. Idress, A. Makhoul, and C. A. Jooude, “A Bi-Level Data Lowering Method to Minimize Transferring Big Data in the Sensors of IoT Applications,” *Karbala International Journal of Modern Science*, vol. 8, no. 2, pp. 123–138, 2022, doi: 10.33640/2405-609X.3228.

الخلاصـة

في شبكة أجهزة الاستشعار اللاسلكية (WSN) ، يلزم أن تتدفق البيانات التي يتم جمعها بواسطة تطبيقات المراقبة بشكل مستمر. على الرغم من أن شبكات الاستشعار اللاسلكية WSN قد حققت نجاحًا في العديد من التطبيقات ، إلا أن تنفيذها على نطاق واسع في المناطق التي يتعذر الوصول إليها والنائية يواجهه عقبات كبيرة – مثل قيود الطاقة. وهذا يرجع في الغالب إلى حقيقة أنها تعمل بالبطارية في معظم الحالات. صيانة تلك الشبكات مثل ؛ استبدال / إعادة شحن البطاريات يعتبر تحديًا في كثير من الحالات. لذلك أصبح الحفاظ على الطاقة من القضايا الرئيسية في معظم التصاميم لشبكات الاستشعار اللاسلكي WSN.

تركز هذه الأطروحة على إطالة العمر الافتراضي للشبكة من خلال استثمار طرق تقليل البيانات في WSN. في هذه الأطروحة ، نقترح بروتوكولًا موفرًا للطاقة يسمى نهج لتقليل البيانات يستند على التنبؤ المزدوج (DRDPS). اقترح DRDPS التحكم في كمية البيانات المنقولة عن طريق تجنب نقل البيانات الزائدة عن الحاجة وبالتالي تجنب الاستهلاك المفرط للطاقة. يتم تنفيذ النهج المقترح على مستويين ؛ مستوى عقدة الاستشعار ، ومستوى رأس الكتلة. يتم تطبيق النهج المقترح على ثلاث مراحل ؛ مرحلة التجميع ومرحلة التنبؤ المزدوج بالبيانات (DDP) والذي يعمل في كلا من المستشعر ورأس الكتلة و مرحلة التنبؤ بالبيانات القائم على الإحصاء والذي يعمل في رأس الكتلة CHs. يتم استخدام ثلاث خوارزميات لتقليل استهلاك الطاقة وإطالة عمر الشبكة WSN ، وهذه الخوارزميات هي خوارزميات K-mean ، ومتوسط متحرك متكامل ذاتي الانحدار (ARIMA) ، وخوارزمية سببية جرانجر (GC).

خوارزمية K-mean عبارة عن خوارزمية تجميع تستخدم لتقسيم شبكة المستشعرات إلى مجموعات ، وبالتالي تقييد نقل البيانات إلى أصغر مساحة ممكنة ، يتم تحديد العدد الأمثل للعناقيد باستخدام طريقة Elbow. تتنبأ ARIMA بقراءات البيانات للجولات التالية على مستوى عقدة المستشعر بطريقة مزدوجة. بينما طريقة GC هي طريقة إحصائية تُستخدم على مستوى رأس الكتلة لتقليل كمية مجموعات البيانات التي ترسل من CH إلى BS. تم تقييم أداء DRDPS من خلال حساب المقاييس: عمرالشبكة، واستهلاك طاقة الشبكة ، وكمية البيانات المرسلة ، ودقة التنبؤ بالبيانات. أيضًا ، تمت مقارنة نتائج DRDPS مع طرق P-DPA و ELR. أظهرت نتائج التجارب أن النهج المقترح يوفر تحسن ملحوظ في اداء شبكة WSN ، ولديه القدرة على تعزيز العمر والحفاظ على دقة مقبولة للبيانات مقارنة بالشبكة التقليدية المسطحة. يتم تقييم نتائج المحاكاة وفحصها باستخدام لغة برمجة Python 3.7.

2022 International Conference for Natural and Applied Sciences (ICNAS), AL-Hikma University College, Baghdad, Iraq

An Improved Energy Efficient Clustering Protocol for Wireless Sensor Networks

Wisal Bassim Nedham
dept. of Dentistry
Al-Mustaqbal University College
Babylon, Iraq
wisal.basim@mustaqbal-college.edu.iq

Ali Kadhum M. Al-Qurabat
dept. of Computer Science
College of Science for Women
University of Babylon
Babylon, Iraq
alik.m.alqurabat@uobabylon.edu.iq

Abstract—**The energy resources available to nodes in wireless sensor networks are limited, so they must be wisely used. Clustering is a useful technique for reducing energy consumption and extending the life of a network. In this study, we presented an Energy-Saving Clustering Algorithm (ESCA) to reduce energy consumption and increase the network's lifetime. The clustering phase is based on cluster construction that is centralized and cluster heads that are distributed. The clustering is stationary and determined using a centralized K-means method, with the created clusters remaining static throughout the process. Subsequently, according to the varying amounts of energy in the nodes, it chooses and rotates the cluster heads (CHs) within those clusters to reduce energy expenditure before the data transmission phase to the base station (BS). The suggested ESCA is compared to the current two MOFCA, and IGHND clustering methods using a Python-based custom simulator. As a consequence, the suggested ESCA efficiently tackles the energy use issue while also greatly extending the network's lifetime.**

Keywords— WSN, Clustering, Energy consumption, K-Means

I. INTRODUCTION

This Several resource restrictions, including the range of transmission, storage, computing power, and energy, are present in wireless sensor networks (WSNs). The most significant resource restriction of WSNs seems to be the sensors' energy. A great number of studies have been conducted over recent years to solve this issue. WSNs are spatially distributed for collecting data for applications that cover a vast region, like solar photovoltaic cell monitoring in a grid, rail tunnel monitoring, coal mines, forestry, agriculture, and so on, and WSNs demand data from all regions [1] [2]. In just about all situations, the base station (BS) is located a distance from the field of sensing. The BS collects data from these kinds of networks on a periodic basis [3]. For the creation of networks for continuous monitoring, hierarchical topology with clustering has been proven effective [4]. It has been demonstrated that network clustering has a long life compared to a network using straight data transfer [5]. Clustering techniques have a number of advantages in data collection networks. The clustering of sensors is intended to decrease the count of long-distance transfers, resulting in energy savings. Clustering extends the sleep durations of regular sensor nodes (cluster members (CMs)), whereas cluster heads (CHs) regulate the actions of their CMs, leading to energy savings once again. The majority of the action scheduling is carried out via TDMA-based schedules [6] [7].

Additionally, clustering improves the aggregation of data at the CH by reducing the count of sent packets, resulting in lower sensor node energy consumption. In clustering protocols, communication takes place in two stages: intra-cluster, which occurs inside clusters, and inter-cluster, which occurs between clusters and the BS. Moreover, in a WSN clustering protocol, communication could be accomplished via either single- or multi-hop transmission routing. Because the distance between sensors inside the cluster is generally low, most clustering protocols employ single-hop communication to communicate within the cluster, e.g. HEED [8], LEACH [9], etc. Direct transmission, on the other hand, is helpful since the radio wastes energy during both reception and transmission. However, direct transmission has its own limitations. It's only suitable for usage until to a specified threshold distance [10]. That's also due to the fact that once transmission distance exceeds a certain threshold, the energy cost rises in proportion to the distance's fourth power. A CH is severely loaded in a clustering protocol since it must execute many duties including cluster creation, data aggregation, data transfer, and relaying. In comparison to non-CH nodes, CHs spend more energy [11] [12].

The optimization of network lifespan is one of the fundamental considerations in WSNs, since once the network becomes dysfunctional, a substantial quantity of energy must not stay in the nodes, as this would be waste [10]. The network lifespan has been specified in several research papers as when the first node dies (FND). The death of the first node is not considered a general definition because the lifetime criteria are application-specific. Because there are several sorts of network applications, the network's lifetime was already assessed at various phases, such as the moment once the first node dies or a particular proportion of nodes fails. In any situation, it's more crucial that the network operates independently and reliably till the end of its lifespan [13].

In this article, the network clustering is handled by the BS. This is a centralized method in which the BS splits the network into clusters at the start of the protocol and allocates the function of CH to the node in every cluster having the greatest energy level. Subsequently, according to the varying amounts of energy in the nodes, it chooses and rotates the CHs within those clusters to reduce energy expenditure before the data transmission phase to the BS. The clustering within that article is stationary and determined using a centralized K-means method, with the created clusters remaining static throughout the process. It's indeed important to mention that the proportion of CHs in the network is determined by various network deployment characteristics; the most significant of those is the topology of the network and the number of k (for the K-means method) that meets the criteria demanded by the application employing this WSN.

Forthcoming and Online First Articles

International Journal of Computer Applications in Technology

- **A comprehensive review of clustering approaches for energy efficiency in wireless sensor networks**

  by Wesal Bassem Nedham, Ali Kadhum M. Al-Qurabat

  Abstract: Wireless Sensor Networks (WSNs) have become more popular in recent years due to their vast range of applications. The use of WSNs is an absolute requirement for future revolutionary domains such as smart cities, the Internet of Things, or ecological fields, where hundreds or thousands of sensor nodes are placed. Moreover, because WSNs are energy-constrained networks, implementing energy-aware protocols is critical. Hierarchical techniques enhance network performance and extend network lifetime in large-scale WSNs. Within a WSN, hierarchy is achieved by dividing the network into sub-networks known as clusters, which are directed by Cluster Heads (CH). Clustering is the most common energy-efficient approach, and it offers several benefits, such as reduced latency, scalability, lifetime, and energy efficiency. This study presents a detailed assessment of several clustering techniques, together with their aims, features, etc. Furthermore, clustering techniques are classified and evaluated based on numerous cluster features, cluster head attributes, and clustering procedures.

  Keywords: *wireless sensor networks; energy consumption; energy efficiency; clustering techniques; IoT.*

Forthcoming and Online First Articles

International Journal of Computer Applications in Technology

- **A review of current prediction techniques for extending the lifetime of wireless sensor networks**

by Wesal Bassem Nedham, Ali Kadhum M. Al-Qurabat

Abstract: The possibility for broad usage of wireless sensor networks (WSNs) in many various sectors, such as environmental monitoring, security, home automation, and many others, has increased research interest in WSNs. Although its successes, the broad proliferation of WSNs, especially in distant and inhospitable areas where their usage is most advantageous, is hindered by the primary obstacle of limited energy, as they are often battery operated. To provide these energy-hungry sensor nodes with a longer life expectancy, one technique to achieve this aim is to reduce the frequency of data transfer. Conversely, a portion of the observed data could be predicted to avoid initiating communications that might overwhelm the wireless channel. In this paper, we classify and analyse current prediction-based data reduction strategies for WSNs. Our key contribution is a systematic technique for choosing a prediction model in WSNs based on WSN limitations, prediction technique features, and observed data.

Keywords: *wireless sensor networks; prediction models; time series models.*

وزارة التعليم العالي و البحث العلمي
جامعــة بابــل/كليــة العلــوم للبنــات
قســــــم علــــــوم الحاســــبات

طريقة محسنة للحفاظ على الطاقة في شبكة الاستشعار اللاسلكية

رســـــالة قُـــــدِمَتْ

الى كلية العلوم للبنات في جامعة بابل وهي جزء من متطلبات الحصول على شهادة الماجستير في العلوم/علوم الحاسبات

الطالبــة

وصـــال باسم نظــام المــرزوك

دبلــوم عالـــي في علــوم الحاســوب / كليــة العلــوم للبنات / جامعــة بابــل 2019

إشــراف

أ.م.د. علـــي كاظم محمـــد الغرابي

1444 هـ

2023 م