IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Data Redundancy Reduction for Energy-Efficiency in Wireless Sensor Networks: A Comprehensive Review

GUL SAHAR[12], KAMALRULNIZAM ABU BAKAR[1], FATIMA TUL ZUHRA[1], SABIT RAHIM[2], TEHMINA BIBI[3], AND SYED HAMID HUSSAIN MADNI[1]

[1]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia 81310, Johor Bahru, Malaysia.
[2]Department of Computer Sciences, Karakoram International University, Gilgit-Baltistan 15100, Pakistan.
[3]Institute of Geology, University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan.

Corresponding author: Gul Sahar (Email: gulsahar@kiu.edu.pk).

ABSTRACT Wireless Sensor Networks (WSNs) play a significant role in providing an extraordinary infrastructure for monitoring environmental variations such as climate change, volcanoes, and other natural disasters. In a hostile environment, sensors' energy is one of the crucial concerns in collecting and analyzing accurate data. However, various environmental conditions, short-distance adjacent devices, and extreme usage of resources, i.e., battery power in WSNs, lead to a high possibility of redundant data. Accordingly, the reduction in redundant data is required for both resources and accurate information. In this context, this paper presents a comprehensive review of the existing energy-efficient data redundancy reduction schemes with their benefits and limitations for WSNs. The entire concept of data redundancy reduction is classified into three levels, which are node, cluster head, and sink. Additionally, this paper highlights existing key issues and challenges and suggested future work in reducing data redundancy for future research.

INDEX TERMS Cluster-based, data redundancy, energy efficiency, reduction, wireless sensor networks

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are extensively used in hostile environments and large-scale applications [1][2], such as terrestrial, underground, underwater, multimedia [3]. Other applications include volcanoes, military issues [4][5], glaciers, earthquakes, agriculture [6][7], industry, environmental issues [8]–[10], and healthcare [11] etc. However, distribution deployment, designing, and energy consumption are the most common issues, and challenges of these applications as millions of sensor nodes are distributed in these large-scale areas. The life span of each sensor is entirely dependent on the battery power to perform different tasks such as sensing, computation, processing, and transmission for data collection. Data transmission, on the other hand, uses more energy than other processes.

In WSNs, data-driven models are used for various applications and are classified into four fundamental data-driven models. A query-driven model is obtained for specific knowledge items required from different places, such as home applications and logistic applications [12]. In a query-driven model, data is gathered, stored locally and transmitted on a requested suitable module for a certain knowledge item that is required from multiple locations. Event-driven data is inactive non-continuous, and transmitted with high energy consumption when the events occur, such as forest fire, mass movement, surveillance, earthquake, and forecasting of the flood, etc. Time driven is also known as periodic sensor data, mostly used for monitoring of a particular phenomenon such as melting glaciers, earthquakes, and healthcare. Furthermore, the periodic sensor continually collects data from the physical environment and reports it to the base station [13]. However, energy is mainly consumed due to the continuous sensing and reporting data to the sink nodes. Environmental objects move quickly or slowly, yet the data is identical or duplicated in both cases, increasing transmission costs. Different methods or processes, such as data aggregation and network hoping, are utilized to minimize transmission costs. The two forms of network hopping are single and multiple hopping. In single hopping, the data is directly sent from sensors to the sink node. Due to long-distance, transmission cost increases. Single hop is not suitable for large-scale regions. Thus, multi hop is used in large-scale regions. Multi hop is primarily used in hierarchical routing protocols such as chain based, tree based, and cluster based [14]. The best energy-saving protocol is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

cluster based architecture [15]. The data transmission in nodes clustering covers a network view reduced between the nodes and the sink for an extended network lifetime.

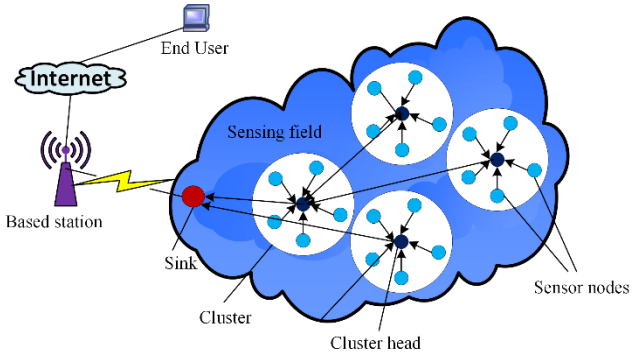

FIGURE1. Cluster-based architecture of the WSN.

Figure 1 presents the cluster-based network both single and multiple hop routing. In general, cluster-based architecture consists of sensing area, sensor nodes, cluster head, base station, a sink node, and end-user. Clusters of sensor nodes are classified into distinct categories. Each cluster group has a cluster head which is responsible for identifying the nodes that constitute a cluster. A node in a cluster is responsible for gathering data from the member nodes in its cluster and for transmitting these data to the Base Station. The data is kept in three separate places in the cluster-based architecture: at the node level, at the cluster head, and finally, at the sink level, with the entire network's data.

Previous survey and review studies focused on specific areas such as a survey on wireless sensor networks [13]–[18], energy-efficient hierarchical routing protocols [19]–[22], energy efficiency data aggregation techniques [23]–[27], challenges and design goals [28]–[31], big data [32]–[34], energy-efficient scheduling [35]–[37], wireless sensor network applications [4], [7], [44]–[46], [8], [11], [38]–[43] and data redundancy [47]–[49] in WSNs (detailed in Section II). However, this review article aims to classify the existing data collection and transmission mechanisms used by sensor nodes in WSNs. As a result, the state of the art of WSNs is described in terms of energy efficient data redundancy. The existing methods are classified into three levels: the node, cluster head, and sink. For each classification level, the performance of comparative and simulation parameters are also described based on existing studies with suggested future works.

In this review article, our contributions are mentioned as follows: This review article gives a comprehensive literature analysis of energy efficiency for WSNs, emphasising energy-efficient data redundancy reduction strategies.

- It elaborates the schemes and methods of data redundancy for reducing in cluster-based architecture with the classification into various levels.
- It analyzes the different approaches, methods, and schemes used in the energy-efficient data redundancy in WSN as well as highlights their benefits and weaknesses.
- This review article also highlights all the performance metrics used to evaluate existing works of WSN.
- Lastly, the summary of the suggested future works from the previous research is assembled, with the ambition that it will help new researchers follow new and innovative directions of energy efficiency in WSN.

However, analyzing the existing approaches and considering their core ideas helps develop some additional applicable and enhanced techniques that might be an improved version of the existing techniques. This review article will assist future researchers in understanding status, needs, and future requirements and finding the loopholes responsibly for energy efficiency in WSNs.

The rest of the paper is divided into the following sections: Section II provides the former survey and reviews researchers based on energy efficiency in WSNs. Section III presents the data redundancy problems in WSNs; Section IV shows the detailed classification of various data redundancy schemes. In Section V, the study focuses on the analysis of parameters with statistical determination. In Section VI, some direction of suggested future research areas in data redundancy reducing for the energy of WSNs are discussed, recommendation and conclusion are presented in Section VII.

II. RELATED WORKS

This section presents a comprehensive review of various existing surveys, comparative studies, and reviews focused on energy efficiency in the field of various perspectives for WSNs.

Various programming approaches and model techniques of WSNs design methodologies are aggregated and explained by [16]. The study discusses two main approaches including low-level-based and high-level-based approaches. Designing environment, power supply design, reconfiguration scenario, and non-functional property (NFP) verification are some of the issues stated in research for WSNs design. Furthermore, the main purpose of the review is to evaluate the architectures, different types, applications, and challenges of wireless sensor networks [50][17]. Researchers also study various techniques for energy and lifetime of WSNs and explain some general issues. Connectivity, coverage, node deployment, environment, fault tolerance, scalability, data aggregation, quality of service, hardware limits, and energy are some of the challenges identified in past studies [18].

Chan et al. [22] survey the literature about hierarchical routing protocols for WSNs and present a comparative analysis of each routing protocol's advantages, disadvantages, and performance issues. On the other hand, a comprehensive review of the classical and swam intelligence approach

delivers the basics of using inside hierarchical energy-efficient routing protocols [19]. The recaps on hierarchical routing protocols are revised by considering energy efficiency, fault tolerance, location awareness, load balancing, quality of service (QoS), scalability, data aggregation, multipath, and query-based. The issues of objectives, methods, classifications, performance metrics, and other open issues are highlighted in detail for further research. Similarly, energy consumption is studied for WSNs lifetime [44]. The study explores the specifics of WSNs applications, design, and network structure and examines the energy efficiency of proactive routing algorithms considering the strengths and flaws. The energy efficiency protocols on periodic sensor networks based on their performance (life duration) and stability parameters are also compared. Furthermore, Sabor et al. [20] give a comprehensive survey about hierarchical-based routing protocols (HBRP) for WSNs based on communication paradigm, control method, routing approach, mobile element, mobility pattern, network architecture, clustering attributes, protocol operation, path establishment, energy model, protocol objectives, and applications. The comparison between survey protocols is based on delay, network size, energy efficiency, and scalability. Furthermore, the drawbacks and advantages are also evaluated. Abbasian et al. [23] present the different data aggregation methods and protocols. The ground, multimedia, underwater, underground, and the human body all employ data aggregation in network applications in distinct ways. With the IoT scenario in WSNs, the study compares data aggregation and non-aggregation approaches. Individually emphasized data reduction approaches are used for reducing data volume size and communication costs.

A detailed review of existing techniques of distributed data aggregation problems under network settings is presented [24]. In this regard, the distributed data aggregation problems are divided into two main categories, communication and computation. For the communication category, routing protocols, network topologies, and all protocols are taken for aggregation process support, while the computation process is used to compute the aggregation function of algorithms. The review discusses the issues and recent approaches as well as advantages of data aggregation for underwater [25]. There are three types of underwater data aggregation techniques: cluster-based, non-cluster-based, and other strategies. These techniques are compared through metric performance. In the study [27], different data aggregation issues of existing studies are conferred and compared with the previous solutions as well for data aggregation issues. Also, it covers the comparative analysis of various data aggregation techniques based on delay, average energy consumption, redundancy, strategy, and traffic load. In the study [31], WSNs applications along with their classifications are explained. The research challenges and concerns are also explained. Mallick and Satpathy [28] claim that the WSNs structure, applications, and characteristics are the biggest challenges during implementation and designing. The WSNs applications are divided into two groups and are explained in detail. The study mentions that the main challenges and design goals in WSNs depend upon resource constraints, including data redundancy, storage, integration, QoS, and topologies.

Moreover, the survey is presented in the state-of-art of WSNs architecture, design and requirements, routing protocol, and its applications [29]. Further, for future designs of algorithms and protocols, some directions are also given. WSNs bring a tremendous change in agriculture monitoring by introducing smart farming, which has replaced traditional farming with its technology and applications. Farmers benefit from smart farming, such as water utilization, ease of agricultural land monitoring, and high yield. However, there are still issues with WSNs implementation in agriculture, but in the future, the entire agricultural system is automatic and sustainable owing to technologies like the internet of things, fog computing, and cloud computing that save time and resources [30]. In the same way, the big data challenges between wireless sensor networks and data aggregation strategies are reviewed and addressed [34]. The open issues, including the evolution of the IoTs, network architecture, real-time communications on fog computing, extensive, flexible framework, modelling, and simulation are also discussed. The big data concept is presented by integrating the dimension and tools and addressed issues.

Furthermore, a new classification for big data is created by WSNs requirements. In WSNs for big data aggregation, the different existing aggregation strategies are surveyed in detail. On the other hand, a comprehensive survey is used to investigate how big data is introduced in WSNs through its state of art research [33]. Moreover, for large-scale WSNs coverage, there are many challenges and opportunities. These challenges and opportunities are important to explore to increase the WSNs' lifetime efficiency. Moreover, in [32], Dai et al. present information on state-of-the-art big data and make recommendations for large-scale wireless networks to attain this aim. The authors concentrated on four phases of big data analytical (BDA) approaches: data acquisition, data pre-processing, data storage, and data analytics, rather than outlining the details of big data for WSNs. According to BDA, the life cycle categorized into four consecutive data stages (acquisition, pre-processing, storage, and analytics) is also presented and open research issues and future directions.

According to Pagar and Mehetre [36], the energy consumption is a basic challenge for WSNs applications. Different methods and techniques used to save energy, such as energy efficient sleep scheduling (EESS) algorithm for WSNs are discussed. Scheduling is also known as packet scheduling in WSNs by which packet schedules are managed, transmitted, and received from queue forms. The WSNs scheduling types are discussed in detail, along with their benefits and drawbacks. Bagaa et al. [37] focus on data aggregation scheduling algorithms for WSNs for energy efficiency, network lifetime, and accuracy. Data aggregation scheduling protocols are classified into two types according to waiting for

time nature, such as unslotted data aggregation scheduling protocols and slotted data aggregation scheduling protocols. Furthermore, each category is divided into subcategories based on its objectives. The unresolved challenges and directions for future study in data aggregation scheduling techniques are also explored.

There are several studies published on WSNs applications explained in detail in [4], [7], [44]–[46], [8], [11], [38]–[43]. Ali et al. [38] focus on real-time WSNs applications for criminal activities on borders, surveillance, traffic monitoring, water level, pressure, vehicular behaviour on roads, real-time intelligent observation of temperature, and remote monitoring of patients. WSNs types are according to different situations, and applications on modern society as well as the implementation concerning different fields are explained with strength, weakness, opportunities, and threat (SWOT) to identify the merit and demerits of WSN's real-life application. A specific application of WSNs for water pipeline monitoring is focused in [39]. The motivation of using WSNs for water pipeline monitoring is presented because being underground, the pipelines are supposed to phase different geological phenomena, including sinking, sliding, shaking, fracturing, and displacement of beds, which ultimately cause rupture and disruption in pipelines. A special application of WSNs in precision agriculture (PA) is presented by [41].

WSNs are used in agriculture to minimize labour. The technology uses wireless communications protocols in agriculture to identify communication distance and energy consumption. In agriculture, the energy harvesting technique for WSNs as well as energy-efficient techniques are used to solve the power consumption issues and identify more suitable methods. The existing techniques are compared, and their limitations are identified. However, recent studies in WSNs in PA are based on the Internet of Things (IoT) which compares and surveys some fields such as IoT end devices, IoT application layer, IoT platforms, type of sensors, and actuators. Premalatha and Prathap [43] highlight the underwater sensor fields as a new field for research, which is an easier way to get information from hostile areas. The study uses sensors to explore the underwater endangered species and discusses the approaches, challenges, and issues. Some researches [8], [40], [51] focus on reviews and surveys regarding WSNs applications for environmental monitoring systems as well. These applications are divided into two types: environmental monitoring systems and environmental monitoring applications. The existing environmental monitoring system techniques are compared and then the challenges and limitations of these techniques are identified. The challenges include power consumption, communication cost, scalability, remote management, and data transmission method.

Large-scale WSNs are randomly, densely deployed due to increase in data size for two reasons. First, the generated data at each sensor node are highly correlated and redundant due to the unchanged natural condition of the physical environment. There is a significant historical correlation among each consecutive data of a sensor node. For example, if temperature data readings are captured on sensor nodes every five seconds every day, the temperature readings may not change significantly. Due to this reason, it is not necessary to count the new reading at five-second intervals; otherwise, the previous reading matches the actual one. Second, when sensor nodes are randomly and densely deployed inside or close to the geographical phenomenon, a large volume of data size is generated and accessible for transmission as data is captured by all the sensor nodes in the area. In such a situation, all these nodes transfer a lot of redundant data. Another issue and challenge phased by WSNs is data redundancy. The similarity in the sensed data by a sensor is known as redundant data. As a result of the data redundancy process, sensor nodes waste most of their energy. However, to save energy, different methods and techniques are used. Generally, data redundancy has a huge influence on the quality of the data [49]. Curiac et al. [48] survey the impact of data redundancy by including and excluding the data redundancy from WSN. Two methodologies, which are fault tolerance and save operations for spatial and temporal data redundancy, are also discussed.

Energy saving is one of the main issues of WSNs, which is caused by data redundancy. Although redundancy is used to boost the data security in WSNs, it utilizes a lot of energy. Data redundancy reduction in WSNs might be the only solution to save sensor energy. In redundancy reduction, the removal of useless data ultimately improves storage efficiency and reduces the transmission cast. Some algorithms and techniques are surveyed and designed for data reduction, which can improve the lifetime of WSNs and increase the energy [47].

Therefore, existing studies state the surveys of WSNs and their application, design routing protocols, implementation designing, and specific real-time application such as underwater, underground, multimedia, and terrestrial application etc. Hence, this review article elaborates the classification of energy consumption by data redundancy in WSNs where it occurs and elaborates the parameters used for energy consumption in their classifications. It also includes the mathematical equations for energy consumption in WSNs. Some of the existing studies established on the classification of energy consumption by data redundancy in WSNs are detailed in section III.

TABLE I
SUMMARY OF EXISTING REVIEWS AND SURVEY FOR WSNS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

Refferences	Contributions	Data Classification			Performance evaluation metrics	Data prediction and nonprediction	Simulation parametric	Data Reductancy	Architecture	Data management	Challenges and future work
		At node level	At CH level	At sink level							
[4]	Security issues and military specificities in WSNs								✓		
[7]	New development and various issues WSNs in precision agriculture								✓		✓
[8]	WSNs environment monitoring systems such as indoor, outdoor, and greenhouse										
[11]	Wireless body area sensor networks										
[19]	Classical and swarm intelligence hierarchical routing protocol				✓		✓		✓		✓
[20]	Hierarchical-Based Routing Protocols for Mobile Wireless Sensor Networks								✓	✓	✓
[21]	Data aggregation protocols for structured and structure-free WSNs							✓	✓	✓	✓
[22]	Routing protocols such as: flat routing algorithms, hierarchical routing algorithms, and location-based routing algorithms				✓			✓	✓		
[23]	Data aggregation architecture in terrestrial, underground, underwater, and wireless body sensor networks in WSNs					✓		✓	✓		✓
[24]	Distributed Data Aggregation Algorithms such as structured, unstructured, and hybrid	✓			✓			✓	✓		✓
[25]	Data aggregation in underwater WSNs architecture such as Cluster-based techniques, and Non-cluster based techniques	✓	✓		✓			✓	✓	✓	✓
[26]	Data aggregation approaches challenges and security issues such as flat networks and hierarchical networks		✓					✓	✓	✓	✓
[27]	Issues of data aggregation methods and strategies such as centralized, in-network, tree-based and cluster-based		✓		✓			✓			✓
[28]	Characteristics, requirements, constraints, applications, and types challenges and design goals of WSNs								✓		✓
[29]	routing protocols and application for WSNs and their design goals and challenges							✓			
[30]	WSN used. Smart Farming d in agriculture and challenges involved in the deployment								✓		
[31]	Classification and types of WSNs										✓
[32]	Big data analytics includes data acquisition, data preprocessing, data storage, and data analytics							✓		✓	✓
[33]	Applications, Network System, and Data System							✓	✓		

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

[34]	big data concept, its dimensions, and analytics tools integrated						✓		✓		
[36]	Types of sleep Energy Efficient Sleep Scheduling inWSNs	✓	✓				✓	✓	✓		
[37]	Data Aggregation Scheduling Algorithms in WSNs	✓	✓				✓	✓	✓	✓	
[44]	WSNs Applications and Energy Efficient Routing Protocols for design									✓	
[45]	Applications of WSNs									✓	
[46]	WSNs: used in environmental monitoring and challenges									✓	
[38]	Types and Requirements of WSN applications										
[39]	WSNs for Water Pipeline Monitoring Applications						✓				
[41]	Energy-efficient WSNs for precision agriculture							✓	✓		
[43]	Underwater WSNs: processes, applications, and challenges						✓				
[47]	Data reduction in WSNs data-driven approaches are classified such as data acquisition, data reduction, in-network processing, data compression reduce and data prediction predicts	✓		✓		✓	✓		✓	✓	
[48]	Redundancy and its applications in WSNs for temporal and spatial redundancy	✓	✓		✓				✓	✓	
[49]	Impact of data redundancy when excluding and including in WSN	✓	✓		✓			✓		✓	✓
Current Review	Classification of Data Redundancy Reduction for Energy-Efficiency in WSNs	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

This article is a comprehensive review of WSNs in various levels of data reduction. We classify the data reduction methods and algorithms proposed in the literature for energy efficiency in WSNs. This classification is based on the most important objectives used for developing and solving energy constraints. The data reduction methods on WSNs are classified into three main levels: data reduction at the node level, at cluster head level, and the sink node level. To the best of our knowledge, a comparative study on the data reduces energy-efficient issues considering these classifications has not been conducted yet. However, previous survey and review studies that focus on the specific areas on wireless sensor networks include military, agriculture, environmental monitoring, and wireless body area network, which also handled the architecture along with the limitations and challenges [13]–[18]. Various existing studies investigate based on routing protocols for data aggregation of types and applications [19]-[27]. Some of the literature focuses only on specific types along with their challenges and issues [28]- [31]. Big data generated by WSNs [32]-[34] are analyzed, including application and management of data. [36]-[38] provide a detailed explanation of scheduling sleep algorithms for energy-efficient in WSNs. Data reduction schemes' impact is investigated by including and excluding in WSNs, including at nodes level and aggregator level data reduction as shown in Table I. Hence, this article explored state-of-the-art strategies of data redundancy reduction and classified into three categories: node level, cluster head level, and the sink level, as shown in Figure 2. The purpose of these classifications is built to the basis for future researchers in WSNs. These classifications are based the illustrated their advantages and disadvantages, and also discribe various presented methods and schemes based on some important parameters regarding cluster-based architecture, such as percentage of data after applying aggregation phase, percentage of data sets sent to the CH, duplicate sets of data, sampling rate adaptation, energy consumption at the node level, the lifetime of a sensor node, data gathered a number ratio sent data set, gathered data readings two consecutive periods, and so on. As well, a side-by-side comparison of all discussed strategies is presented, and some suggested future work are addressed.

III. DATA REDUNDANCY REDUCTION

This section describes the issues and problems related to data redundancy reduction in WSNs. Data redundancy is the repetition of a single entity more than two times. It is also known as similarity or the exact value [52]. Redundancy is found during the sensing process when sensor nodes sense a physical object. Due to some constraints in WSNs, there are approximate issues of redundancy, typically in hostile or harsh areas where the sensors cannot be replaced or recharged [53]. Conversely, another issue related to WSNs is big data, as thousands of sensors collect and compile the data in a wide

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

area and produces a significant portion of big data. Nowadays, WSNs are one of the main sources of big data in IoT because the sensors sense a huge amount of data in a minute before sending it to a base station. However, Big data processing is quite complex to manage [54]. The issues and challenges of data redundancy are stated in the section below, and different levels of redundancy in cluster-based architecture for WSNs are described. The main data challenges in WSNs are classified into different categories such as clustering [55], security, processing, data analysis, data aggregation, and energy saving.

In addition, big data is classified into two major areas: network systems and data systems [32]. Usually, a network system delivers censored data which converts to an extensive form of data. Though many resources are required to save data, a huge amount of energy is also required for its processes, sensing, and transmission. After the network system delivers censored data to an extensive data network, the data system processes the data [56]. Generally, the data is received by data network in a redundant form for analysis, which causes multiple issues during data processing and analysis [57]. Another issue of WSNs is battery limitation, as the sensor lifetime relies on its battery. The sensor uses battery power in several operations with different quantities [58]. Battery power is not only important for sensors life but is also important for sensing, collecting, and communicating data. When the sensor uses redundant or raw data in the operations above, the battery`s energy depletes quickly [59]. As battery power saving is one of the most challenging issues, reducing redundancy could help save battery power. Moreover, data redundancy in WSNs raises some other issues, such as high workload, conjunction in-network, and high transmission cost.

IV. CLASSIFICATION OF DATA REDUNDANCY REDUCTION

In this review article, the data redundancy reduction schemes are classified into three levels: the node level, the cluster head level, and the sink level, as shown in Figure 2. These data redundancy reduction strategies are based on the factors that have been employed in the estimate of performance in numerous researches.

WSNs are based on cluster-based architecture, and it is possible to identify the particular levels where the redundant data is formed [60]. Usually, the cluster-based architecture deploys in a large network which is further divided into small cluster groups [61]; each group has its cluster head and member node [62][63] and each cluster is supposed to send data to the sink node [64]

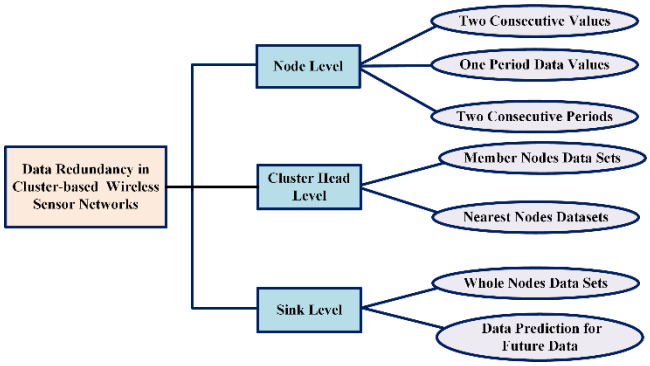

FIGURE 2. Classification of data redundancy reduction in WSNs.

In addition, redundant data is found at three different levels such as at node level, cluster head level, sink level. First, while the sensor is sensing data at its fixed time, where there are no dynamic changes in the environment, the data is redundant at the node level. Second, when the cluster head collects data from its member nodes, there is a big chance that the data is redundant as nodes are randomly distributed, i.e., some nodes might be close to each other or there are no dynamic changes in the environment. Lastly, nodes far away from the sink send their sensed data to nodes near the sink; thus, big data or redundant data are formed as the sensor near the sink has data of nodes away from the sink plus their own sensed data. When nodes near the sink node must forward big or redundant data to the sink, plenty of energy is used, thus, to conserve the energy used by the sensors near the sink, the sink node invents some mechanisms. Recently several researchers are worked to find the solution to preserve the battery power and proposed different methods and techniques. Below are the work of the recent researchers and some issues with their work. The taxonomy of various data redundancy schemes classification is presented in Figure 2. The details of these classification itemized, based on existing techniques used for energy efficiency in WSNs, are shown in Table II, VI, and VIII.

A. DATA REDUNDANCY REDUCTION AT NODE LEVEL

This section explains the existing studies of the data redundancy reduction schemes and algorithms used at the node level for the WSNs. However, the current techniques are analyzed and considered the necessary parameters for the emergence of data redundancy in WSNs. Table I presents the various problems of data redundancy reduction at the node level with proposed schemes by combining the models and strategies. The contributions and limitations in WSNs are also presented.

TABLE II
EXISTING STUDIES ARE BASED ON DATA REDUNDANCY REDUCTION SCHEMES AT NODE LEVEL

References	Schemes	Problems Addresses	Proposed Models / Strategies	Improvements/ Enhancements	Limitations/ Weaknesses
[65]	Aggregation and transmission protocol (ATP)	Conserving energy, eliminating data redundancy, and reducing communication	1. Aggregation phase using a similar function and measure frequency. 2. Transmission phase us one-way ANOVA model and fisher test	Energy consumption, data quality	Focus only on two consecutive readings and assumptions
[66]	Data collecting and aggregation with discerning transmission (DCADT) technique	Redundancy in the collected data to sending it to the sink	1. Data capturing for dimensionality reduction, the SAX symbolic method, and piecewise aggregate approximation (PAA) with adaptive piecewise constant approximation (APCA) DTW are used for two periods. 2. Selective transmission using the notification (NOTIFY_ PKT) or (MEASURE_ PKT) 3. Changes in sampling rate	Consumed energy and data accuracy	Complex computation and high memory used
[67]	The least number of bits	Reducing energy consumption is one of the main issues in WSNs	1. Sending the difference between new reading and previous reading for less size by least bits number	Energy consumption and extends lifetime of sensor network	Single value comparison due to this high computational rate
[68]	Min and max stratum	Intra-temporal and inter-spatial correlation data generated by the dense distribution of sensor nodes	1. First, comparison between predefined means values with new capture data 2. Second, obtained from the previous step value compare either it is min or mix with predefined stratum values	Control packet collision, communication cost network, congestion, and enhanced network lifetime	Approximating the mean and variance are additional difficult than for simple random sampling
[69]	Energy-efficient adaptive distributed data collection method (EADiDaC)	A crucial issue in PSNs is the continuous collection of a large volume of data	1. First, data collection adaptively 2. Second, dimensionality reduction using adaptive piecewise constant approximation (APCA) technique 3. Third, frequency reduction using the symbolic aggregate approximation (SAX) approach. 4. Lastly, sampling rate adaptation is based on dynamic time warping (DTW) similarity.	Preserve the energy at the sensor nodes and extend the PSNs lifetime	There is overly complex and require huge processing
[70]	Kruskal–Wallis test	Data reduction for WSNs	1. First, eliminate similar reading from the vector by similar function and measures redundancy used reading weight 2. Second, after ending each period searches for redundancy of new gather value if it is redundant just add its weight	Reducing the size of data transmitted over the network and thus saving energy	Do not apply correlation between neighboring sensor nod

A cluster architecture that contains a cluster head and member nodes are shown in Figure 3. Moreover, each sensor node collects the data periodically. This periodic data is further divided into small intervals known as slots, and every slot senses unique data. However, the slot has a short, fixed period to collect the data. During this period, if the physical environment shows change (rapid or slow), then there are chances of similar or redundant data coverage [71]. Ultimately, the redundant data consumes a lot of energy at each sensor due to periodically sensing data. Even though

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

many researchers have worked on energy conservation caused by data redundancy and have come up with some mechanisms and algorithms, energy issues continue to require attention [72].

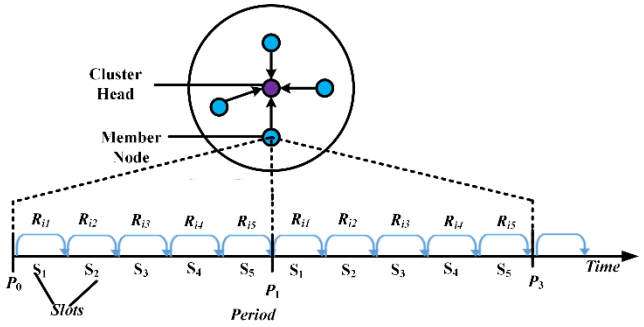

FIGURE 3. Process of data collection at the node level.

In periodic sensor networks (PSNs), the network lifetime is based on the energy of each sensor node. Hassan et al. [65] propose an aggregation and transmission protocol (ATP) construction for each sensor node to reduce data transmission and ultimately preserve energy. In PSNs, each node gathers sense data in vector form. Each period is divided into fixed time slots. At the node level, captured data is a form of vector. The period is divided into equal fixed time slots. However, when the time interval is small, the node may have collected redundant or closely similar data due to the unintentional changes in the physical monitoring object. The study proposes the node level data redundancy reduction based on two phases, aggregation and transmission, to reduce redundant data. The aggregation phase at each node aims to reduce redundancy or size from raw data which increases the energy power of each sensor node. Moreover, the aggregation phase is classified into two functions. The first function is to identify the similarities between two measurements with application defined threshold. Furthermore, if these two measurements are similar, then their function adds 1 to the first measurement and discards the second measurement. If these two values are not similar, then it considers the new value. After the aggregation phase, a node compiles the data sets observed by each sensor and decides whether to send these aggregated data to the cluster head (CH) for the transmission phase. Once the data is sent to the transmission phase, it is checked again to find out the data redundancy between two successive periods. Every period has its own data sets and time interval frequency measures. A statistical model is used for data redundancy reduction between two periods by using the analysis of variance (ANOVA) one-way model and Fisher test. Furthermore, the Fisher test shows whether all prior and new periods are similar. During the transmission phase, each sensor node uses the Fisher test to calculate the variance between the prior period of data sets and the current data sets. It checks if and only if the variance is not significant, then it discards the new period data sets and just sends a notification to the CH. The notification packet is empty, showing that the new dataset and the previous dataset are equal to avoid sending redundant sets to decrease the power consumption. However, while the ANOVA model identifies redundancy between two data sets, different data sets still need to be measured. Meanwhile, the Fisher test is usually needed when the small data size has high computational issues.

The data collecting and aggregation by discerning transmission technique (DCADT) are proposed in [66] to increase the lifetime of PSNs. DCADT finds correlations among the collected data in every sensor by sampling rate in a dynamic way. This technique works in rounds, with each round divided into two periods and each period consisting of four phases: data aggregation, data gathering, frequency adjustment, and selective transmission. To gather samples of data, every sensor node uses dynamic time warping (DTW) to measure distance. Both data gathering and data transmission are used for data sampling. The usage of both data transmission and data gathering is used for data sampling for the removal of redundant data within sensors while sending data to the base station. Before transmission, the redundant data first aggregate data. In the data-gathering phase to get adoptive data, a sampling rate is fixed in every sensor node. Second, in the data aggregation phase, the DGAST protocol uses a symbolic aggregate approximation (SAX) algorithm to eliminate redundant data from temperature reading before transferring to the CH. This phase is further distributed into two more stages. The first stage, dimensionality reduction, and adaptive piecewise constant approximation (APCA) approach use different lengths but constant values for the segment. In the second phase, the SAX method is used to reduce reading repetition by making a table of readings from segments representing symbols for breakpoints specification. This phase is a decision-making phase that decides whether or not to forward the data between two complete periods to the CH. Suppose there is redundant data between the two complete periods. In that case, it only sends notification packets, while if there is no redundancy in data, then it sends new period's data to the CH. The final phase is the adoption of sampling rate, which finds redundant reading percentage between two consecutive periods per round for new redundant rate. For this, the DTW distance base is adopted for the measurement using a similarities function redundancy.

The study presents methods to conserve energy and reduce data size for low transmission cost in every sensor [67]. It finds the differences in new sensed reading and last reading value and then uses the least number bit, which is used for transmission. Rather than sending all values to the start-up phase, every sensor sends its first reading to CH and saves it in its memory. When CH receives the first reading from every sensor node in cluster reading, it saves it in memory. Next, in the data collection phase, sensor node calculates the differences between the new and previous reading. The differences are called the least number of bits. It assumes that if the last sensor reading is 25°C (110012), then the new temperature in this reading is increased by 5°C. The new

reading then becomes 30°C (111102). Then the difference (30-25=5°C) (1012) is computed. The total number of bits is decreased and evaluated by using bits that save energy and reduce the data size.

Similarly, the study proposes an energy-efficient and computational lightweight aggregation technique. The main advantage of this technique is node processing cost where it reduces the amount of data being sent to the base station which ultimately controls the data conjunction. In this scheme, every cluster member node places data according to a stratum that exists in every sensor node within the buffer. Seven stratums are used for ranging the temperature value from 26^0C to 32.99^0C. These ranges are common for all temperature readings captured by every sensor node. These seven strata are based on previously acquired historical data. In this approach, there are two steps. The values of new readings are compared to the stratum mean value in the first phase. The received value from the last stratum is compared to either the minimum or maximum value for a particular stratum in the second phase. However, each stratum is divided into the minimum or maximum values, amongst which each sensor node compares its values from minimum or maximum and then forwards it to cluster head in a predefined time interval [68].

Al-Qurabat et al. [69] propose an energy-efficient adaptive distributed data collection method (EADiDaC) for data aggregation, which collects data periodically to increase sensors lifetime. This method is divided into cycles and four different stages are built in each cycle. The first stage of the cycle is data collection, in which the process shows how the sensor collects the data in a network and its transmission process to the base station. Each cycle is divided into two periods and each period is divided into slots. EADiDaC method collects the sensor readings in time-series form, and it is called temperature readings. The redundancy in these temperature readings increases only under two conditions; when the time slot decreases or the changes in the area's physical environment are slow. The second stage is dimensionality reduction, where the adaptive piecewise constant approximation (APCA) technique is used to decrease dimensionality. There is a period fixed to measure sensor readings at this stage to reduce the size of data by dimensionality reduction technique using APCA technique. The study presents some modifications in APCA i.e., the length of the segment is not fixed and with the help of user-specified reconstruction error, the adoption base is set. Then to build different segments, sliding windows through user-specified reconstruction errors are used. The third stage is frequency reduction, which is reduced with the help of SAX. By using this method, redundancy from temperature readings before sending it to the base station is reduced. The EADiDaC method is used to build a reduced vector by imposing a variable length. Now the temperature readings are divided into an unspecified number of segments by using a sliding window that varies in length. Each segment calculates a mean which is called length. The mean values are turned into symbols and the alphabet to get a fixed size. It puts a breakpoint on the symbols which are predefined in a Table form. Before converting to APCA, the mean values are converted into symbols in which redundant symbols are also included. To remove the redundant symbols EADiDaC method uses a function to find the redundancy between the symbols before sending them to the base station. The final stage is sampling rate adoption which is based on dynamic time warping (DTW). This method finds the redundancy percentage from the temperature reading in a period and decides the sampling rate. Initially, it finds data similarities between two periods. At the end of each period, the EADiDaC method changes APCA, and it selects a different number of a segment whose length is different in every period. EADiDaC method uses similar functions to find redundancy between every two APCA temperature readings and then it verifies the number of data similarities in a period at each cycle. Thus, periods have different lengths in each cycle, so the reading percentage is calculated per cycle. Suppose the redundancy percentage is high in the readings at different periods. In that case, it means that there are only minor changes in the environment, but if the redundancy percentage is less, it means there are minimal changes in the environment.

In the study [70], an online data reduction method is proposed in which the sensing rate of sensor node depends upon data variances Kruskal–Wallis test. The Kruskal–Wallis test is built in every node, so it reduces redundant data. The first phase at the node level is known as acquisition. It uses three periods in a cycle where data is organized in an order form in a table and an ordered rank is fixed for every reading. If the received reading is redundant, which is named as tied, a mean of the tied readings is calculated. Additionally, a threshold value of an application risk level is taken. In every round, each node decides after differentiating between risk level and sampling value that either the sampling rate should increase or decrease. To find the redundancy level, a behavioral function is used to identify the differences between the crucial value threshold and sensing data. Furthermore, suppose the values are higher than the risk level. In that case, it is labeled as one (1), else if it is lesser than risk level, then it is labeled as zero (0), and these values are placed in value R. There are chances of data redundancy in R and similar function is used to identify it. Three conditions are used to identify similar data. In the first condition, if the result is (0) zero after comparison, it labels as one (1). In the second condition, if it is lesser than the threshold value, then it is referred to as redundant readings. In the third condition, weight is fixed on it, and if it is redundant, then the weight generates a vector. Finally, the sensor sends a set of readings with its weight to the sink node.

In Table II, simulation parametric values are based on a threshold value, measures readings, E_{elec}, ßamp, K, MINSAMP, data, and field; simulator and sensor at the node level are described. Most of the researchers used the 0.03 and 0.05 threshold values with 50 nJ/bit on the intel Berkeley

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

research lab and due to using a small threshold value, the measures readings range is also less like 20-100. Some of the researchers used the large threshold value like 0.07 and 0.1 with large measures readings ranged from 100-2000. Table III shows the existing proposed methods and schemes with the benchmark's methods and schemes used for comparison. Also, it mentions the various performance comparison parameters for evaluating the existing schemes and methods for data redundancy reduction in WSNs.

TABLE III.
SIMULATION PARAMETRIC VALUES AT THE NODE-LEVEL

References	Schemes/ Methods	Threshold Values	Measures Readings	E_{elec}	ßamp	K	$MINS_{AMP}$	Data and Field	Simulator and Sensor
[65]	ATP	0.03, 0.05 and 0.07	20, 50, 100			2		Intel Berkeley Research Lab and temperature	Custom Java Based and Mica2Dot
[73]	DaT	0.03, 0.05, and 0.07	20, 50, and 100	50 nJ/bit	100 pJ/bit/m2	ρ/2		Intel Berkeley Lab	OMNeT++ and Mica2dot
[74]	TLDA	0.03, 0.05, 0.07, and 0.1	200, 500, 1000 and 2000	50 nJ/bit	100 pJ/bit/m2	2		Intel Berkeley Research Lab and temperature	OMNeT++ and Mica2dot
[66]	DGAST	0.07, 0.1, 0.2, 0.03 and 0.05	20, 50, 100, and 200	50 nJ/bit	100 pJ/bit/m2	2	20, 40, and 60	Intel Berkeley Research Lab and temperature	OMNeT++ and Mica2dot
[75]	Divide-and-Conquer Algorithm		50, 100 and 200					Intel Berkeley Research Lab	Mica2dot
[76]	TTDR		20, 50 and 100 sensed data	50 nJ/bit	100 pJ/bit/m2				OMNeT++
[77]	ESTS and Reliable-(RESTS)	0.03, 0.05, 0.07 and 0.1	200, 500 and 1000					Intel Berkeley Research Lab	Mica2dot
[67]	(Exact) and difference value (Diff) approaches		43086, 14246 and 3213					Le Gènepi, Le Borien, and PlaineMorte Switzerland	Java event-driven
[78]	AAS	0.05, 0.07, 0.1, 0.2, 0.3, 0.4 and 0.5	50, 100, 150, and 200.					Intel Berkeley Research Lab	Mica2dot
[79]	EK-Means	0.0 0 05 to 0.025	256 to 2048	50 nJ/bit	100 pJ/bit/m2	5 and 50.		Argo project	Java based Simulator
[68]	Min and Max Stratum		1500 packets		1.064 µ joule	Packet size 128 bytes	Sampling rate is 1 packet/se		
[80]	Dual Prediction (DP)	0.5 ºC	Data blocks sizes (m, 8) and (m, 10)			data buffer size N = 10		Intel Berkeley Research Lab	Mica2dot
[81]	REDA			50.10−9J	100pJ/bits/ m2				MATLAB
[69]	EADiDaC	0.07, 0.1and 0.2	20, 50, 100 and 200	50 nJ/bit	100 pJ/bit/m2		20, 40 and 60	Intel Berkeley Research Lab	Mica2dot OMNeT++
[82]	KNN		50					Real sensor laboratory	Crossbow telosb motes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

[83]	Two-level data reduction approach	0.4, 0.5, 0.6 and 0.7	100, 200, 500 and 1000					Intel Berkeley Research Lab, and real sensor nodes laboratory	Crossbow telosb motes and Mica2dot
[70]	Kruskal–Wallis Test	0.05, 0.01 and 0.025	50						
[84]	KAB, Euclidean, cosine distance and PFF	0.03, 0.05, 0.07, 0.1, 0.35, 0.4, 0.45 and 0.5	200, 500 and 1000.					Intel Berkeley Research Lab	Mica2dot and Java based simulator
[85]	Prediction model ECR	0.5 and 1	500	50 nJ/s for 1-bit	150 nJ/s for 1-bit, 10 m	10			NS-2.34
[86]	Extended (DPS)								International Airport Tlemcen (Algeria) MATLAB

Note: E_{elec} is the energy consume on the transmitter and the receiver, ßamp, is a transmitter amplifier, K means total periods, $MINS_{AMP}$ means a minimum sample rate, nJ/bit the circuitry is to path the source or receiver, pJ/bit/m2 is on the transmitter amplifier.

TABLE IV.
PERFORMANCE COMPARATIVE PARAMETER'S FOR SAVE ENERGY AT THE NODE LEVEL

References	Schemes/methods	Compared methods	Performance Comparative Parameter's: Percentage of data after applying aggregation phase	Percentage of data sets sent to the CH after applying transmission phase	Data accuracy / percentage of lost data	Energy analysis	Percentage of data after aggregation every node	Duplicate sets of data	Sampling rate adaptation	Energy consumption at the node level	Lifetime of sensor node	Data gathered number	Ratio sent data set	Gathered data readings two consecutive periods	Transmitted data number to CH	Percentage of energy-saving every node
[65]	ATP	PFF	✓	✓	✓	✓										
[73]	DaT	ATP and PFF			✓	✓	✓									
[74]	TLDA	[71] and PFF			✓		✓	✓								
[66]	DGAST	PFF, and Harb		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
[75]	sensor tier: divide-and-conquer algorithm	PFF													✓	
[76]	TTDR	ATP and PFF			✓	✓	✓								✓	
[77]	ESTS and reliable-RESTS	ESTS and RESTS	✓		✓		✓				✓					
[67]	Exact value (Exact) and Difference value (Diff)	(Exact) and difference value (Diff) approaches								✓					✓	
[78]	AAS	Naïve approach												✓		

[79]	EK-Means	Naïve approach				✓		✓						
[68]	aggregation technique			✓	✓	✓		✓	✓					
[80]	Dual Prediction (DP) with TRP and MSE	OSSLMS, NNs, and LSTMs and DP										✓		
[81]	REDA	ESPDA and SRDA	✓							✓			✓	
[69]	EADiDaC	PFF and Harb et al. (2016)			✓	✓			✓	✓	✓			
[82]	KNN	S-LEC			✓						✓			
[83]	two-level data reduction	S-LEC		✓								✓		
[70]	Kruskal–Wallis Test	Bartlett test and S-LEC											✓	
[84]	With aggregation	Without aggregation				✓	✓							
[85]	Prediction Model ECR	P-DPA and OSSLMS			✓	✓								
[86]	Extended (DPS)	DPS			✓							✓		

B. DATA REDUNDANCY REDUCTION AT THE CLUSTER-HEAD LEVEL

This section describes the current studies of the data redundancy reduction schemes and algorithms at the node to the CH level for the WSNs. Although the existing techniques are examined and measured, the used parameters are essential for emergence of data redundancy in WSNs. Table V highlights that some of the current existing methods problems are addressed with their progress models, and proposed energy efficiency schemes to reduce data at the CH level for WSNs. Table V also deliberates the contributions and boundaries of existing methods.

TABLE V.
EXISTING STUDIES ARE BASED ON DATA REDUNDANCY SCHEMES AT THE CH LEVEL

References	Schemes	Problems identify	Proposed strategies	Improvements/ enhancements	Limitations/ weaknesses
[73]	1. Data transmission (DAT) 2. Energy-efficient two-layer data transmission reduction (ETDTR)	Reducing the data volume transmitted to the final base station or sink	1. Node level: KNN modified k-nearest neighbor) algorithm 2. CH level: grouping vectors into sets according to their length	Data reduction rate, energy consumption, and lost data	It does not work with big data sets
[74]	Two level data aggregation (TLDA)	1. Reduce the amount of data captured by each sensor individually. 2. Identifying closely matching sets, integrating replica readings, and transferring aggregated data to the sink	1. Node level: data collection, segments generated by sliding window and adaptive piecewise constant approximation (APCA) technique used for data aggregate 2. CH level: finding similar data sets by using a hash function and symbolic aggregate approximation (SAX)	Reduce sensed data, reduce energy consumption, and data quality extend the lifetime of the PSN while retaining.	It is based on the mean value which is causes to miss some important information

[76]	Two-tier data reduction (TTDR)	Sensor nodes create data and transmit it to a gateway (GW), which consumes high energy and storage.	1. Node level: delta encoding with differences between two readings and run-length encoding add repeated value number 2. CH: minimum description length (MDL) set hypothesis by discrete normalization, description, and conditional length	Control data transmission and energy	length of a vector is fixed of all reading which causes data loss
[77]	Reliable-ESTS (RESTS)	Densely distribution and the dynamic objects offer high correlation among sensor nodes	1. Node level: Euclidean distance for search sensor data similarities 2. CH: Geographical closeness neighbouring nodes calculated by Euclidean distances	Saving energy consumption, network lifetime, and coverage of the monitored range	It does not consider adaptive sample rate
[78]	Aggregation and adaptive sampling (AAS)	Two major challenges: the huge amount of collected data, and the absence of a replenishable source of sensor energy	1. Node level: Similarities detect by similar function between two values reading redundant data for measures frequency 2. CH level: Physical distance between two sensors is determined by Euclidean distance and adjust sampling rate for the next period by a correlation between sensors spatial temporal	Saving sensor energy, data accuracy and coverage network	When a similar sampling rate operate between two sensor nodes collision occur between packets
[79]	EK-Means	Sensor nodes is huge volume of data collect	1. Node level: Every data determined two consecutive points by Euclidean distance and two similar vectors found by similar threshold between two measures 2. CH level: the difference between two equal vectors data sets determined by Euclidean distance if two different length vectors added new data point to equal them	Less energy consumption	Euclidean distance does not support a different length of data vectors
[81]	Redundancy elimination data aggregation (REDA)	Redundancy elimination data	1. Node level: received lookup table from CH and compared sensed value 2. Generated lookup table for their member nodes	Energy consumption and bandwidth occupancy	Only for small area networks
[82]	K-Nearest neighboring (KNN)	1. The first challenge is big data collection 2. The second challenge is limited sensor energy	1. Node level: big raw data reduce by Pearson's coefficient metric 2. CH level: remove data redundancy collected by neighbouring nodes with k-nearest neighbouring	Energy consumption and data accuracy	Very slow algorithm for large data requires high memory and
[83]	Two-Level data reduction	WSNs are one of the big data givers, someplace data are being gathered at anomalously	1. Node level: selective reading sends by Pearson coefficient 2. CH level: the combination among Euclidean distance and classic k=mean for reducing similarities between each cluster member nodes data sets	Improving Network Lifetime	When data has higher values, it can be easily misinterpreted
[84]	KAB, Euclidean Distance, Cosine Distance, and PFF	The creation of a high volume of big data is WSNs	1. Node level: similar function and weighted cardinality 2. CH level: one-way ANOVA model with statistical tests and the distance functions,	Energy consumption, data latency, and accuracy	Geographically sensor nodes distance does not consider

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

[85]	Extended Cosine Regression (ECR)	Data aggregation to reduces duplicate data transmission	1. Node level: two vectors: one is the actual data vector (adv) and the other is the predicted data vector (PDV_{SN}) 2. CH: CH is built PDVCH	Expand the network's lifetime and accuracy	Data accuracy is dependent on threshold

In WSNs, scalability improves with energy efficiency by preparing a hierarchical design. The typical architecture technique is called clustering, which replaces single-hop transmission with multiple hop transmission for improved scalability. If the clustered-based architecture is considered in a periodic sensor network, then the network is supposed to be categorized into different clusters. Each cluster has a CH that receives its member node data and takes responsibility for further transmission as shown in Figure 4. Whenever its member nodes sense data and the nature of the environment is constant, the data have redundancy. The data is aggregated and sent to the CH by the member node. Subsequently, the CH receives different data sets of each member node. Due to short geographical distance and environmental changes, there is a greater chance of having redundant data. Among the member nodes in a cluster, this redundant data contributes to high traffic, high workload, high memory loss, and rapid depleting sensors' battery [68]. However, several researchers have proposed different techniques and methods to resolve these problems, which are discussed in Table III.

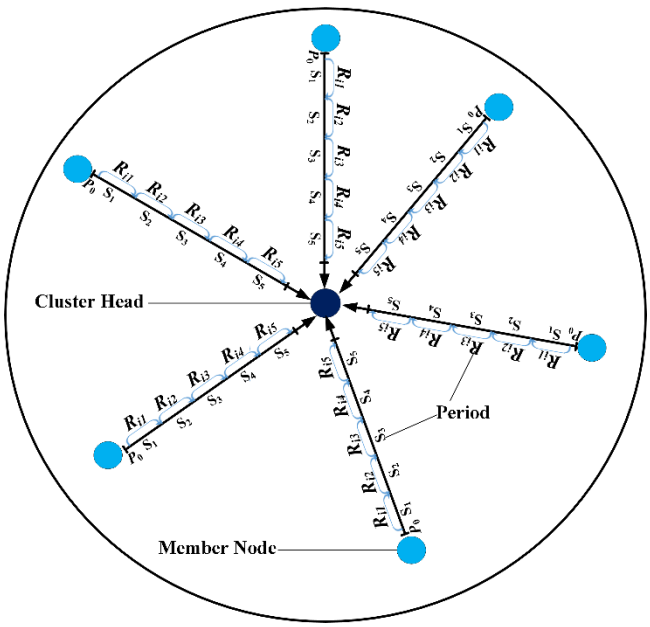

FIGURE 4. Data sets were collected at the CH level from the member nodes.

For the data redundancy at the CH level, Ying et al. [87] offer an energy-efficient data collection technique with cluster-based WSNs to identify spatial-temporal correlation data. To detect temporal redundant data, a dual prediction technique is applied to reduce intra-cluster transmission. Then, hybrid-compress sensing is built in clusters to identify spatial redundant data among the sensor nodes for inter-cluster transmission. The prediction model also presents an error threshold selection scheme to enhance energy consumption and accuracy recovery. In the prediction method, CH sends a forecasted value to its all member nodes. To decrease intra-cluster transmission, when cluster member nodes received the forecast value from the CH, the observed and forecasted values are compared within a special threshold. If there is a large error between forecast value and observed value, then it ends at a threshold. Cluster member sends its observed value to the CH, but the CH considers a predicted value. Afterward, the inter-cluster transmission starts its work where the CH gathers the data from all member nodes in the cluster, and it transmits if it is lesser than a threshold value (M). If it is more than M, it aggregates data using compressing sense. Finally, the sink node uses the CS recovery method to restore all of the data to its original form.

Moreover, Idrees et al. [73] provide an expanded version of the KNN (Modified k-Nearest Neighbour) method at the sensor node level to minimize energy usage in WSNs. In addition, to extend sensor lifespan, it uses DaT protocol that is divided into two stages: data categorization and data transmission in each interval. Modified k-Nearest Neighbours is used to classify the obtained data into multiple groups while instead of delivering all data, the DaT protocol picks the best illustrative data from each class and transmits it to the sink [88][87]. Likewise, similar classes are combined into a single class. Finally, the best representative readings of all classes with reduced vector are transferred to the sink at the data transmission stage. The KNN protocol is unfair by the value of K and has high computational complexity and less memory for large data sets.

Al-Qurabat et al. [74] propose two-level data aggregation (TLDA) protocol for extending the lifespan of WSNs. At the node level, the data aggregation is constructed at an initial stage. Different lengths of segments for data are created using slide windows, and aggregate data is collected using the adaptive piecewise constant approximation (APCA) technique, which helps to decrease data collection size at each node level. The second level is created at aggregators or CH. The aggregator gathers a set of data by chaining a hash table together with the SAX method. At this point, it searches and decreases redundant sets by merging the redundant readings. Henceforth, it transmits the aggregate data to the sink node. TLDA protocol enhances the PSN lifetime, decreases redundancy data, saves node power, and maintains accuracy.

In a similar way, Al-Qurabat et al. [76] propose a two-tier data reduction (TTDR) technique that works in two-tier

networks such as sensor nodes and gateway. The first tier is at the node level, where a straightforward data compression strategy is applied ideal for node restrictions. For determining temporal correlation in sensor node data, delta encoding followed by run-length encoding (RLE) is utilized. The purpose of delta encoding is to reduce the dynamic range in a data set. Delta encoding finds a difference between current and previous sensor data, then compresses it, and sends it to the CH. However, the redundant data is also included at this stage and to remove the redundant data, RLE encoding method is used to compress more data. In the second tier, when the period ends, it sends all gathered data to the aggregator or CH. Subsequently, when the aggregator receives the data set from its member nodes in the cluster, it finds the data redundancy and compresses data size before sending it to the sink by minimum description length (MDL). Furthermore, there is a hypothesis through (MDL) at the CH level, and each cluster follows it. Ultimately, the CH sends the difference between data and hypothesis to the sink. By this procedure, they identify the redundant data sets and non-redundant data sets and send them to the sink. According to (MDL), if two data sets are similar, then it compresses them; otherwise, they are considered different.

A Reliable-ESTS (RESTS) in spatio-temporal scheduling (ESTS) technique is proposed on spatio-temporal correlation of sensor data and ultimately increases the lifetime of PSNs [77]. In addition, a new model is introduced to use Euclidean distance to search periodical correlation between spatio-temporal data of nodes neighbours. Thus, at this level, a temporal correlation at the sensor node level is identified. There is redundancy at each consecutive reading and to check this redundancy, a local temporal correlation is suggested. By using local temporal correlation, redundancy is found between two consecutive readings; if the redundancy is found, then only one reading is used and the other discarded, while in a replace of the discarded value and keep weighted values. Further, the CH receives data sets with the weights of each member node at the end of each period. Before delivering data to the sink, the CH seeks a spatio-temporal correlation for each node, removing duplicated data among neighbouring nodes. By using the Euclidean distance, the CH finds the distance between sensor nodes with the help of a specific threshold. First, if some nodes are nearer to the predefined threshold distance, then it considers them in a spatial correlation. Moreover, the CH checks the temporal correlation of the nodes which are near to each other. After finding the spatial-temporal correlation, the CH applies a scheduling algorithm such as sleep/active sensors mode. Second, suppose two nodes are closer to each other and produce similar data but the remaining energy level amongst one of them is weak. In that case, the algorithm keeps the node at an active mode with more energy, and the other with less energy is turned into sleep mode. Now, the active node senses collect data and send it to the sink.

For PSN applications, [78] recommends a novel adaptive sampling methodology. Aggregation and adoption are two steps of this method. The primary goal of the first step is to limit the amount of data acquired by the node. In this stage, the data is collected in a vector form. Redundant data from vectors is removed by a data similar function. A new adaptive sampling technique identifies redundancy between two values by a special threshold that is given based on an application. If it identifies the two as similar, then their similar function is equal to one (1); otherwise, they are recognized differently. In the case of constantly similar data, it adds one (1) in their frequency through the frequency measuring function. In the end, a set of data is collected by each node and transferred to the sink node. In the second stage, the CH receives a set of data and its frequency weight. Subsequently, the CH found a special correlation between sensor nodes by using two techniques including closer geographical sensor nodes and highly spatial correlation between collected data by their member nodes. However, to find geographical distance it takes the help of Euclidean distance between two nodes. On the other hand, to find spatial correlation, it uses some functions (overlap coefficient, Jaccard similarity, and cosine similarity) on a dataset of both nodes and data sets. To find similarity between both data sets and sensor nodes, two different values, zero (0) and one (1) are taken, whereas zero (0) is considered as different data sets and one (1) is considered as redundant data sets. Finding spatio-temporal correlation between sensors helps to decrease energy consumption between sensors.

By using the EK-means, a new data processing strategy is given that reduces data transfer without compromising data integrity. This strategy works in two stages. Data redundancy is removed at the node level in the first stage, using a linear interpolation function and the Euclidean distance methodology. First, a point is captured as a vector and then two vectors are considered to identify redundancy between them if these two vectors of measures are with constant data sets size. Suppose the measurements of these two vectors are similar in the threshold value. In that case, the quantity of the data similarities via threshold value concerning the two vectors is created on Euclidean distance. Moreover, every node finds its representative point at each period. However, to find the representative point, a starting point and an endpoint are selected. The use of Euclidean distance calculates the distance between starting point to the endpoint. After the collection of a set of representative points at one period, the data is sent to the CH. In the second step, the aggregator receives the dataset of representative points by member nodes. At this level, the main task is to check the data redundancy among data sets between member nodes and to reduce the data amount before sending it to the sink node. A CH approach introduced in the K-mean algorithm is enhanced to make a cluster for the data sets to accomplish this task. The new approach decreases the data latency. In the EK-mean, there are two main differences from classical k mean. First, Euclidean

distance is used to calculate a distance between a dataset of points instead of data vectors. Second, the Euclidean distance is measured only if the radius value is higher than the threshold value. Finally, the EK-mean builds up in clusters for each period and then identifies special information from each period. In the end, the aggregator sends all the cluster's centroid values to the sink node [79].

An efficient data aggregation technique is done at the node level, which is also known as local aggregation. At this stage, most of the data collected by the sensor are strongly dependent on monitoring conditions, whereas there are greater chances of having redundant data. The measurement selected in a period contains dissimilar data, and at end of the period that measurement comes out in form of a vector [84]. Although the vector includes a lot of redundant data, users can use a similar technique to find redundancy data between two measures by specifying a threshold value. Only if their comparable function is equal to one (1), then two successive redundant measurements exist. The redundant data found in the vector is presented by the weight measurement function for information integration. The sensor takes a set of measurements without redundancy after each period and sends it to the CH. In the second stage, aggregation starts at the CH level by using similar functions. However, the CH also receives a set of measurements and their weight from its member nodes. Aggregation at the CH level aims to deduce redundancy between the member nodes using a specific threshold value. The Jaccard (similarity function) is used to identify similarities and their weights between two data sets. Whereas, for the comparison of the weight and data, the prefix frequency filtering (PFF) technique is proposed, and it works in two steps. The first step is candidate pairs generation in which the sensor searches candidate pairs at every dataset. The CH chooses a candidate pair only if the calculations are greater than β. The second step is the Candidates' verification. After finding candidate pairs, it considers a candidate pair between two data sets in case if the similarities of both data sets are greater than the Jaccard threshold. The analyzed variances between measurements are calculated using K-mean and adopting the ANOVA model and the Bartlett test. Both distance functions, Euclidean distance and Cosine distances, are used to determine data redundancy between data sets.

Extended cosine regression (ECR) data aggregation is proposed to reduce energy depletion during data collection [85]. It is a prediction model, and the key objectives are used to reduce data redundancy, maintain accuracy, and enhance network lifetime. ECR model is based on two vector models which stop inter-cluster transmission by making a data sequence that is implemented both at the member node and the CH in each cluster. A precondition is established for the prediction model. The node lifetime of each sensor is further divided into a mixed identical slot and in that slot, a sensor only has to store one identical value. Moreover, both the CH and member node have the same prediction algorithm and if both receive any knowledge, acknowledge each other. The base station broadcasts two threshold values to the CH and member node. One of the values is the application-specific error threshold, whereas the other is the user predefined threshold error. Furthermore, each member node forms two types of vectors. The first one is the actual data vector (ADV) which saves actual values. The second one is the predicted data vector (PDV) which saves the prediction values that are similar between CH and member node. The length of ADV and PDV remain the same at each cluster, however, the CH forms a corresponding vector for each of its member nodes. Every member node first saves the sensed value into ADV and PDV then sends it to the CH, which forms its private PDV. The ECR method is implemented on variations of cosine distance on linear regression which is divided into three phases. The first phase is the initialization phase. The based station (BS) broadcasts an acceptable prediction error to every CH and member node. In every cluster, its member node transfers its sensed data to CH in a cluster by using single-hop communication. Now the CH saves all the member node's data in an ordered form. Before predicting each sensor node is a new value that forms a vector from its old values using an ERM technique. At the end of the start-up phase, the CH has a huge amount of data that it uses in the ERM model. The second phase is the modelling phase that combines linear regression and cosine distance to increase the prediction accuracy of each member node. Generally, the linear regression does not pass any data unless the sequence line has perfect coordination. The differences between the two data sets are found by using signal differences between two vectors distance similarities. The third phase is the working phase where every member node calculates prediction value and prediction error. At this point, it compares the value with an error; if the value is lesser than or the same as the error, then it does not send it to the CH, while on the other hand, if the value is more than the error, then it sends actual value to the CH. Furthermore, the same procedure applies to CH when it receives the values.

WSNs also play a key role in big data collection. The major challenge of big data collection is the consumption of sensor energy which affects network lifetime. The two-level data reduction approach in WSN by Harb et al. [82] is proposed to reduce data communication and enhanced network lifetime. At the first stage, the data compression model is implemented at the node level where the coefficient of Pearson is used to identify a correlation between two data sets. If the Pearson coefficient is equal to 1, it means a positive correlation; otherwise, it considers no correlation when the indicator is equal to 0 or -1. The vector of the reading collected at each sensor node level is to reduce the Pearson Coefficient. The representation readings selection algorithm is applied on a vector to divide into sub vectors until these sub vectors show high correlation. The divide function divides the readings vector into two equal sub vectors. If the correlation is less than a certain level, the Pearson Coefficient is used to set a limit. This vector evaluates the final vector of readings, then consists

of mean reading values and weight of values, which is the number of readings represented by the mean value. At the end of the process, each sensor node provides a vector reading to the CH. Second, the data clustering model is used at the CH level to identify redundant data sets when received from member nodes. CH has used the K-mean algorithm to orderly allocate redundant data into data sets. The group data sets in K clustering by K-mean algorithm with Euclidean distance are used to group similar data sets in the same cluster.

To minimize communication and extend the network's lifetime, a two-level data reduction strategy is presented. Each node continues to compress data obtained at the first level using the Pearson Coefficient [55]. When the Pearson coefficient is equal to -1, there is no correlation between the two data sets; otherwise, there is 0. When the Pearson coefficient is equal to -1, there is a negative correlation between the two data sets. The high correlation is determined only in case if the predefined threshold is closely similar or equal to the data sets. Moreover, the data compression algorithm is implemented at each member node to compress the collected data vector and find a subset reading with the whole vector. Further, the vector readings are divided into sub vectors and to find a high correlation between vector readings by applying the Pearson coefficient. At the end of each process, every node contains mean values and weighted values representing a repeated number of readings. After receiving the data sets to CH from their cluster member nodes, the clustering data model is applied to identify the redundancy between grouping data sets at the same cluster based on K-mean and TopK nearest algorithm. The EK-mean algorithm is the combination of classic k-mean and Euclidean distance. EK-mean is used for checking the similarities between in cluster data sets to determine high and low correlation. The main objective of this process is to eliminate data redundancy from collected sensory data and nearest neighbouring nodes to reduce big data and enhance network lifetime.

Energy conservation is one of the critical issues in WSNs. To preserve energy for WSNs. Khriji et al. [81] propose a redundancy elimination data aggregation algorithm (REDA). The algorithm has two main characteristics, better data aggregation and enhanced network lifetime for reducing energy consumption. REDA is used to reduce data redundancy and communication based on the pattern code generation approach. The pattern code generation algorithm is applied on all sensor nodes for predefined sensed data. Each CH generates a ranges number of intervals called lookup table and then sends it to cluster members. Moreover, each member node compares its sensed data with a look-up table that was received from CH. According to the look-up table, each member node computes its pattern and sends the first iteration to the CH. The sensor node then computes a new pattern code and compares it to the old one. It does not transmit if both patterns are the same; else, it is sent to CH.

In Table VI, simulation parametric values of data redundancy at the CH level based on threshed value, measures readings, E_{elec}, ßamp, K, data and field, simulator and sensor are presented. Proposed methods and schemes are also mentioned with their resource parameters. Table VII shows the current proposed estimation the existing schemes and methods for data redundancy reduction in WSNs.

TABLE VI.
SIMULATION PARAMETRIC VALUES OF DATA REDUNDANCY REDUCTION AT THE CH LEVEL

References	Schemes/ Methods	Threshed Value	Measures Readings	E_{elec}	B_{amp}	K	Data and Field	Simulator And Sensor
[73]	DAT	0.03, 0.05, and 0.07	20, 50, and 100	50 nj/bit	100 pj/bit/m2	ρ/2	Intel Berkeley Lab	Omnet++ and Mica2dot
[74]	TLDA	0.03, 0.05, 0.07, and 0.1	200, 500, 1000 and 2000	50 nj/bit	100 pj/bit/m2	2	Intel Berkeley Lab and temperature	Omnet++ and Mica2dot
[75]	Grid-leader tier: support-confidence		50, 100 and 200				Intel Berkeley Research Lab,	Mica2dot
[78]	AAS	0.0 0 05 to 0.025	512					
[88]	Dual Prediction	0.2	500,and 1000	One bit is 600 Nj	Per clock cycle is 3.5 Nj			Matlab
[79]	Ek-Means	0.0 0 05 to 0.025	512	50 nj/bit	100 pj/bit/m2	5 and 50.	Argo Project	Java Based Simulator

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

TABLE VII.
PERFORMANCE COMPARATIVE PARAMETER'S FOR ENERGY SAVING AT THE CH LEVEL

References	Schemes/methods	Compared methods	Performance Comparative Parameter's						
			Percentage of data sent from sensor node to CH	Percentage of sets sent to the sink	Data accuracy	Energy consumption at CH	Number of redundant pairs	Data latency: execution time	Cost of communication energy
[73]	ETDTR	[71] and PFF	✓	✓		✓	✓		
[74]	TLDA	[71] and PFF			✓	✓			
[75]	Grid-leader tier: support-confidence	PFF technique		✓	✓				
[76]	TTDR				✓	✓			✓
[78]	AAS (Aggregation and Adaptive Sampling)	na¨ıve approach.		✓	✓				
[88]	dual prediction	CHCS and DPPCA							✓
[79]	EK-means	K-means				✓			
[80]	Data Compression				✓				✓
[82]	K-nearest neighboring (KNN)	S-LEC		✓	✓				
[83]	Two-Level Data Reduction	S-LEC and PFF							✓
[84]	KAB, Euclidean distance, cosine distance and PFF	KAB, Euclidean distance, cosine distance and PFF		✓	✓	✓	✓	✓	✓

C. DATA REDUNDANCY REDUCTION AT NODE TO CH AND CH TO SINK LEVEL

This section describes the recent studies of the data redundancy reduction algorithms and schemes at node to the CH and the sink level for the WSNs. The current techniques are studied and the absorbed important parameters measured for improvement of data redundancy in WSNs. Table VIII displays how some of existing methods addressed their problems with improved models, and proposed schemes for energy efficiency to reduce data at the sink level for WSNs. Table VII also debates the contributions and weaknesses of existing methods.

From node to the CH and then the sink level data reduction model is shown in Figure 5. There are two different processes of data aggregation in WSNs, containing the simple data redundancy reduction (DRR) and DRR with prediction. First, at simple DRR, a sensor periodically captures and aggregates the data by interacting with the environment and transferring data to the CH. Then the CH receives the member nodes' data and aggregates the data between data sets and sends it to the sink. After receiving data from CHs, the sink further processes and checks the data accuracy. Second, in many existing types of research, the same prediction models are implemented at both levels (sensor node or CH and the sink) for data redundancy reduction. The study [86] presents the adaptive dual prediction scheme (DPS) to reduce data transmission. To update the model's perimeter, history, data tables are avoided, and the old collection models that are already activated from past sequences are used to build DPS. A new prediction is started at sensor node and sink level which updates the perimeter models from time to time by using new data history tables and maintain the accuracy. First, an ordinary adoptive DPS is generated and computed on the sensor node and sink. As the data increases, the prediction models are supposed to activate and all the data is saved at the sink node. At the initial phase, the previous sample is eliminated, and a new sample is considered for the first data set for new model perimeter prediction.

However, the data size is based on a threshold. For this, WSNs is used in a ringing model and the sink node is present at the centre, whereas source to destination transmission occurs through its intermediate node. In the data routing scheme, data is sent from one node to another in ring, and this process goes on until the data reaches the sink node. A time interval is set to separate two consecutive transmissions on the sensor node, which is 30 minutes in current research. A lightweight algorithm is used to set at the node and the sink levels. The lightweight algorithm does not consume useless data due to which their storage and running time is increased. Also, the transmission model informs the number of exchanged data on the sensor at a period 1 / f is the number of transmissions found. The prediction model then forms a unicast transmission and predicts an accuracy at every sensor node. During transmission, if the data captured by any sensor

match, then that data is not transferred to the sink node where minimum accuracy is maintained so that the average is also maintained. The prediction data model depends on a sink or sensor node and gathers data from the cloud. The sink is responsible for generating the prediction model and spreading it towards sensor nodes. For spatial-temporal correlation between sensor data, a novel data reduction strategy called spatio-temporal correlation-based approach for sampling and transmission rate adaptation (STCSA) is developed. The data decreases the overall sampling and reduces transmission rate and maintains data quality [87].

Second, a backend reconstruction algorithm is proposed at the sink level to maintain data accuracy. However, at the node level, the algorithm still needs to perform a unique sampling rate and reduce data transmission at all sensors. At the end of every round, the CH runs the algorithm to find the spatial correlation between member nodes data that was sent to the CH. Next, the CH transfers the data to its sensors and provides a command to make a new sampling rate or the next round between cluster heads. To find a sampling rate, a high correlation is shown between many sensor nodes which are in a specific number.

TABLE VIII.
EXISTING STUDIES ARE BASED ON DATA REDUNDANCY SCHEMES AT SINK LEVEL

References	Schemes	Problems addressed	Proposed strategies	Improvements/ Enhancements	Limitations/ weaknesses
[80]	Dual Prediction (DP) and Data Compression (DC)	High data traffic generated in WSNs through spatial and temporal correlation	1. CM level: DP data store in buffer and use it predicts for the next value then evaluate the predictive value and new observation value. If accuracy is yes storing the value in buffer otherwise it sent to the CH 2. CH level: DP same illustrate in CH and to exploit spatial correlation by DC scheme 3. Sink level: same dc scheme constructs at the sink for data accuracy	Data transmission reductions	High affects bandwidth, energy, and congestion
[86]	Extended Dual Prediction Schemes (DPS)	Data transmission control	Ring model: The sensor nodes and the sink generate activity new data history table for update the parameters of a model	Communication reduction and accuracy	Synchronization problem between the sensor nodes and their neighbors is not considered
[88]	Dual prediction and hybrid compressed	Data redundancy temporal for intra cluster transmission and spatial for inter cluster transmission	1. Node level: It compares current value and predicts value when the error between them is equal or large 2. CH level: Decision for transmitting data if data is less than m it transmits directly else CH aggregate and compressed these data m dimension 3. Sink level reconstructs the original data by cs recovery algorithm	Reduce transmission energy cost and data recovery accuracy	Only reduce data size but still need to reduce data redundancy
[89]	Long short-term memory (LSTM)	Volume of raw data sent on the network	1. Node level: first some periods directly send to the sink node for training data at the sink level. After that sensor received predicted training data values and compared them with their original sensed data 2. After received sensor data sets and manage, sequential manner and training these data sets by 1st to predict remain periods of each sensor nodes	Decreasing data transmission and saving network lifetime	Does not consider the geographical distance between sensor nodes in WSNs
[90]	Reinforcement learning-based signal predictor (RLSP)	Exploit the signal prediction issue in a learning pattern at the sensor side	1. Node level: RLSP runs the same Q-table and parameters that are used at the sink level. If the predictive value and sensed value are not equal the sensor transmit the value to the sink 2. Sink level: received data use for next signal value and compare its own predicted data and update predicted model	Extremely low data transmission and energy consumption	It is needed to a lot of data and a lot of computational which is not feasible at the node level
[91]	Hybrid linear model (HLM)	High delay, high transmission cost, and complex model training	1. At node level: a forward stagewise algorithm for training of a hybrid linear model 2. Sink level: data reconstruction and prediction by using received hybrid model from the sensor nodes	Energy efficiency with controllable delay	

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

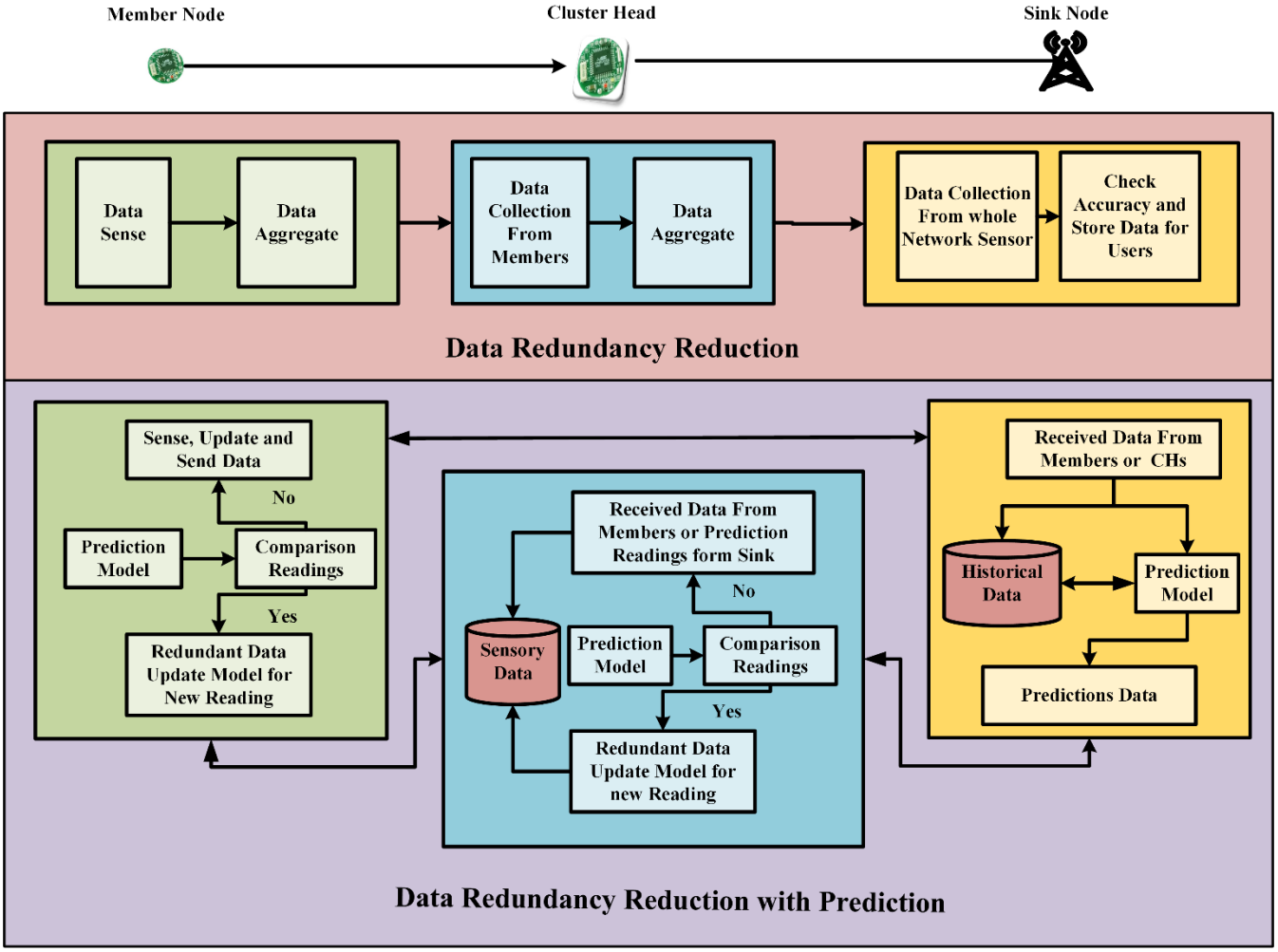

Figure 5. Energy efficiency data redundancy reduction process at Node, CH and Sink Level in WSN.

The stata is a form of vector of sensor nodes when it sends its data to CH which has the same size for each node. The Stata does not compute nan values for correlation; instead, it replaces the nan values with a number representing how often those nan values occur. After finding every sensor's correlation it prepares a table and divides the table into two columns: max (maximum) correlation for sensor j^{th} which forms sensor i^{th} to find the correlation degree. Some sensor data does not appear in the second column in the table because their data might not match. The aim of this process is to find correlation degrees between the sensor nodes which are found at the CH. Now the sink node is responsible for making prediction models and disseminating them to sensor nodes through unicast.

A distributed round-based prediction model for hierarchical large-scale sensor networks is adopted in the study [89]. The network lifespan is divided into a series of rounds in a distributed round-based prediction model, with each round including multiple periods. The periods are separated into defined slots, and the sensor takes some of the data from each period and transmits it to the sink node in each round, while the remaining sensor nodes are in sleep mode. When data is received, the sink nodes then apply a prediction model based on the long short-term memory (LSTM) time series on it and finds which sensor nodes are on sleep mode. The main purpose of finding data prediction is to reduce data transmission in sensors to save energy, decrease data, and improve lifetime. The data that is sent to the sink node is converted into the training data vector. If the size of training vectors is big, the maintained accuracy is high. When the sensor collects data, it sends the active packet name and collected data to the sink node; otherwise, if the packet is empty, it is considered as a sleep packet. In addition, when the sink node receives the training data vectors from the sensor, it tries to predict data for the next round. After receiving data from sensor nodes, it uses this data for normalization. Normalization is formed at different scales which are defined in the different ranges (0 to 1). The minmax scaler algorithm is used on received data. The sink processes Minmax scaler before the data prediction process and then neural network parameters are determined by three indicators: the selection of the values of the number of blocks, the number of time steps, the number of features, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3128353, IEEE Access

depend on simulation setup that is fixed. After the training process, the LSTM model starts to find variation between training data and expected future data. It is divided into two concepts: loss function and optimizer. Loss function calculates variations of data between training data and optimization uses an iterative method which samples randomly from data and finds an error in the loss function. To find the error, mean square error (MSR) is used. The first step is that the data is predicted. In the training period, when the sink first receives the data, it considers the last *d* stands for period values and forms an order to make a prediction value for the rest of the periods.

Moreover, Nazaktabar et al. [90] prepare a framework that learns the relation of signal behaviour to find the next value. The approach is used to reinforce learning-based signal predictor (RLSP). The purpose behind the preparation of the framework is to transfer the sensor data to the sink node when the signal fails at the sink side. The RLSP model is applied in DSP at both sensor node and the sink level through 0 initialization as well as in Q-learning algorithm, which is used to Q-table where the perimeter and configuration remain the same. In addition, on the sensor node side, if there is a difference between prediction and sensory data, it is sent to the sink node; otherwise, it discards. Once the data is received at the sink side, the sink uses it to set the next data values; in other words, the sink uses it for prediction. If the prediction values and the data are equal, then it is considered by signal action, otherwise it is considered by sensory data. On the other hand, if the sink node does not receive data from the sensor node, updating the Q-table utilizes old data and predicts a value. Now it is time to correct the sink node's projected data. The getinitiate action is used to declare the vector's final element as the initial prediction. The sensor gathers data from the environment and uses its prediction model to forecast it on the sensor side. If the data is rectified during prediction, the sensor does not transmit data to the sink and instead modifies the model to reflect the new information. Otherwise, the data is sent to the sink and its model updated by using the new readings. When the sink receives sensory data from (#k), it also uses its prediction model to predict data from the sensor that sends data. If the data gets predicted, then it is confirmed that the data is predicted and updated the number of k, and if the data does not get predicted, then that data is used for text prediction and updates of its prediction.

In the study [91], two models are combined to build a hybrid model based on historical data reconstruction and future data prediction for decreasing additional transmission, controlling delay, and improving the energy efficiency of sensors. This hybrid model is implemented in real-world WSNs and two algorithms have been proposed. The first one is on node-level which is the stage-wise algorithm. This algorithm avoids the computation load and creates flexibility in the hybrid model. In the second algorithm, data reconstruction and the prediction implemented at the sink node level evaluates hybrid models' performances with the help of rough experiments stimulated based on different data sets taken from real-world WSN applications. A hybrid linear model is presented which counts the continuous readings of the certain physical environment which are captured into time series by sensors. The sensor senses the data and uploads it without prediction. The linear model assumes that the environmental data have a short-term linear behaviour. It builds a training model in each section. Subsequently, instead of sending the original value, it sends the parameters of the training model to the sink node, which then construct pure data. Likewise, the same model is established at the sink node and sensor node level for future data.

However, if the prediction error increases, a pre-set threshold is set in the hybrid model, and the sensor sets as a retainer. There are two data points in each model and each data point has two data values and two reconstruct values. Using these data values, three parameters for a data model are used to stop additional transmission. To train Hybrid Linear models least square method is used which optimizes the error. After that, the new value is compared to the projected value, and if the difference is more than the absolute error, a new error is created. When the sensor node sends a hybrid model to the sink, the data is automatically built for reference, and the sensor predicts a value based on it.

WSNs have a challenging problem in terms of energy conservation and complicated decision-making for large amounts of data. Marwa et al. [75] present an energy-saving adaptive approach and decision-making approach. The technique is composed for grid-based architecture network consisted of three tiers. The first tier at the node level mostly works with redundant readings among the period for all slots. The collected readings vector is divided into three equal divisions by a divide-and-conquer algorithm. After that, the mean value is calculated for each division and a vector of these mean values dataset sent to the GL (grid-leader/CH). At the CH level layer, the GL gets a mean set of data from each sensor node at the conclusion of each session. After collecting important information from surrounding member nodes at each grid, the GL reduces the redundant data from data sets among the member nodes in grid. GL uses a mean support and frequent mean support algorithm with a predefined threshold to look for received mean values. It only sends the mean value which are equal to or greater to the defined threshold by the sink. At the sink level tier, the decision-making model is used for real-time decision-making, consisting of two main tables. The first table scores the decision table which is used for specific application such as a normal range of temperature, light, wind speed, and humidity. The second table is the early decision table which the users prepare to predetermined value matches with the collected value range for data aggregation. The technique worked efficiently and achieved the goals of saving energy and accuracy of the data.

The spatial and temporal correlation data flow among sensor nodes is one of the most serious problems in WSNs. In a two-tier data reduction system, dual prediction DP and data

compression DP [80] techniques are given to minimize data transmission in network traffic. The DP technique is used at the node level to identify redundant data at each node. The goal of the DP method is to minimize traffic between two points, such as a cluster member node and the cluster head. This scheme algorithm is constructed at both ends of link points. At the endpoint of CM, the last observation is held in buffer, but some collected observations are initially transmitted into the other endpoint CH. Later the observation value is predicted by using buffer held values which are then compared with the new observed value. If the new observed value is fairly similar to the prediction value, then the data transmission of the other endpoint does not happen; otherwise, it is sent. However, at the other ending point, CH saves the previous observation values in the buffer which have the same length as that other endpoint CM. CH receives a new observed value, which it saves in the buffer, but if no observation value is received, then it is considered as accuracy prediction value and both values are in same conditions, then it updates the DP model. DC scheme is implemented at each CH to exploit the spatial correlation data collected from CMs. CH makes data blocks on CMs transmitted data and sends these blocks to the sink. After receiving these blocks, the sink uses inverse operation to recover the block data. In Table IX, simulation parametric values such as period size (τ), measures readings, number of steps (*S*), number of blocks (*B*), number of features (F), size of round (ρ), training data size (α) and energy cost of message sending at the sink level are described.

Table X displays that the existing proposed methods and schemes with the benchmarks methods and schemes are used for comparison. Also, it indicates the various performance comparison parameters for the evaluation of the existing schemes and methods for data redundancy reduction in WSNs.

TABLE IX.
SIMULATION PARAMETRIC VALUES OF DATA REDUNDANCY REDUCTION AT SINK LEVEL

Reference	Schemes/ Methods	Period Size (τ)	Measures Readings	No of Steps (*S*)	No of Blocks (*B*)	No of Features (*F*)	Size of Round (ρ)	Training Data Size (α)	Energy Cost of Message Sending	Data and Field	Simulator and Sensor
[75]	Tier: Decision-Making Model		50, 100, and 200							Intel Berkeley Research Lab	Mica2dot
[89]	LSTM	100, 500, 1000 and 2000		50 and 200	5 and 50	1	10	1		Intel Berkeley Research Lab	Mica2dot
[90]	RLSP									Intel LAB and NDBC	Mica2dot
[91]	HLM	13 bytes, and 11 bytes,							0.0144mJ /Byte	Intel Berkeley Lab and Life Under Your Feet (LUYF) project	Mica2dot

TABLE X.
PERFORMANCE COMPARATIVE PARAMETER'S SAVE ENERGY AT SINK LEVEL

References	Schemes/ Methods	Compared Methods	PERFORMANCE COMPARATIVE PARAMETER'S						
			Fixed the Period Size	Decision Results at the Sink	Data Transmission Ratio from Sensor to Sink	Data Accuracy Prediction in Function Of T, S And B	Memory Complexity	Energy Consumption and Time Complexity	Delay Control
[75]	Sink Tier: Decision Making Model	GL1 with GL2	✓	✓					
[89]	Long Short-Term Memory LSTM	SFDC	✓		✓	✓			
[90]	RLSP	AR and CM					✓	✓	

[91]	HLM	(DBP) and (OSSLMS)	✓		✓	✓

In the data collection, thresholds are fixed depending upon small and large data readings values. There are two types of scenarios found in existing studies: the small range of readings that give small threshold values while the large range of readings give the large threshold values. However, work needs to be done on large range of readings, which should give the small threshold values in terms of date redundancy reduction in WSNs. At the node level, energy analysis, percentage of data sent after aggregation, data transmission ratio to the CH and data loss are most useable comparative performance parameters found, while redundant data calculation, adoption sample, energy saving at each node and data sets sent ratio are less considerable comparative performance parameters and need to be more focused in further research. At the CH level, energy consumption at the CH, data accuracy and number of data sent to the sink are most applied performance comparative parameters found while data received from member nodes, data latency and communication cost are less substantial performance comparative parameters and need to be more focused in further research. At the Sink level, period size fixed, calculation of data transmission ratio, and energy consumed are most effective performance comparative parameters found while data delay and memory usage are less extensive performance comparative parameters and need to be more focused in further research.

Omnet++ and java-based simulator are the most suitable and used simulators considered for the simulation. Mica2dot is the most appropriate sensor device and readings from the temperature are used as a dataset for the simulation of date redundancy reduction in WSN.

V. -ANALYSIS OF PERFORMANCE METRICS USED IN EXISTING STUDIES

In this section, performance metrics used for assessing in the previous research studies are presented in Tables II, V, VIII and III, VI, and IX. There are several performance parameter matrices are used to measure data redundancy reduction at sensor nodes and CHs level. Numerous performance matrices are presented, which are used in current studies to determine energy saving in WSNs.

The main goal is to maximize profits and revenue from WSN's. For this purpose, different techniques/algorithms or schemes are used which increase the user satisfaction, avoid raw data transmission, decrease energy consumption, and enhance network lifetime. On the other hand, data redundancy reduction minimizes the overheads and increases the overall performance. The performance parameters are used for data redundancy reduction for saving energy in existing studies are described below:

A. NETWORK LIFETIME

Network lifetime consider in various definitions with the time in which the network performs the desired task include time till network becomes disjoint link, first node fail, certain percentage of nodes fail given predefined threshold, largest links disconnected, some percentage of data rate loss, and all nodes fail. In clustering architecture, the network lifetime is defined as the time till the nodes in the network entirely deplete their energy in the network [92][93][77]. The network lifetime is calculated when all rounds of network fail due to the discontinuation of one or more sensors with the help of Equation (1).

$$Nlif_n^n = \underset{sn \in SN}{mim}\ Nlif_{sn} \tag{1}$$

where $Nlif_{sn}$ is the lifetime of *sn* is a sensor node, *mim* is the minimum energy sensor nodes and *SN* is the sets of nodes without including the sink, $Nlif_n^n$ represents the failure of the first node's life in a network.

B. DATA AGGREGATION RATE

Data aggregation rate is an amount to remove duplicate values and allows sensor nodes to decrease the quantity of data gathered. The amount of decrease is determined by the threshold value selected as well as the total number of recorded readings. The threshold value can be increased to identify additional data redundancy [94][65]. The primary motivation for using a data aggregation at the node level is to conserve energy d-bit data is shown in equation (2)

$$DA_E = dE_{TC} \tag{2}$$

Where DA_E is the energy consumption while data aggregating, d is d-bit data and E_{TC} is energy consumption at transceiver.

C. DATA AGGREGATION AT THE CH

Overall energy usage in WSNs can be reduced by decreasing transmission costs. For reducing the inter clustering communication by data redundancy elimination at CH level in a cluster; otherwise, redundant data is influencing the whole network transmission [76][94]. Correspondingly, the total energy consumption at cluster head in a cluster are receiving, aggregating, and transmitting data to the sink node with the preference of Equation 3.

$$CH_{TE} = CH_{dre} + DA_E + CH_{dtr} \tag{3}$$

Where CH_{TC} is total energy consumption at CH level, CH_{dre} is energy consumption of data received at CH, DA_E energy consumption on data aggregation, CH_{dtr} is energy consumption for data transfer.

D. DATA ACCURACY

Error-free data is a term used to describe data correctness. The data accuracy is calculated by dividing the proportion of data lost by the amount of data supplied by sensor nodes. The data loss measurements increase as the sensing range and reconstruction error threshold value between the data readings. Hence, as the quantity of obtained readings increases over time [77][85]. The data accuracy is calculated by the quantity of data successfully transmitted and the total amount of data sent on sensor nodes using Equation (4) [94].

$$D_{AC} = \sum \frac{(EsM - AcM)}{\sum AcM} \times 100 \tag{4}$$

Where EsM is estimation mean, and AcM is actual mean of data.

E. ENERGY ANALYSIS

Energy analysis is an important concern for WSNs due to its resource-constrained network [73][68]. The total energy consumption is calculated in-network energy consume include the function perform data aggregation, data received and response, transmissions, and computation. The energy consumption for data transmission is determine equation (5).

$$DT_E(d,D) = E_{elec-Dt} + E_{amp} = \begin{cases} D.E_{elec} + D.\varepsilon_{fs} \times d^2, \left(d \le d_0\right) \\ D.E_{elec} + D.\varepsilon_{amp} \times d^4, \left(d \ge d_0\right) \end{cases} \tag{5}$$

Where DT_E is energy consumption of data transmit, d is distance consist of sending and response, sending D bit data to the node, $E_{elec-Dt}$ data transmission link loss and is amplifier link loss E_{amp}

F. DATA AGGREGATION AT EVERY NODE

Each sensor node is found data redundancy between the data sensed measurements at each time during the aggregation process. Aggregation is thus dependent on the threshold, the number of collective measurements each period, and changes in the monitoring object. However, if data aggregation is not performed at each sensor node, a higher volume of data is sent, resulting in increased network energy usage. Aggregation is advantageous on a network because it reduces transmission costs while increasing network lifespan. As a result, the proportion of data aggregation focuses on all other factors such as node lifespan, transmission cost, energy usage, and network lifetime [66][79]. As stated in equation (6), the data aggregation at each node is determine by following formula .

$$N_{Da} = \frac{\sum DT_A}{DT_t} \times 100 \tag{6}$$

Where N_{Da} is data aggregation at node, DT_A total aggregated data before transmitted and DT_t is the total data collection before transmitted

G. DATA DUPLICATE SETS

At the conclusion of each period, the CH gets all the data sets from each member node in a cluster. The most common issue in aggregator/CH is that huge data sets are gathered, and redundant data sets must be removed before being delivered to the sink node. To increase data accuracy and network lifespan, redundant data sets must be eliminated [74]. The duplicate data examines variations between two dataset's comparisons by using Equation (7).

$$DR_S\left(D_{s1}, D_{s2}\right) = \frac{\sum_{k=1}^{n}\left(D_{s1k} \times D_{s2k}\right)}{\sqrt{\sum_{k=1}^{n} D_{s1k}^2} \times \sqrt{\sum_{k=1}^{n} D_{s2k}^2}} \tag{7}$$

Where DR_S is redundant data sets, $\left(D_{s1}, D_{s2}\right)$ are two data sets, n is the length of data sets and if and only if D_{s1} and D_{s2} are considered redundant it is $\left(D_{s1}, D_{s2}\right) \le \delta$ less than threshold.

H. ENERGY CONSUMPTION AT THE NODE LEVEL

When each sensor node collects data during most of the data is redundant. it means that the monitoring condition speeds up and slows down as a result, more readings from each sensor node are redundant [74][79]. As a result, energy consumption is increased because the sensor sends all collected readings to their CH. However, the network lifespan is quickly depleted energy of each sensor node in network. The energy consumption of data receiving node receiving d bit data is $NE(R)$ determine by Equation (8) and (9).

$$NE(R) = d \times E_{elec} \tag{8}$$

and energy consumption of data transmit on sensor node d bit data is $NE(DT)$.

$$NE(DT) = DT \times d_E \tag{9}$$

Where Da_E is required energy for bit data transfer.

I. DATA REDUNDANCY BETWEEN TWO CONSECUTIVE READINGS

The data redundancy between two consecutive readings are calculated by similar functions [94][61]. Therefore, two readings are redundant if and only if the similar function is 1 or 0 as shown in Equation (10).

$$Similar\left(R_1, R_2\right) = \begin{cases} 1, & if \left\| R_1 - R_2 \right\| \le \delta, \\ 0, & otherwise \end{cases} \quad (10)$$

where (R_1, R_2) are the two consecutive readings and R_1 is compared to R_2 if both are same added 1 on first reading; otherwise, the second value is considered a new reading with δ user-defined threshold.

J. ENERGY CONSUMPTION AT THE CLUSTER HEAD

Energy consumption in the CH is illustrated as the energy needed for data management, data received from member nodes and transfer data toward the sink [84][95]. The energy consumption at CH is determined by the total energy consumed at CH is measured by Equations (11).

$$CH_E = DE_{elec} MN/C + DE_{Da} MN/C + DE_{amp}\, d_{SK}^4 \quad (11)$$

where DE_{elec} energy consumed of data transmit, energy consumption in aggregation is DE_{Da}, MN/C is a number of average nodes per cluster, and d_{SK}^4 is the distance from the sink node to the cluster head.

K. DATA SENT RATIO:

Data quantitative analysis by ratio functions determine in each node, cluster head and the sink node. It uses various data analysis parameters such as data sent, data aggregate, data received, data reduction, and data transfer ratio.

$$DS_{ratio} = \frac{D_R}{D_{TR}} \times 100 \quad (12)$$

where D_R data reduction in total data received D_{TR}. The data ratio is calculated by Equation (12).

L. NUMBER OF REDUNDANT PAIRS:

In two or more than two nodes, inter data redundancies are correlated. Various studies used Pearson correlation [96] to determine the percentage of data redundancy in the same type of data collected by various neighbour nodes [15]. Suppose two vectors' data V1 and V_2, with the help of Equation (13), are determined by the data redundancy correlation coefficient.

$$\rho\, v_1\, v_2 = \frac{\vec{v_1} \cdot \vec{v_2}}{\left\| v_1 \right\| \cdot \left\| v_2 \right\|} \times \delta \quad (13)$$

Hence $\rho\, v_1\, v_2$ the Pearson correlation between two data vectors where δ is a sign parameter and it takes two values such as positive 1 or negative -1 values.

VI. SUGGESTED FUTURE WORK

The main common issue associated with energy efficiency in WSNs are sensor nodes correlation, threshold values define, energy efficiency at hostile environmental conditions issue, high transmission, data reduction and aggregation, prediction system, data accuracy, various level data transmission, check sum error or bit modulation, data collision and data redundancy etc. Figure 6 is a bubble graph that narrates the future directions in energy efficiency as pointed out in previous research articles by other authors.

A. SENSOR NODES CORRELATION:

Networks are randomly dense nodes deployed as the distance nearest two or more sensor nodes, known as neighbouring sensor correlation. When sensor nodes cross the predefined distance limit and are geographically close to one another due to this they generate duplicate data as a result high network traffic still need to enhance for dense deployment nodes, as recommended by [65] [66] [82] [73] [75] [79] [89][83].

B. PREDEFINED THRESHOLD:

Predefined threshold values control the data limits. There is no standard threshold for data size and shape and comparing research with various applications is challenging. Control and termination data links are more difficult to achieve [65].

C. ENERGY EFFICIENCY AT HOSTILE ENVIRONMENTAL CONDITIONS ISSUE

Glaciers, floods, environmental monitoring, and health care are examples of real-world applications. In WSNs, energy-efficient data redundant reduction is a prerequisite for real-world applications such as underwater and healthcare applications. Therefore, achieving an enhanced lifetime of the network is still considered in a real-world application as a most challenging issue in the research [66][97][98].[68].

D. HIGH TRANSMISSION:

The WSNs transmission means the data or any message that travels on a link from one node to another node or among the nodes which reach its end user. It is estimated in real-life test-bed applications. In WSNs, due to data redundancy, high bandwidth is required, thus high usage energy and congestion occur due to high data transmission rate and duration. These are key challenge in WSNs indicators for future work by [99]–[102] [80] [87] .

E. DATA REDUCTION AND AGGREGATION:

Data reduction and aggregation means combining multiple sensor nodes data set to reduce data size and maintain accuracy as well as energy saving by an aggregator or cluster head. However, data processing need more capacity by different data redundancy processing including data merging with as key-value stores [103], blockchains [104], and big data [105] in future work [106].

F. DATA PREDICTION SYSTEM:

In WSNs, the sensors predict future data based on the previous sensed data, reducing transmission and increasing energy.

Previous studies focused on data prediction algorithms on each sensor level and CH level where there is a need for high memory, high data analysis processing. As a result, there is still need to focus on data prediction system for WSNs to improve energy and transmission cost for further recommendations [107][86] [75][108].

G. DATA ACCURACY:

Data accuracy is depending upon data aggregation and data reduction ratio. In WSNs, data accuracy is the major component of information quality. Data redundancy occurs in WSNs due to inefficient transmission and data process complexity. In order to measure data redundancy reduction, sufficient amount of data must be available for maintaining the data accuracy, it is necessary to fix the amount of data redundancy reduction in scientific way. So that research needs to work or fined the accurate amount of data redundancy reduction. The data accuracy improvements still recommendation for further research [109]–[111] [75].

H. VARIOUS LEVEL DATA TRANSMISSION:

Various level data transmission means three-phase. In the first phase, data transmits from member node to CH; in the second phase, the data transmits from the CH to the sink are known as forward data transmission; and in the last phase, data transmits from sink to CH or each node in the network are known as backward data transmission. As future work, data reduction techniques to reduce data transmission at three-tiers require more focus such as fog tier, gateways tier, and each sensor node tier as recommended by [112] [76].

I. CHECK SUM ERROR OR BIT MODULATION:

Bit representation means sensor collected data is converted from digital to binary. The data transmission in WSNs is also considered as a part of bit representation, which is proposed by [67] to reduce data transmission for save energy. However, bit presentation only detects odd bit numbers, and does not detect when the data values are in even form. Thus, some errors occur in data, indicating the need for improvement as a future recommendation.

J. DATA COLLISION:

When two sensor nodes are sent similar packets at the same time, the collision occurs. In WSNs data collision detect more difficult for determining the sensor nodes location of a fault can be challenging in [78] [84] [69] for further enhancement.

K. DATA REDUNDANCY:

In WSNs, data redundancy occurs for different reasons. primarily, two consecutive readings may be the same and between two periods readings are same, neighboring sensor nodes data sets are the same, nearest neighbor nodes correlation sensed data are same because environmental conditions have speed up or slowed down. In [105], [113], [114] [78] [66] [75] [76], data redundancy is suggested, especially in terms of energy consumption, data quality, and network lifetime. Thus, there is a need to design efficient data reduction algorithms and the enhancement the lifetime WSNs.

Refer to Figure 6, there is still a need to improve data transmission cost. Data redundancy reduction is highly recommended for future research in various applications while less research is required for data accuracy, data prediction and criticality of application in 2021. In 2020, further research is focused on the neighbouring node regions correlation, real application, and data redundancy reduction. Also, more research is presently focused on the data different sampling rates, data prediction system, transmission reduction neighboring nodes correlation and data redundancy reduction and fusion in 2019. Similarly, in 2018, there is a further suggestion to focus research activities towards the area of neighboring nodes correlation, data different sampling rates, and accuracy. Previously based on 2017-2015, there is an enormous interest in areas such as three-tier data transmission, neighboring nodes correlation, different sampling rates, data redundancy, and data redundancy and fusion. These exposed issues and future works present a conclusive role in relating the technological strategies for further improving energy efficiency in WSN.

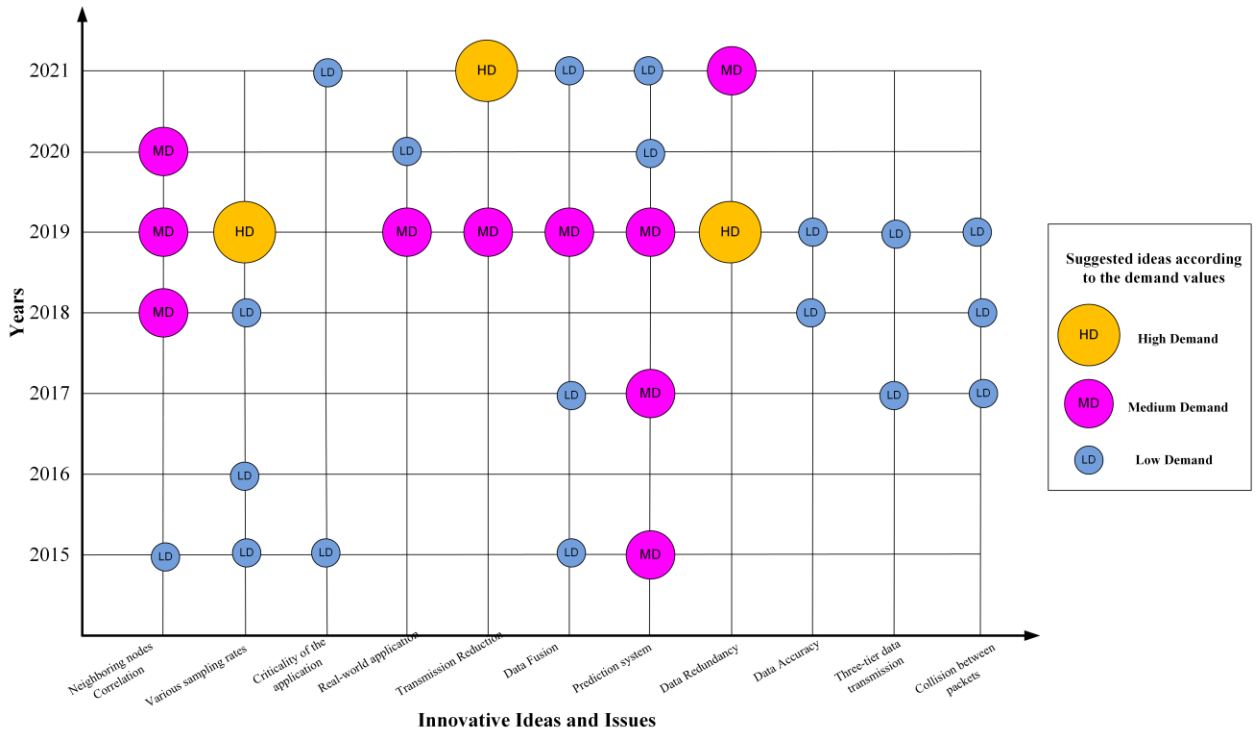

Figure 6. Innovative ideas of energy efficiency in WSNs.

VII. CONCLUSION

WSNs is an inspiring field of research, with the aim of proportional modernization and high performance in terms of technological express. The data redundancy reduction schemes or methods should be as simple as possible, requiring less computation, processing, and transmission so that they consume less energy and increase data accuracy to work as part of IoT, cloud computing, and smart networks. There are some challenges and limitations. Scalability, dynamic environment, mobility, node localization, and user satisfaction in wireless sensor networks are also aided by IoT applications, enhancing the network lifetime and saving energy. In this review article, a classification scheme and an expressive literature review of data redundancy reduction are briefly described for energy efficiency in WSNs. However, existing data redundancy reduction schemes for energy-efficiency in WSNs research is still contradictory due to the application`s requirements and technological concerns, such as application-oriented data redundancy, QoS for designing and implementing approaches, spatial correlation as a lot of resources such as transmission, bandwidth, data accuracy, and energy are wasted in case of data redundancy (like big data and IoT resources) regarding monitoring environmental condition of WSNs are involved . The classification and descriptive review provide a detailed description and open loopholes with their benefits and limitations for researchers and experts of the current assertion of data redundancy reduction and stimulate the further research interest of energy efficiency in WSNs.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

[1] P. Prasad, “Recent trend in wireless sensor network and its applications: A survey,” *Sens. Rev.*, vol. 35, no. 2, pp. 229–236, 2015, doi: 10.1108/SR-08-2014-683.

[2] S. Khan, A. S. K. Pathan, and N. A. Alrajeh, *Wireless sensor networks: Current status and future trends*. 2016.

[3] W. Dargie and C. Poellabauer, *Fundamentals of Wireless Sensor Networks: Theory and Practice*. 2011.

[4] T. Azzabi, H. Farhat, and N. Sahli, “A survey on wireless sensor networks security issues and military specificities,” *Proc. Int. Conf. Adv. Syst. Electr. Technol. IC_ASET 2017*, pp. 66–72, 2017, doi: 10.1109/ASET.2017.7983668.

[5] I. Ahmad, K. Shah, S. Ullah, A. Wali khan University Mardan, and K. Pakhtoonkhwa, “Military Applications using Wireless Sensor Networks: A survey,” *Int. J. Eng. Sci. Comput.*, 2016.

[6] S. A. Kumar and P. Ilango, "The Impact of Wireless Sensor Network in the Field of Precision Agriculture: A Review," *Wireless Personal Communications*. 2018, doi: 10.1007/s11277-017-4890-z.

[7] K. Goel and A. K. Bindal, "Wireless sensor network in precision agriculture: A survey report," in *PDGC 2018 - 2018 5th International Conference on Parallel, Distributed and Grid Computing*, 2018, pp. 176–181, doi: 10.1109/PDGC.2018.8745854.

[8] T. Alhmiedat, "A Survey on Environmental Monitoring Systems using Wireless Sensor Networks," *J. Networks*, vol. 10, no. 11, 2016, doi: 10.4304/jnw.10.11.606-615.

[9] K. S. Adu-Manu, C. Tapparello, W. Heinzelman, F. A. Katsriku, and J. D. Abdulai, "Water quality monitoring using wireless sensor networks: Current trends and future research directions," *ACM Transactions on Sensor Networks*. 2017, doi: 10.1145/3005719.

[10] A. Saad, A. E. H. Benyamina, and A. Gamatie, "Water Management in Agriculture: A Survey on Current Challenges and Technological Solutions," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.2974977.

[11] R. A. Khan and A. S. K. Pathan, "The state-of-the-art wireless body area sensor networks: A survey," *Int. J. Distrib. Sens. Networks*, vol. 14, no. 4, 2018, doi: 10.1177/1550147718768994.

[12] S. Chelbi, C. Duvallet, M. Abdouli, and R. Bouaziz, "Event-driven wireless sensor networks based on consensus," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 0, pp. 1–6, 2016, doi: 10.1109/AICCSA.2016.7945684.

[13] A. Maheshwari and N. Chand, "A Survey on Wireless Sensor Networks," in *Lecture Notes in Networks and Systems*, vol. 46, no. 11, 2019, pp. 153–164.

[14] P. Kuila and P. K. Jana, *Clustering and routing algorithms for wireless sensor networks: Energy efficiency approaches*. 2017.

[15] K. T. M. Tran and S. H. Oh, "UWSNs: A round-based clustering scheme for data redundancy resolve," *Int. J. Distrib. Sens. Networks*, vol. 2014, 2014, doi: 10.1155/2014/383912.

[16] M. S. Bensaleh, R. Saida, Y. H. Kacem, and M. Abid, "Wireless Sensor Network Design Methodologies: A Survey," *Journal of Sensors*, vol. 2020. 2020, doi: 10.1155/2020/9592836.

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

[18] E. Nagarajan, "A Survey on Wireless Sensor Network : Energy and Lifetime Perspective," *Taga J.*, vol. 14, no. April, pp. 1099–3113, 2018, doi: 10.13140/RG.2.2.11629.69606.

[19] K. Guleria and A. Kumar, "Comprehensive review for energy efficient hierarchical routing protocols on wireless sensor networks," *Wirel. Networks*, vol. 25, no. 3, pp. 1159–1183, 2019, doi: 10.1007/s11276-018-1696-1.

[20] N. Sabor, S. Sasaki, M. Abo-Zahhad, and S. M. Ahmed, "A comprehensive survey on hierarchical-based routing protocols for mobile wireless sensor networks: Review, taxonomy, and future directions," *Wireless Communications and Mobile Computing*, vol. 2017. Hindawi, pp. 1–23, Jan. 10, 2017, doi: 10.1155/2017/2818542.

[21] S. Yadav and R. S. Yadav, "A review on energy efficient protocols in wireless sensor networks," *Wirel. Networks*, vol. 22, no. 1, pp. 335–350, 2016, doi: 10.1007/s11276-015-1025-x.

[22] L. Chan, K. Gomez Chavez, H. Rudolph, and A. Hourani, "Hierarchical routing protocols for wireless sensor network: a compressive survey," *Wirel. Networks*, vol. 26, no. 5, pp. 3291–3314, 2020, doi: 10.1007/s11276-020-02260-z.

[23] 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," *Wirel. Networks*, vol. 26, no. 2, pp. 1243–1263, 2020, doi: 10.1007/s11276-019-02142-z.

[24] P. Jesus, C. Baquero, and P. S. Almeida, "A Survey of Distributed Data Aggregation Algorithms," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 1, pp. 381–404, 2015, doi: 10.1109/COMST.2014.2354398.

[25] N. Goyal, M. Dave, and A. K. Verma, "Data aggregation in underwater wireless sensor network: Recent approaches and issues," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 3, pp. 275–286, 2019, doi: 10.1016/j.jksuci.2017.04.007.

[26] V. I Puranikmath, S. S Harakannanavar, S. Kumar, and D. Torse, "Comprehensive Study of Data Aggregation Models, Challenges and Security Issues in Wireless Sensor Networks," *Int. J. Comput. Netw. Inf. Secur.*, vol. 11, no. 3, pp. 30–39, 2019, doi: 10.5815/ijcnis.2019.03.05.

[27] S. Sirsikar and S. Anavatti, "Issues of data aggregation methods in Wireless Sensor Network: A survey," in *Procedia Computer Science*, 2015, vol. 49, no. 1, pp. 194–201, doi: 10.1016/j.procs.2015.04.244.

[28] C. Mallick and S. Satpathy, "Challenges and Design Goals of Wireless Sensor Networks: A

Sate-of-the-art Review," *Int. J. Comput. Appl.*, vol. 179, no. 28, pp. 42–47, 2018, doi: 10.5120/ijca2018916667.

[29] S. Kalantary and S. Taghipour, "A survey on architectures, protocols, applications, and management in wireless sensor networks," *J. Adv. Comput. Sci. Technol.*, vol. 3, no. 1, pp. 1–11, 2014, doi: 10.14419/jacst.v3i1.1583.

[30] T. Rajasekaran and S. Anandamurugan, *Challenges and Applications of Wireless Sensor Networks in Smart Farming—A Survey*, vol. 750. Springer Singapore, 2019.

[31] T. Bala, V. Bhatia, S. Kumawat, and V. Jaglan, "A survey: Issues and challenges in wireless sensir network," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 53–55, 2018, doi: 10.14419/ijet.v7i2.4.10041.

[32] H. N. Dai, R. C. W. Wong, H. Wang, Z. Zheng, and A. V. Vasilakos, "Big data analytics for large-scale wireless networks: Challenges and opportunities," *ACM Comput. Surv.*, vol. 52, no. 5, 2019, doi: 10.1145/3337065.

[33] B.-S. Kim *et al.*, "Wireless Sensor Networks for Big Data Systems," *Sensors*, vol. 19, no. 7, p. 1565, Apr. 2019, doi: 10.3390/s19071565.

[34] S. Boubiche, D. E. Boubiche, A. Bilami, and H. Toral-Cruz, "Big Data Challenges and Data Aggregation Strategies in Wireless Sensor Networks," *IEEE Access*, vol. 6, pp. 20558–20571, 2018, doi: 10.1109/ACCESS.2018.2821445.

[35] S. Kumar and H. Kim, "Energy efficient scheduling in wireless sensor networks for periodic data gathering," *IEEE Access*, vol. 7, pp. 11410–11426, 2019, doi: 10.1109/ACCESS.2019.2891944.

[36] P. D. C. M. Ajit R. Pagar, "A Survey on Energy Efficient Sleep Scheduling in Wireless Sensor Network," , *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 5, no. 1, pp. 4–8, 2015, [Online]. Available: www.ijarcsse.com.

[37] M. Bagaa, Y. Challal, A. Ksentini, A. Derhab, and N. Badache, "Data Aggregation Scheduling Algorithms in Wireless Sensor Networks: Solutions and Challenges," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 3, pp. 1339–1368, 2014, doi: 10.1109/SURV.2014.031914.00029.

[38] A. Ali, Y. Ming, S. Chakraborty, and S. Iram, "A comprehensive survey on real-time applications of WSN," *Future Internet*, vol. 9, no. 4. 2017, doi: 10.3390/fi9040077.

[39] M. S. Bensaleh, S. M. Qasim, A. M. Obeid, and A. Garcia-Ortiz, "A review on wireless sensor network for water pipeline monitoring applications," in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, 2013, pp. 128–131, doi: 10.1109/CTS.2013.6567217.

[40] M. F. Othman and K. Shazali, "Wireless sensor network applications: A study in environment monitoring system," in *Procedia Engineering*, 2012, vol. 41, pp. 1204–1210, doi: 10.1016/j.proeng.2012.07.302.

[41] H. M. Jawad, R. Nordin, S. K. Gharghan, A. M. Jawad, and M. Ismail, "Energy-efficient wireless sensor networks for precision agriculture: A review," *Sensors (Switzerland)*, vol. 17, no. 8. 2017, doi: 10.3390/s17081781.

[42] S. Ghosh, S. Mondal, and U. Biswas, "Enhanced PEGASIS using ant colony optimization for data gathering in WSN," *2016 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2016*, no. Icices, pp. 1–6, 2016, doi: 10.1109/ICICES.2016.7518930.

[43] J. Premalatha and P. M. Joe Prathap, "A Survey on Underwater Wireless Sensor Networks: Progresses, Applications, and Challenges," *MATEC Web Conf.*, vol. 57, pp. 1147–1154, 2016, doi: 10.1051/matecconf/20165702007.

[44] R. E. Mohamed, A. I. Saleh, M. Abdelrazzak, and A. S. Samra, "Survey on Wireless Sensor Network Applications and Energy Efficient Routing Protocols," *Wirel. Pers. Commun.*, vol. 101, no. 2, pp. 1019–1055, 2018, doi: 10.1007/s11277-018-5747-9.

[45] D. Kandris, C. Nakas, D. Vomvas, and G. Koulouras, "Applications of Wireless Sensor Networks: An Up-to-Date Survey," *Appl. Syst. Innov.*, vol. 3, no. 1, p. 14, 2020, doi: 10.3390/asi3010014.

[46] L. M. L. Oliveira and J. J. P. C. Rodrigues, "Wireless sensor networks: A survey on environmental monitoring," *Journal of Communications*, vol. 6, no. 2. pp. 143–151, 2011, doi: 10.4304/jcm.6.2.143-151.

[47] M. P. Mashere, S. S. Barve, and P. D. Ganjewar, "Data Reduction in Wireless Sensor Network : A Survey," *Int. J. Comput. Sci. Technol.*, vol. 8491, pp. 86–88, 2015.

[48] D. I. Curiac, C. Volosencu, D. Pescaru, L. Jurca, and A. Doboli, "Redundancy and its applications in wireless sensor networks: A survey," *WSEAS Trans. Comput.*, vol. 8, no. 4, pp. 705–714, 2009, Accessed: Oct. 18, 2019. [Online]. Available: http://www.aut.upt.ro/~curiachttp://www.cs.utt.ro/~dan/http://www.ee.sunysb.edu/~adoboli/.

[49] N. Verma and D. Singh, "Data Redundancy Implications in Wireless Sensor Networks," 2018, doi: 10.1016/j.procs.2018.05.036.

[50] M. Singh Manshahia, "Wireless Sensor Networks: A Survey," *Int. J. Sci. Eng. Res.*, vol. 7, no. 4, 2016, Accessed: Feb. 23, 2020. [Online]. Available: http://www.ijser.org.

[51] L. M. L. Oliveira and J. J. P. C. Rodrigues,

"Wireless sensor networks: A survey on environmental monitoring," *J. Commun.*, vol. 6, no. 2, pp. 143–151, 2011, doi: 10.4304/jcm.6.2.143-151.

[52] X. Fan, W. Wei, M. Wozniak, and Y. Li, "Low Energy Consumption and Data Redundancy Approach of Wireless Sensor Networks with Bigdata," *J. Inf. Technol. Control*, vol. 47, no. 3, p. 47, 2018, doi: 10.5755/j01.itc.47.3.20565.

[53] K. Maraiya, K. Kant, and N. Gupta, "Application based Study on Wireless Sensor Network," *Int. J. Comput. Appl.*, vol. 21, no. 8, pp. 9–15, 2011, doi: 10.5120/2534-3459.

[54] B. Fateh and M. Govindarasu, "Energy minimization by exploiting data redundancy in real-time wireless sensor networks," *Ad Hoc Networks*, vol. 11, no. 6, pp. 1715–1731, Aug. 2013, doi: 10.1016/j.adhoc.2013.03.009.

[55] H. Harb and C. A. Jaoude, "Combining compression and clustering techniques to handle big data collected in sensor networks," *2018 IEEE Middle East North Africa Commun. Conf. MENACOMM 2018*, pp. 1–6, 2018, doi: 10.1109/MENACOMM.2018.8371009.

[56] H. Harb and A. Makhoul, "Energy-efficient scheduling strategies for minimizing big data collection in cluster-based sensor networks," *Peer-to-Peer Netw. Appl.*, vol. 12, no. 3, pp. 620–634, 2019, doi: 10.1007/s12083-018-0639-z.

[57] M. Medlej, "Big data management for periodic wireless sensor Maguy Medlej To cite this version : HAL Id : tel-01228515," 2015.

[58] A. K. Idrees, A. Kadhum, and M. Al-Qurabat, "Distributed Adaptive Data Collection Protocol for Improving Lifetime in Periodic Sensor Networks." [Online]. Available: http://www.iaeng.org/IJCS/issues_v44/issue_3/IJCS_44_3_10.pdf.

[59] Arvind and D. Singh, "IMPORTANCE OF ENERGY IN WIRELESS SENSOR NETWORKS : A SURVEY," *Int. J. Eng. Sci.*, vol. 17, no. January 2016, pp. 500–505, 2016.

[60] H. Harb, A. Makhoul, D. Laiymani, A. Jaber, and R. Tawil, "K-means based clustering approach for data aggregation in periodic sensor networks," *Int. Conf. Wirel. Mob. Comput. Netw. Commun.*, pp. 434–441, 2014, doi: 10.1109/WiMOB.2014.6962207.

[61] G. B. Tayeh, A. Makhoul, C. Perera, and J. Demerjian, "A Spatialoral Correlation Approach for Data Reduction in Cluster-Based Sensor Networks," *IEEE Access*, vol. 7, pp. 50669–50680, 2019, doi: 10.1109/ACCESS.2019.2910886.

[62] M. Wu, L. Tan, and N. Xiong, "Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications," *Inf. Sci. (Ny).*, vol. 329, pp. 800–818, Feb. 2016, doi: 10.1016/j.ins.2015.10.004.

[63] O. Zahwe, O. Majed, H. Harb, M. Hamze, and A. Nasser, "A Fast Clustering Algorithm for Analyzing Big Data Generated in Ubiquitous Sensor Networks," *2018 Int. Arab Conf. Inf. Technol.*, pp. 1–6, 2018.

[64] R. K. Yadav and J. Singh, "Energy Efficient Clustering Based Data Gathering Using Hybrid DB-EMGM In Distributed Sensor Networks," pp. 601–606, 2017.

[65] H. Harb, A. Makhoul, R. Couturier, and M. Medlej, "ATP: An aggregation and transmission protocol for conserving energy in periodic sensor networks," *Proc. - 2015 IEEE 24th Int. Conf. Enabling Technol. Infrastructures Collab. Enterp. WETICE 2015*, pp. 134–139, 2015, doi: 10.1109/WETICE.2015.9.

[66] A. K. M. Al-Qurabat and A. Kadhum Idrees, "Data gathering and aggregation with selective transmission technique to optimize the lifetime of Internet of Things networks," *Int. J. Commun. Syst.*, no. November 2019, pp. 1–31, 2020, doi: 10.1002/dac.4408.

[67] M. Hammad, M. Bsoul, M. Hammad, and M. Al-Hawawreh, "An efficient approach for representing and sending data in wireless sensor networks," *J. Commun.*, vol. 14, no. 2, pp. 104–109, 2019, doi: 10.12720/jcm.14.2.104-109.

[68] S. R. U. Jan, M. A. Jan, R. Khan, H. Ullah, M. Alam, and M. Usman, "An Energy-Efficient and Congestion Control Data-Driven Approach for Cluster-Based Sensor Network," *Mob. Networks Appl.*, vol. 24, no. 4, pp. 1295–1305, 2019, doi: 10.1007/s11036-018-1169-x.

[69] A. K. M. Al-Qurabat and A. K. Idrees, "Energy-efficient adaptive distributed data collection method for periodic sensor networks," *Int. J. Internet Technol. Secur. Trans.*, vol. 8, no. 3, pp. 297–335, 2018, doi: 10.1504/IJITST.2018.093660.

[70] A. Makhoul and H. Harb, "Data Reduction in Sensor Networks: Performance Evaluation in a Real Environment," *IEEE Embed. Syst. Lett.*, vol. 9, no. 4, pp. 101–104, 2017, doi: 10.1109/LES.2017.2749333.

[71] H. Harb, A. Makhoul, D. Laiymani, O. Bazzi, and A. Jaber, "An Analysis of Variance-Based Methods for Data Aggregation in Periodic Sensor Networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9430, Springer Verlag, 2015, pp. 165–183.

[72] H. Harb, A. Makhoul, R. Tawil, and A. Jaber, "A suffix-based enhanced technique for data

aggregation in periodic sensor networks," *IWCMC 2014 - 10th Int. Wirel. Commun. Mob. Comput. Conf.*, pp. 494–499, 2014, doi: 10.1109/IWCMC.2014.6906406.

[73] A. K. Idrees, R. Alhussaini, and M. A. Salman, "Energy-efficient two-layer data transmission reduction protocol in periodic sensor networks of IoTs," *Pers. Ubiquitous Comput.*, 2020, doi: 10.1007/s00779-020-01384-5.

[74] A. K. M. Al-Qurabat and A. K. Idrees, "Two level data aggregation protocol for prolonging lifetime of periodic sensor networks," *Wirel. Networks*, vol. 25, no. 6, pp. 3623–3641, 2019, doi: 10.1007/s11276-019-01957-0.

[75] M. Ibrahim, H. Harb, A. Nasser, A. Mansour, and C. Osswald, "Adaptive Strategy and Decision Making Model for Sensing-Based Network Applications," *Proc. - 2019 19th Int. Symp. Commun. Inf. Technol. Isc. 2019*, pp. 96–101, 2019, doi: 10.1109/ISCIT.2019.8905211.

[76] A. K. M. Al-Qurabat, C. A. Jaoude, and A. K. Idrees, "Two tier data reduction technique for reducing data transmission in IoT sensors," *2019 15th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2019*, pp. 168–173, 2019, doi: 10.1109/IWCMC.2019.8766590.

[77] H. Harb, A. Makhoul, A. Jaber, and S. Tawbi, "Energy efficient data collection in periodic sensor networks using spatio-temporal node correlation," *Int. J. Sens. Networks*, vol. 29, no. 1, pp. 1–15, 2019, doi: 10.1504/IJSNET.2019.097547.

[78] A. Karaki, A. Nasser, C. A. Jaoude, and H. Harb, "An adaptive sampling technique for massive data collection in distributed sensor networks," in *2019 15th International Wireless Communications and Mobile Computing Conference, IWCMC 2019*, 2019, pp. 1255–1260, doi: 10.1109/IWCMC.2019.8766469.

[79] M. Rida, A. Makhoul, H. Harb, D. Laiymani, and M. Barhamgi, "EK-means: A new clustering approach for data sets classification in sensor networks," *Ad Hoc Networks*, vol. 84, pp. 158–169, Mar. 2019, doi: 10.1016/j.adhoc.2018.09.012.

[80] A. Jarwan, A. Sabbah, and M. Ibnkahla, "Data Transmission Reduction Schemes in WSNs for Efficient IoT Systems," *IEEE J. Sel. Areas Commun.*, vol. PP, no. c, pp. 1–1, 2019, doi: 10.1109/jsac.2019.2904357.

[81] S. Khriji, G. Vinas Raventos, I. Kammoun, and O. Kanoun, "Redundancy elimination for data aggregation in wireless sensor networks," *2018 15th Int. Multi-Conference Syst. Signals Devices, SSD 2018*, pp. 28–33, 2018, doi: 10.1109/SSD.2018.8570459.

[82] H. Harb, A. Makhoul, and C. A. Jaoude, "En-Route Data Filtering Technique for Maximizing Wireless Sensor Network Lifetime," *2018 14th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2018*, pp. 298–303, 2018, doi: 10.1109/IWCMC.2018.8450348.

[83] H. Harb, A. Makhoul, and C. A. Jaoude, "A real-time massive data processing technique for densely distributed sensor networks," *IEEE Access*, vol. 6, pp. 56551–56561, 2018, doi: 10.1109/ACCESS.2018.2872687.

[84] H. Harb, A. Makhoul, S. Tawbi, and R. Couturier, "Comparison of Different Data Aggregation Techniques in Distributed Sensor Networks," *IEEE Access*, vol. 5, pp. 4250–4263, 2017, doi: 10.1109/ACCESS.2017.2681207.

[85] K. Jain and A. Kumar, "An energy-efficient prediction model for data aggregation in sensor network," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 11, pp. 5205–5216, 2020, doi: 10.1007/s12652-020-01833-2.

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

[87] G. B. O. U. Tayeh, A. Makhoul, and J. Demerjian, "A Spatial-Temporal Correlation Approach for Data Reduction in Cluster-Based Sensor Networks," *IEEE Access*, vol. 7, pp. 50669–50680, 2019, doi: 10.1109/ACCESS.2019.2910886.

[88] Y. Zhou, L. Yang, L. Yang, and M. Ni, "Novel energy-efficient data gathering scheme exploiting spatial-temporal correlation for wireless sensor networks," *Wirel. Commun. Mob. Comput.*, vol. 2019, 2019, doi: 10.1155/2019/4182563.

[89] G. Saad, H. Harb, C. A. Jaoude, and A. Jaber, "A Distributed Round-Based Prediction Model for Hierarchical Large-Scale Sensor Networks," in *International Conference on Wireless and Mobile Computing, Networking and Communications*, 2019, vol. 2019-Octob, pp. 1–6, doi: 10.1109/WiMOB.2019.8923312.

[90] H. Nazaktabar, K. Badie, and M. Nili, "RLSP : a signal prediction algorithm for energy conservation in wireless sensor networks," *Wirel. Networks*, vol. 23, no. 3, pp. 919–933, 2017, doi: 10.1007/s11276-016-1200-8.

[91] X. Xu and G. Zhang, "A Hybrid Model for Data Prediction in Real-World Wireless Sensor Networks," *IEEE Commun. Lett.*, vol. 7798, no. c, pp. 1–1, 2017, doi: 10.1109/lcomm.2017.2706258.

[92] P. Chaturvedi, … A. D.-N. F. for I. and 5G, and undefined 2021, "Priority Encoding-Based Cluster Head Selection Approach in Wireless Sensor

Networks," *igi-global.com*, Accessed: Feb. 01, 2021. [Online]. Available: https://www.igi-global.com/chapter/priority-encoding-based-cluster-head-selection-approach-in-wireless-sensor-networks/265034.

[93] S. Randhawa and S. Jain, "Data Aggregation in Wireless Sensor Networks: Previous Research, Current Status and Future Directions," *Wirel. Pers. Commun.*, vol. 97, pp. 3355–3425, 2017, doi: 10.1007/s11277-017-4674-5.

[94] J. M. Bahi, A. Makhoul, and M. Medlej, "Data aggregation for periodic sensor networks using sets similarity functions," *IWCMC 2011 - 7th Int. Wirel. Commun. Mob. Comput. Conf.*, pp. 559–564, 2011, doi: 10.1109/IWCMC.2011.5982594.

[95] A. Al Qurabat and A. K. IDREES, "Energy-efficient Adaptive Distributed Data Collection method for Periodic Sensor Networks," *Int. J. Internet Technol. Secur. Trans.*, vol. 1, no. 1, p. 1, 2017, doi: 10.1504/ijitst.2017.10007731.

[96] C. Salim and N. Mitton, "K-predictions based data reduction approach in WSN for smart agriculture," *Computing*, vol. 103, no. 3, pp. 509–532, 2021, doi: 10.1007/s00607-020-00864-z.

[97] S. Kumar and V. K. Chaurasiya, "A Strategy for Elimination of Data Redundancy in Internet of Things (IoT) Based Wireless Sensor Network (WSN)," *IEEE Syst. J.*, vol. 13, no. 2, pp. 1650–1657, 2019, doi: 10.1109/JSYST.2018.2873591.

[98] H. Ramezanifar, M. Ghazvini, and M. Shojaei, "A new data aggregation approach for WSNs based on open pits mining," *Wirel. Networks*, vol. 27, no. 1, pp. 41–53, 2021, doi: 10.1007/s11276-020-02442-9.

[99] S. Pasupathi, S. Vimal, Y. Harold-Robinson, M. Khari, E. Verdú, and R. G. Crespo, "Energy efficiency maximization algorithm for underwater Mobile sensor networks," *Earth Sci. Informatics*, vol. 14, no. 1, pp. 215–225, 2021, doi: 10.1007/s12145-020-00478-1.

[100] X. Xiao, H. Huang, and W. Wang, "Underwater wireless sensor networks: An energy-efficient clustering routing protocol based on data fusion and genetic algorithms," *Appl. Sci.*, vol. 11, no. 1, pp. 1–24, 2021, doi: 10.3390/app11010312.

[101] W. K. Yun and S. J. Yoo, "Q-Learning-based data-aggregation-aware energy-efficient routing protocol for wireless sensor networks," *IEEE Access*, vol. 9, pp. 10737–10750, 2021, doi: 10.1109/ACCESS.2021.3051360.

[102] W. Feng, J. Wang, Y. Chen, X. Wang, N. Ge, and J. Lu, "UAV-aided MIMO communications for 5g internet of things," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1731–1740, 2019, doi: 10.1109/JIOT.2018.2874531.

[103] T. Li *et al.*, "GRAPH/Z: A key-value store based scalable graph processing system," 2015, doi: 10.1109/CLUSTER.2015.90.

[104] A. Al-Mamun, T. Li, M. Sadoghi, and D. Zhao, "In-memory Blockchain: Toward Efficient and Trustworthy Data Provenance for HPC Systems," 2019, doi: 10.1109/BigData.2018.8621897.

[105] P. Mehta *et al.*, "Comparative evaluation of big-data systems on scientific image analytics workloads," 2017, doi: 10.14778/3137628.3137634.

[106] J. Wang, O. T. Tawose, L. Jiang, and D. Zhao, "A new data fusion algorithm for wireless sensor networks inspired by hesitant fuzzy entropy," *Sensors (Switzerland)*, vol. 19, no. 4, 2019, doi: 10.3390/s19040784.

[107] A. K. Idrees and A. K. M. Al-Qurabat, "Energy-Efficient Data Transmission and Aggregation Protocol in Periodic Sensor Networks Based Fog Computing," *J. Netw. Syst. Manag.*, vol. 29, no. 1, pp. 1–24, 2021, doi: 10.1007/s10922-020-09567-4.

[108] V. Chandran and P. G. Scholar, "Elimination of Data Redundancy and Latency Improving in Wireless Sensor Networks." Accessed: May 15, 2019. [Online]. Available: www.ijert.org.

[109] Z. Yemeni, H. Wang, W. M. Ismael, Y. Wang, and Z. Chen, "Reliable spatial and temporal data redundancy reduction approach for WSN," *Comput. Networks*, vol. 185, no. June 2020, p. 107701, 2021, doi: 10.1016/j.comnet.2020.107701.

[110] M. M. Warrier and A. Kumar, "Energy efficient routing in Wireless Sensor Networks: A survey," in *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, 2016, pp. 1987–1992, doi: 10.1109/WiSPNET.2016.7566490.

[111] S. Mehrjoo and · Farshad Khunjush, "Accurate compressive data gathering in wireless sensor networks using weighted spatio-temporal compressive sensing," *Telecommun. Syst.*, vol. 68, pp. 79–88, 2018, doi: 10.1007/s11235-017-0376-2.

[112] A. K. Idrees, H. Harb, A. Jaber, O. Zahwe, and M. A. Taam, "Adaptive distributed energy-saving data gathering technique for wireless sensor networks," *Int. Conf. Wirel. Mob. Comput. Netw. Commun.*, vol. 2017-Octob, pp. 55–62, 2017, doi: 10.1109/WiMOB.2017.8115805.

[113] A. Kadhum, M. Al-Qurabat, A. K. Idrees, A. K. M. Al-Qurabat, and A. Kadhum Idrees, "Distributed Data Aggregation protocol for improving lifetime of Wireless Sensor Networks Advanced Networks and Distributed Algorithms View project Routing in the Internet of Things (IoTs) Networks View project QALAAI ZANIST JOURNAL Distributed

Data Aggreg," *A Sci. Q. Ref. J. Issued by Leban. French Univ.*, no. 2, 2017, doi: 10.25212/lfu.qzj.2.2.22.

[114] S. Kandukuri, J. Lebreton, R. Lorion, N. Murad, and J. Daniel Lan-Sun-Luk, "Energy-efficient data aggregation techniques for exploiting spatio-temporal correlations in wireless sensor networks," *Wirel. Telecommun. Symp.*, vol. 2016-May, 2016, doi: 10.1109/WTS.2016.7482055.

GUL SAHAR is working as lecturer in Department of Computer Sciences, Karakoram International University Gilgit, Pakistan. She received the M.S. degree in mobile ad hoc network from International Islamic University at Islamabad, Islamabad, in 2012. She is currently pursuing her Ph.D. at Universiti Teknologi Malaysia. Her current research interests include wireless sensor networks, big data, and the Internet of Things.

Email: gulsahar@kiu.edu.pk

KAMALRULNIZAM BIN ABU BAKAR received the B.Sc. degree in computer science from the Universiti Teknologi Malaysia, Malaysia, in 1996, the M.Sc. degree in computer communications and networks from Leeds Metropolitan University, U.K., in 1998, and the Ph.D. degree in computer science from Aston University, U.K., in 2004. He is currently a Professor in computer science with the Universiti Teknologi Malaysia, and a member of the Pervasive Computing Research Group. His research interests include Mobile and Wireless Computing, Adhoc and Sensor Networks, Information Security, and Grid Computing. He is a member of the ACM, the Internet Society, and the International Association of Engineering. He involves in many research projects and is a referee for several scientific journals and conferences. Email: knizam@utm.my

FATIMA TUL ZUHRA received the BS(CS) degree in computer science from the Quaid-e-Awam University of Engineering, Science & Technology, Pakistan, in 2009 and the MCS degree in computer science from the University of Malaya, Malaysia, in 2016. She received the Ph.D. degree in computer science from the Universiti Teknologi Malaysia, Malaysia in 2020. She is currently a Researcher of Universiti Teknologi Malaysia under the Post-Doctoral Fellowship project: Efficient Route Stability-Aware Routing Scheme for Wireless Body Sensor Networks. She is a member of the Pervasive Computing Research Group. Her research interests include Wireless Sensor Network, Wireless Body Sensor Network, routing algorithms, Artificial Intelligence, and Internet of Things.
E-mail: fatima-tul-zuhra@utm.my

SABIT RAHIM is working as Assistant Professor in Department of Computer Sciences, Karakoram International University Gilgit, Pakistan; He received his MS from Hamdard University Pakistan and PhD from University of Science and Technology Beijing China. His research interest is machine learning; Cloud computing; IoTs and ICT in education in rural areas of developing countries. He is also member of provincial ICT policy for education of Gilgit-Baltistan, Pakistan.
Email: sabit.rahim@kiu.edu.pk

TEHMINA BIBI is working as Assistant Professor at the Institute of Geology, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan. She received her M.Phil in Geology from University of Peshawar Pakistan and PhD from Universiti Teknologi Malayasia, Malaysia in Geoinformatics. Her research interest is Natural Hazards; Risk Assessment; disaster risk reduction and future modelling through IoTs in mountainous area of Asian countries. She is also member of Pakistan Association of Petroleum Geoscientists (PAPG).
Email: tehmina.bibi@ajku.edu.pk

SYED HAMID HUSSAIN MADNI is a currently working as a Senior Lecturer (Assistant Professor) in School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Malaysia. He received his PhD Degree in 2020 from UTM, Malaysia. His area of research is "Optimal Resource Scheduling for Infrastructure as a Service in Cloud Computing based on Cuckoo Search". He has received MS (CS) degree in 2009 from Federal Urdu University Arts, Science and Technology, Islamabad, Pakistan. His areas of interest are cloud computing, analysis of algorithm, computer networks, e-learning, e-health, and Internet of Things. He has published about 15 research papers in in High Impact Journals. He is also conducting training on research publications with different International Universities in various countries. Email: madni4all@yahoo.com