

ORIGINAL RESEARCH PAPER

A reliable and energy efficient dual prediction data reduction approach for WSNs based on Kalman filter

Haibin Wang¹ | Zaid Yemeni¹ | Waleed M. Ismael¹ | Ammar Hawbani²  | Saeed H. Alsamhi³

¹ College of IoT Engineering, Hohai University, Changzhou, Jiangsu, China

² School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

³ Software Research Institute, Athlone Institute of Technology, Athlone, Ireland

Correspondence

Haibin Wang, Yemeni Zaid, College of IoT Engineering, Hohai University, Changzhou, Jiangsu, China.
Email: wanghaibin@hhuc.edu.cn; yemenizaid@hhuc.edu.cn

Funding Information

Fundamental Research Funds for the Central Universities, Grant/Award Number: B200202216; Innovation Foundation of Radiation Application, China Institute of Atomic Energy, Grant/Award Number: KFZC2020010401

Abstract

Wireless sensor networks (WSNs) are critically resource-constrained due to wireless sensor nodes' tiny memory, low processing units, power limitations, and narrow communication bandwidth. The data reduction technique is one of the most widely used techniques to reduce transmitted data over the wireless sensor networks and to minimize the sensor nodes' energy consumption, particularly, the entire network in general. This paper proposes a reliable dual prediction data reduction approach for WSNs. This approach performs data reduction through two phases: the data reduction phase (DRP) and data prediction phase (DPP). The DRP is mainly to decrease the number of transmissions between the sensor node and the sink node, thereby minimizing energy consumption. It also detects faulty data and discards them at the sensor node. The discarded faulty data at the sensor nodes are replaced by estimated values at the sink node to maintain data reliability. DPP runs at the sink node or base station, which works in synchronization with the sensor nodes. This phase is responsible for predicting the non-transmitted data based on the Kalman filter. The simulation results demonstrate that the proposed approach is efficient and effective in data reduction, data reliability, and energy consumption.

1 | INTRODUCTION

Owing to the dramatic change in humankind's lifestyle, WSNs have become a commonly used technology for data collection in various applications. Monitoring is one of the most WSNs applications used to perform over different environments and processes. WSNs consist of thousands of nodes, which are used to sense a specific phenomenon or event. These nodes are usually randomly or systematically deployed in targeted areas to collect physical parameters information, such as temperature, pressure, humidity, vibration, noise level, and vital signs. Monitoring applications are essential in environmental, health care, governmental, industrial, and military applications [1–4]. The data collected by monitoring applications are required to flow continuously over time. However, wireless sensor nodes are tiny devices with extreme computational and energy limitations. Wireless sensor nodes' batteries are limited for a finite period of time, depending on many factors (e.g., the number of data

transmissions). Consequently, prolonging network lifetime is one of the most important research topics in WSNs [5,6].

Due to the limitations mentioned above, the data collected by wireless sensor nodes are enormous, some of which may be unnecessary and faulty. The continuously collected data are highly correlated due to the observed phenomena' physical nature [7,8]. In other words, most of the collected observations are not exclusive, and the deviation between them and the previously collected readings has no significant entropy. Therefore, discarding the transmission of unnecessary collected data substantially decreases energy consumption in WSNs. This results in prolonged network lifetime, as most energy is dissipated in the data transmission process [1,9,10]. Therefore, one of the proposed solutions to this issue was to reduce the number of data transmissions over wireless networks [11,12]. However, it minimizes the network overhead at the same time [13]. To lessen the volume of transmitted data, various state-of-the-art methods have been proposed, including data aggregation (DA) [14–16],

data compression (DC) [17,18], adaptive sampling [19–21], or data prediction (DP) [22–24].

Nevertheless, DP is a more preferred and efficient approach, as it can realize a significant data suppression ratio (SR) compared with other techniques [1]. The DP concept is to exploit the temporal collected data by finding a correlation between a previous set of collected data and building a prediction model that can predict the future measurements to be similarly correlated. This can be achieved by discarding the transmissions of the newly collected data that can be estimated at other ends, such as cluster heads, sinks, base stations, edge nodes, or clouds. The exploiting operation can be done at the sensor nodes by running a prediction algorithm where collected data is compared with predicted values. If the predicted data are accurate enough (based on the system's needs), the current measurement transmission will be cancelled. Thus, the sensor nodes only transmit non-predicted values to the sink node. In this case, the prediction algorithm is no longer accurate unless the sink can reproduce the non-transmitted readings.

Nevertheless, some earlier proposed prediction-based data reduction techniques may cause an increase in nodes' computational operations. Some algorithms are not applicable in real-world monitoring WSNs due to the sensor node computational and memory limitations. This paper aims to explore a new prediction-based data reduction approach to improve the data reduction performance in transmission reduction, data reliability, and energy consumption. The main contributions to this paper are twofold:

- Developing a data reduction algorithm, which aims to decrease the number of data transmissions from the sensor nodes to the sink nodes. The collected data will be discarded transmission if it's redundant, predictable, or faulty, as discussed in Section 3.
- Developing a data prediction algorithm that runs at the sink node based on the Kalman Filter to predict non-transmitted data from end nodes while maintaining data reliability.

The remainder of this paper is organized as follows: Section 2 provides an overview of the recently proposed data reduction algorithms. The system description and assumptions are discussed briefly in Section 3. Next, the proposed data reduction approach is discussed in Section 4. In Section 5, the simulation results are presented in comparison with various counterparts. Finally, Section 6, based on our findings, gives a brief conclusion for the article, along with further study directions.

2 | RELATED WORK

Data reduction is a technique used to decrease data transmission in WSNs, thus prolonging networks' lifetimes and reducing network bottlenecks. There have been diverse state-of-the-art approaches dedicated to increasing WSNs longevity. The authors in [25] proposed a new data transmission reduction algorithm based on the dual prediction model to decrease the amount of data transmitted to the sink node. The proposed

approach aims to prolong the network lifetime by combining data reduction with the adaptive sampling rate technique. The authors in [1] proposed a two-tiers data prediction framework based on dual prediction (DP) and data compression (DC) schemes. The DP tier aims to minimize the data transmissions from sensor nodes to their cluster heads. Simultaneously, the DC tier is proposed to decrease the communication from cluster heads to the base station. Additionally, the authors have used neural networks (NNs) and long short-term memory networks (LSTMs) for data predictions at the DP tier. Although this approach has reduced the data transmissions by up to 54%, the energy consumption and data accuracy were compromised.

For In-networking based data reduction, the authors in [26] proposed error-aware data clustering (EDC). The proposed EDC contains three different adaptive modules that allow users to choose a module that suits their required data quality. Based on temporal data redundancy, the main contribution of this technique was to remove the temporal data correlation using different techniques such as histogram-based data clustering (HDC), Recursive Outlier Detection, Smoothing (RODS), and Verification of RODS (V-RODS). The proposed EDC reduces the amount of data redundancy and detects faulty data that may happen in WSNs. Another in-networking approach is proposed in [27]. The proposed approach contains two layers, namely, data filtering and data fusion layers. The data filtering layer aims to minimize the number of transmissions. At the same time, the data fusion layer is based on the minimum square error criterion to fuse the data. It is worth mentioning that this approach is based on the Kalman Filter for data prediction at the edge node, which is a part of the techniques we used in this paper. Another distributed data predictive model (DDPM) is proposed in [28], aiming to increase energy efficiency by reducing data transmissions over WSNs. The DDPM model is based on a finite impulse response filter combined with a recursive least squares adaptive filter.

The least-mean-square (LMS) algorithm is also used for data reduction in WSNs. In [29], the authors proposed a data reduction method based on two decoupled least-mean-square (LMS) windowed filters that combined convexly with different sizes. The proposed method estimates future readings in both sink and sensor nodes. The data transmission occurs if the current reading value has a significant deviation from a predefined threshold. Another prediction-based data reduction approach is proposed in [30]. This approach exploits hierarchical least-mean-square (HLMS) for data reduction in WSNs. HLMS is used to predict values at sensors and sink nodes individually. In addition, the working mechanism HLMS was proposed, and the mean-square error was analysed. While the interactive HLMS protocol and prediction algorithm are designed for both sink and sensor nodes. In [31], the authors proposed a prediction-based data reduction technique. The proposed method aims to reduce the data transmission by establishing a relationship between sensor readings. Therefore, the sensor node is exempt from sending a massive volume of data during a predefined duration. Instead of transmitted collected data, the CH is supposed to predict the non-transmitted data and thus minimize the energy consumption of WSN. In [32], the authors aim to

prolong network lifetime by decreasing the amount of wireless sensor data transmission. Therefore, a data transmission reduction scheme has been proposed based on machine learning. The presented machine learning-based data transmission reduction scheme is specified for IoT network applications.

In [33], a machine learning-based data reduction approach (MLDR) for WSNs is proposed. The proposed MLDR is mainly proposed for enhancing the environmental data reduction of agriculture applications. Same as many DP schemes, the prediction is done simultaneously in both sensor and sink nodes. Another deep learning-based distributed data mining (DDM) model is proposed in [34]. DDM aims to increase the network energy efficiency by minimizing data transmissions. Firstly, the wireless network is divided into several layers based on a combination of recurrent neural network (RNN) and long short-term memory (LSTM) RNN-LSTM. These layers are then placed on the sensor nodes. In addition, the fusion centre overhead is reduced to decreasing the number of transmissions. In [35], the authors proposed a fuzzy-based clustering and machine learning-based data reduction in WSNs to increase the life span of the networks. The proposed method exploited machine learning to extract similar sensed data and discard it. Also, the Fuzzy Inference System is used to form suitable clusters based on the update cycle calculated. Another machine learning-based data reduction is proposed in [36]. The proposed method aims to decrease energy harvesting by using the K -mean algorithm as the network's basis clustering process. In addition, the number of data transmissions is decreased by using the optimal fixed packet size.

Since approaches in [6] (hereafter referred to as DP_LSTM), [27] (hereafter referred to as DDR-IoT), and [28b,28] (hereafter referred to as LMS) carry out the data reduction process based on dual prediction in both sensor and sink nodes, we chose them as benchmarks for evaluating the performance of the proposed approach.

Unlike the approaches mentioned above, the proposed approach's strength lies in its ability to detect faulty data while increasing data reliability and reducing the number of transmissions. Many state-of-the-art approaches have been proposed for data reduction in WSNs, but the data reliability is compromised. In this paper, the proposed approach considers the fact that WSNs collected data are apt to faultiness and unreliability. Thus, data reduction and faulty data detection are proposed in this paper while enhancing data reliability.

3 | SYSTEM DESCRIPTION AND ASSUMPTION

3.1 | System description

Consider a WSN of N sensor nodes, $S = \{s_1, s_2, \dots, s_N\}$, and M sink nodes located in the same sensing area. These nodes are deployed systematically or randomly such that each sensor node s_i senses and transmits its readings to one of the sink nodes at a predefined time t . In this paper, a data reduction approach is proposed such that the similarly (correlated and redundant)

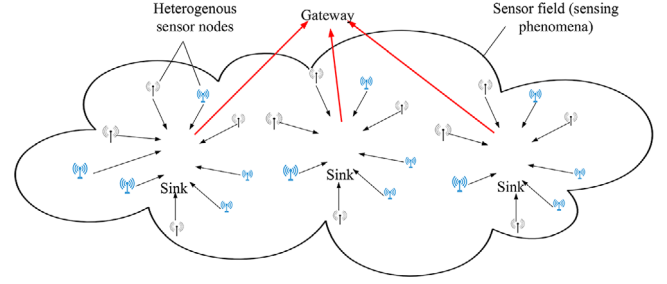


FIGURE 1 The network model example

collected data by the sensor node are discarded. It is assumed that only the deviated readings from the predefined threshold will be forwarded to the sink node. After that, the sink node is supposed to predict the non-transmitted readings and forward all the received and predicted readings to the gateways. The presented network model scenario of this proposed approach is illustrated in Figure 1. As can be seen, the sensor nodes' main task is to monitor the targeted environment, collect observations, and send them to the sink node. On the other hand, the sink nodes aim to receive the transmitted observations, predict the non-transmitted observations, and send them to the gateway. Finally, the gateway received the collected data and sent them to the end-user.

It is worth mentioning that the data transmission protocol has not been considered in this research. Instead, we assumed the communication between the sensor and the sink node is device-to-device communication. So, for instance, the data transmitted from the sensor node reaches the sink node in a timely manner. If no data is received at a given time t , it is assumed to be discarded at the reduction phase (at the sensor node level). Hence, the data prediction phase (DPP) will predict the non-transmitted data. Furthermore, according to [36,37], data transmission is the main factor in energy consumption of the WSNs. Therefore, the proposed approach to energy consumption is calculated based on the number of readings transmitted from sensor nodes to the sink node.

Figure 1 shows that the links between the end nodes (sensor node) and the sink node are depicted in black. The data reduction phase proposed an algorithm to reduce the number of transmissions. In contrast, the links between the sink node and the gateway are depicted in red. The data prediction phase is used to reproduce the non-transmitted readings.

3.2 | System assumption

Due to the fact that each WSN has its own special factors that affect the network performance, such as the targeted sensed phenomena, targeted monitoring events, type of the node (static or mobile), the proposed approach has its assumptions as follows:

- All the sensors in the network are stationary.
- The sensors are randomly or systematically distributed in the sensing environment.

- The data sampling rate is fixed for all nodes.
- Same as the sensing process, the transmission process is done timely. In other words, the data transmissions between sensors and sinks reach timely.
- Unlike the sensor nodes, the sink nodes have no power, memory, or processing limitations.

4 | PROPOSED APPROACH

Transmitting the entire sensed data is not a good idea in many cases. Data reduction is the key to solving some WSNs issues, including minimizing energy consumption and eliminating redundant data. In this regard, a data reduction mechanism has been proposed for many state-of-the-art types of research. In this paper, a two-phases dual prediction data reduction approach is proposed to control and minimize the data transmissions on the sensor side and reproduce the non-transmitted data on the sink side.

4.1 | Data reduction phase (DRP)

The DRP aims to decrease the number of data transmissions between the sensor nodes (end nodes) and the sink nodes. In the presented scenario, the links between every sensor node s_i and its corresponding sink node are used to achieve the aimed reduction. The DRP is based on three techniques that are implemented in every sensor node algorithm. The presented sensor node algorithm steps are as follows:

4.1.1 | Data equality (DE)

Data equality (DE) is the first step of the DRP algorithm to check whether the new sensed reading is equal to the previous reading or not, as specified by Equation (1):

$$z_t - l_{x-1} = 0 \quad (1)$$

where z_t is the current reading, and l_{x-1} is the previous reading.

Firstly, DRP cached a predefined number of readings l of each sensor node in the network and transmitted them to the sink node. After that, each new sensed reading z_t at time t of sensor s_i is compared with the previously collected reading l_{x-1} by the same sensor node. Thus, the current reading z_t is discarded if no change is detected. Otherwise, the second step of the proposed algorithm will start processing.

4.1.2 | Data deviation computation (DDC)

Data deviation computation (DDC) after DRP assures that the current sensed reading z_t has some deviation from the previous reading l_{x-1} . DDC aims to compute the value of this deviation and transmit or discard the reading accordingly. Indeed, in the

proposed DRP, two different processes are proposed to calculate the DDC. The first process aims to calculate the deviation between the current sensed reading z_t and the previous readings l_{x-1} based on Equation (2). If the deviation between z_t and l_{x-1} less than the user predefined e_{\max} , then the data transmission will be discarded, and the cache will be updated. Otherwise, the second DDC process starts. A DCC is presented to calculate the deviation of the current sensed reading from their predicted values. Since Kalman-Filter estimated values are highly accurate, the idea behind this process is to compare the current reading z_t with the Kalman-Filter-based estimated value est_t , which is almost the same as the previous reading. The deviation between z_t and est_t is calculated by Equation (3). If deviation E_{dev} is bigger than the predefined user threshold e_{\max} , then z_t is transmitted to the sink node; else, the z_t data transmission is discarded, and the cache is updated:

$$Vdev_t = z_t - l_{x-1} \quad (2)$$

$$Edev_t = |z_t - est_t| \quad (3)$$

4.1.3 | Faulty data detection (FDD)

FDD is used to eliminate the transmission of faulty sensed readings. Indeed, wireless sensor nodes are prone to failure due to the limitations of their resources. Therefore, assuring the collected data to be fault-free is essential for data accuracy and reliability. In this step, the proposed FDD technique is based on Equations (4)–(6). Since WSNs are apt to faults, fault detection is an essential process. Unlike many state-of-the-art data reduction approaches, the proposed approach considers fault detection processes:

$$dis = \sum_{i=0}^n |z_t - l_i| \quad (4)$$

$$corr = |dis - (l_{\max} - l_{\min})| \quad (5)$$

$$z_t = \begin{cases} \text{transmitted} & \text{if } dis < \theta \\ \text{discarded} & \text{otherwise} \end{cases} \quad (6)$$

where z_t denotes the current sensed reading, θ is a user predefined value based on the application needs, and l_{\max} , l_{\min} are the maximum and minimum cached readings, respectively. dis denotes the distance between the current reading and the cached values. The $corr$ is the deviation between the current sensed reading with the pre-cached readings to the deviation between the maximum cached reading with the minimum cached reading. According to Equations (4)–(6), faulty data transmission is discarded, and the cache is updated with the estimated value.

Algorithm 1 complexity is $O(1)$ and it represents the proposed DRP, and the operational flow chart of DRP is illustrated in Figure 2.

Input: The sensor current readings z_t
Output: The transmitted readings

```

1 set  $\theta$  = User-defined threshold
2 set  $l \leftarrow []$  User-defined cache
3 set  $e_{max} \leftarrow$  User-defined
4 set  $l_{x-1} \leftarrow$  the last cached value
5 set  $CR_t \leftarrow$  the new sensor readings
6 if  $z_t = l_{x-1}$  then
7    $z_t$ .dicarde transmission()
8    $est \leftarrow$  KF.update( $z_t$ )
9 else
10  if  $|z_t - l_{x-1}|$  or  $|z_t - est| < e_{max}$  then
11     $z_t$ .dicarde transmission()
12     $est \leftarrow$  KF.update( $z_t$ )
13    del  $l_0$ 
14     $l.append(z_t)$  or  $l.append(est)$ 
15  else
16    compute dis based on Equation (4)
17    compute corr based on Equation (5)
18    if  $corr > \theta$  then
19       $z_t$ .dicarde transmission()
20       $est \leftarrow$  KF.update( $est$ )
21      del  $l_0$ 
22       $l.append(est)$ 
23    else
24       $CR_t$ .transmitted()
25       $l.append(CR_t)$ 
26       $est \leftarrow$  KF.update( $z_t$ )

```

ALGORITHM 1 Data reduction phase (DRP)

4.2 | Data prediction phase (DPP)

In WSNs, all the data received by the sink node are sensed, produced, and transmitted by the sensor node. Typically, the sink nodes are more capable in terms of processing units, energy, and transmitters and perform tasks better than the sensor nodes [24]. Therefore, the proposed approach's main goal is to balance the data reduction and energy consumption on one side and the data reliability and accuracy on the other side. The proposed DRP focuses on data reduction in WSNs, as illustrated in Algorithm 1, by minimizing the data transmissions to the sink. Thus, the sink node received data are incomplete compared to the data collected by sensor nodes, which, in turn, affects the WSN data accuracy and reliability. To overcome these issues, DPP is proposed to predict the non-transmitted data at each time interval based on the Kalman filter.

DPP was developed to predict the non-received readings of sensor node s_i . In DPP, data prediction is the process of reproducing the non-transmitted readings based on the pre-received cached readings. Similar to DRP, every l reading of sensor s_i is cached by DDP. The l cached values will be updated by the received reading at time t of each sensor node s_i . If there is no reading received, the DPP will predict the reading based on two techniques: neighbouring-based prediction (NP) and self-based prediction (SP). Algorithm 2 presents more explanation of the cached values.

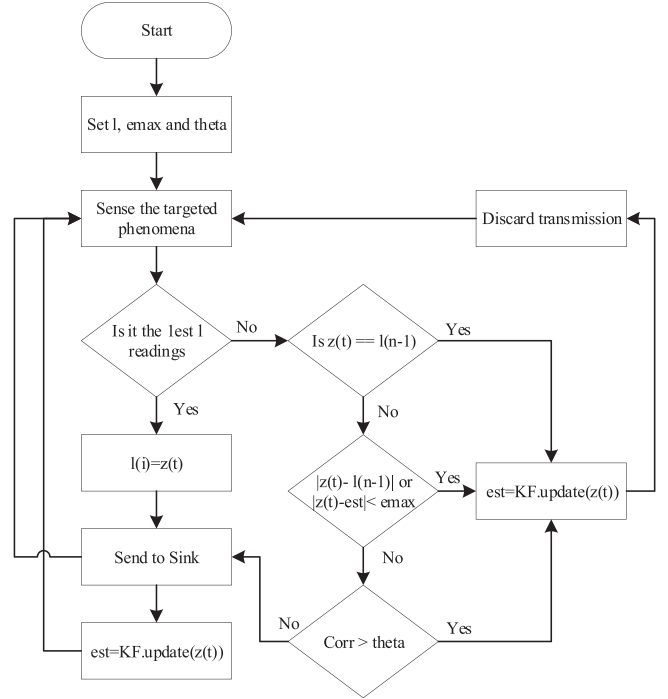


FIGURE 2 The operational flow chart of the proposed data reduction phase (DRP)

Input: The RR_t of total sensors of a given network
 $S = (s_1, s_2, \dots, s_n)$

Output: The transmitted readings

```

1 set  $RR_t \leftarrow$  The received reading from the sensor  $s_i$  at time  $t$ 
2 set  $Nlast \leftarrow$  The last received readings of neighbor
3 set  $Slast \leftarrow$  The last received readings of sensor  $s_i$ 
4 set  $e_{max}$  = User predefined value
5 set  $l$  = User predefined cache value
6 set  $s_{received} \leftarrow []$ 
7 if  $RR_t = \emptyset$  then
8   for each  $s_i$ .Neighbors do
9     if  $Nlast - Slast < e_{max}$  then
10       $Sest \leftarrow$  KF.Update( $Nlast$ )
11      del  $l_0$ 
12       $l.append(Nlast)$ 
13    else
14       $Sest \leftarrow$  KF.Update( $Slast$ )
15      del  $l_0$ 
16       $l.append(Slast)$ 
17 else
18    $s_{received}.append(RR_t)$ 
19   KF.Update( $RR_t$ )
20   del  $l_0$ 
21    $l.append(RR_t)$ 
22 Exit

```

ALGORITHM 2 Data prediction phase (DP)

4.2.1 | Neighbouring-based prediction (NP)

The non-transmission between the sensor node and the sink node may happen for two reasons. It may be reduced during the process of the proposed DRP by faulty or predictable

readings. In contrast, the second reason may happen due to network failure. In case of failure or faulty, the predicted reading may not be accurate enough. To avoid this scenario and increase prediction accuracy, the NP aims to check whether one of the targeted sensor neighbouring nodes is transmitting readings. In the case of one or more neighbouring transmit readings, the algorithm calculates the Jaccard similarity between the targeted sensor and its neighbours based on Equation (7). If the similarity is less than the predefined e_{\max} then the reproduced non-transmitted reading process (predicted value) will be based on the neighbour's received reading. Otherwise, the self-based prediction process will predict the non-transmitted readings:

$$J(s_i, s_j) = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} = \frac{|s_i \cap s_j|}{|s_i| + |s_j| - |s_i \cap s_j|} \quad (7)$$

4.2.2 | Self-based prediction (SP)

This technique aims to predict the non-transmitted readings passed through the first step. Indeed, the nature of the network plays a critical role in determining the prediction technique. In some wireless sensor networks, the sensor nodes are spatially redundant, which increases the similarity. In contrast, some others are established systematically to avoid such redundant problems. Nevertheless, the proposed data prediction algorithm evaluates the prediction in both cases. Both NP and SP used the Kalman Filter for data prediction. As mentioned in the system description section, each sink node received data from n sensor nodes. After the initial setup of the proposed approach at the sink node, the DPP collects the sensor nodes' information such as sensor location coordinates, sensor sensing range, sensor neighbouring, etc. In this paper, both the sensor and sink nodes are required to be preconfigured to operate synchronously. Therefore, the sink node runs each time to check out whether the targeted sensor node readings are received or not. If no readings are received, the Kalman filter will be used to forecast the non-received values. In fact, the Kalman filter estimated values have two different initializations. If the targeted sensor neighbours have readings, then the Kalman filter updated value will be the neighbour readings based on Equation (8). Otherwise, the updated value of the Kalman filter will be the last reading of the targeted sensor:

$$sr = \begin{cases} KF.update(Nlast) & |Slast - Nlast| > e_{\max} \\ KF.update(Slast) & \text{otherwise} \end{cases} \quad (8)$$

where sr is the non-received reading, $Nlast$ is the neighbour's last received reading, and $Slast$ is the targeted sensor's last received reading. Algorithm 2 complexity is $O(n*m)$ and it represents the proposed DPP, and the operational flow chart of the DPP is illustrated in Figure 3.

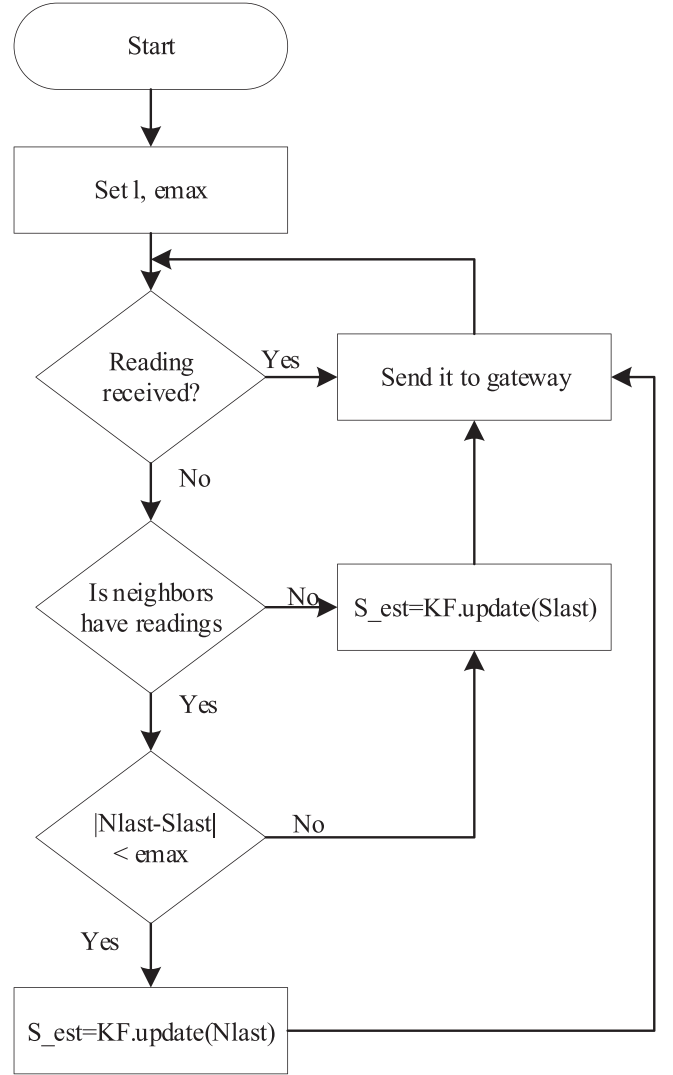


FIGURE 3 The operational flow chart of the proposed data prediction phase (DPP)

5 | IMPLEMENTATION AND RESULTS

The proposed approach is evaluated and compared with three prediction-based data reduction approaches: DP_LSTM, DRR-IoT, and LMS. The comparison is done based on three aspects: data reduction percentage, data prediction accuracy, and energy consumption. One of the main contributions of the proposed approach is detecting faulty readings and discarded transmitting them. Thus, the selected data was tested in its normal situation and with a faulty injection of 10%.

5.1 | Datasets

In this paper, the dataset from Intel Berkeley research lab (IBRL) is used. The dataset consists of real sensor nodes that collected data between February and April, 2004. The 54 Mica2Dot sensor nodes were used to collect the surrounding

weather data such as temperature, humidity, light, and voltage data. These Mica2Dot sensor nodes are located in the IBRL facilities, with a sampling rate of approximately 31 s. Only 10,000 collected humidity readings were included in the simulations. Sensors 1, 2 and 3 were used to evaluate the proposed approach. For more information on the datasets used, the reader is referred to [38].

5.2 | Simulation setting

To validate the proposed approach under realistic WSN deployment conditions, the Intel Berkeley research lab data set was presented for evaluation. The simulation code for the proposed approach is written using the Python programming language. Humidity readings from Intel Lab datasets were selected since they are typical time-stamped weather data. Sensors 1–3 humidity readings are used to evaluate the proposed approach. First, downsampling and fitting are used to fit the missing data and fill the gaps since some data points are faulty or missing in the collected data. Second, we created a faulty dataset of the three targeted sensors using Additive Gaussian noise (1 dB). Finally, since prior information is needed for the Kalman filter to work correctly, the first value of each sensor reading has been selected for initializing the state x of the Kalman filter (the first reading of each sensor).

It is worth mentioning that all benchmark approaches are based on user predefined values. Therefore, the e_{\max} values are set as (0.03, 0.05, 0.07, and 0.09) for the proposed approach, DP_LSTM, DDR-IoT, and LMS. In addition, the aforementioned simulation settings are applied to the benchmark's approaches, including DP_LSTM, DDR-IoT, and LMS.

5.3 | Results and analysis

In this section, the obtained results from data reduction and prediction algorithms are analysed using different e_{\max} values (0.01, 0.05, 0.09). The proposed data reduction performance is evaluated by comparing the input and output data size in percentage terms. Furthermore, to ensure the proposed approach's efficiency, it was evaluated based on ten thousand humidity readings of sensors 1, 2, and 3. Moreover, the obtained results of three well-known data reduction approaches are compared with the results of the proposed approach to ensure the effectiveness of the proposed data reduction approach. Data reduction percentage, data accuracy, and energy consumption are the selected metrics for evaluating the results of the proposed approach.

5.3.1 | Data reduction

To calculate the data reduction, Equation (9) is used to calculate the number of transmitted readings, and Equation (10) is used to calculate the percentage of transmitted readings based on the

total number of readings:

$$TTR = TCR - NTR \quad (9)$$

$$DPR = \left| \left(\frac{TTR}{TCR} * 100 \right) - 100 \right| \quad (10)$$

TTR is the total transmitted readings, TCR is the total collected readings (the size of the dataset), NTR is the non-transmitted readings, and DPR is the data reduction percentage. Tables 1–3 show the data reduction results of sensors 1, 2, and 3, respectively. They exemplify the evaluation of the proposed approach, DP_LSTM, DDR-IoT, and LMS with different e_{\max} values of (0.03, 0.05, 0.07, 0.09). To graph the data reduction in Tables 1–3, Figures 4 and 5 illustrate the proposed approach performance in comparison with DP_LSTM, DDR-IoT, and LMS $e_{\max} = 0.05$ and 0.09 in terms of data reduction. Every figure has three subfigures (a), (b), and (c) to represent sensors 1, 2, and 3, respectively. Moreover, different e_{\max} values of 0.05 and 0.09 are used as shown in Figure 4 and 5, respectively. The figures show that the proposed approach outperformed DP_LSTM, DDR-IoT, and LMS for real data with and without injecting faults for sensors 1–3.

In the same way, Figure 6 represents the data reduction of the proposed approach compared to DP_LSTM, DDR-IoT, and LMS using different values of e_{\max} with the real collected data without fault injection. Besides, Figure 7 illustrates the data reduction of the proposed approach compared to DP_LSTM, DDR-IoT, and LMS using different values of e_{\max} with fault injected data. Figure 7 clarify that the proposed approach achieved a significant data reduction percentage with faulty data up to 6,196 readings with $e_{\max} = 0.09$ while DP_LSTM, DDR-IoT, and LMS used the same e_{\max} and exact data of 2,086, 1,999, and 2,088 readings, in sensor 1, respectively. Besides, DP_LSTM, DDR-IoT, and LMS data reduction percentages decreased dramatically in the case of injecting 10% faulty data into the collected readings. On the other hand, the proposed approach increases the number of reduction readings up to 7514 with $e_{\max} = 0.09$, while DP_LSTM, DDR-IoT, and LMS achieved reduction, of 5499, 5339 and 4471 readings in sensor 1 with $e_{\max} = 0.09$, respectively. Similarly, the proposed approach achieves better results than DP_LSTM, DDR-IoT, and LMS in both sensors 2 and 3, as listed in Tables 2 and 3. Figures 4 and 5 plot the humidity data as the y-axis and the samples as the x-axis. It's noted that the proposed approach outperforms DP_LSTM, DDR-IoT, and LMS in terms of data reduction due to several reasons. First, all DP_LSTM, DDR-IoT, and LMS used data prediction to reduce transmissions, while the proposed approach to data reduction was based on three techniques, as discussed earlier. Secondly, besides data reduction, the proposed approach discarded the faulty collected data transmitted.

Besides, the data reduction percentage is represented based on Equations (9) and (10). Furthermore, Equations (11) and (12) are presented to calculate the data reduction accuracy percentage for the targeted sensors. Figure 6 illustrates the data

TABLE 1 Comparison of the number of transmissions of the proposed approach, DP_LSTM, DDR-IoT, and least-mean-square (LMS) and for sensor 1

e_{\max}	No. of readings	Normal data				Faulty data			
		Proposed approach	DP_LSTM	DDR-IoT	LMS	Proposed approach	DP_LSTM	DDR-IoT	LMS
0.03	10,000	2342	1604	1341	333	2017	611	517	173
0.05	10,000	4840	3498	3271	2031	4278	1464	1339	1118
0.07	10,000	6734	4326	4108	3061	5961	1876	1773	1828
0.09	10,000	7216	5027	4857	3977	6433	2306	2197	2360

TABLE 2 Comparison of the number of transmissions of the proposed approach, DP_LSTM, DDR-IoT, and least-mean-square (LMS) and for sensor 2

e_{\max}	No. of readings	Normal data				Faulty data			
		Proposed approach	DP_LSTM	DDR-IoT	LMS	Proposed approach	DP_LSTM	DDR-IoT	LMS
0.03	10,000	2492	1840	1561	492	2025	568	467	277
0.05	10,000	5034	3961	3755	2692	4173	1420	1299	1589
0.07	10,000	6467	4839	4654	3649	5679	1778	1684	2338
0.09	10,000	7514	5499	5339	4471	6196	2086	1999	2988

reduction percentage of the proposed approach compared to DP_LSTM, DDR-IoT, and LMS. It can be seen that a higher value of e_{\max} increases the reduction percentage and vice versa. Figure 6a represents the reduction percentage of sensor 1. The achieved data reduction percentage of the proposed approach is the highest compared to DP_LSTM, DDR-IoT, and LMS. As can be seen in Table 4, the achieved data reduction percentage of the proposed approach is between 24.92% and 75.14% for the values of e_{\max} (0.03, 0.05, 0.07, 0.09). Concerning DP_LSTM, it achieves a data reduction percentage between 18.4% and 54.99% using the same e_{\max} values of (0.03, 0.05, 0.07, 0.09). In contrast, the data reduction percentage achieved by DDR-IoT is between 15.61% and 53.39%. With regards to LMS, it shows a data reduction percentage ranging between 4.92% and 44.71%.

Tables 4–6 summarize the simulation results of the archived data accuracy and data reduction percentage for sensors 1, 2, and 3, respectively. They clarify the comparison results of the proposed approach with DP_LSTM, DDR-IoT, and LMS using different e_{\max} values of (0.03, 0.05, 0.07, 0.09).

In the same way, Tables 5 and 6 illustrate the achieved data reduction percentage of the proposed approach compared

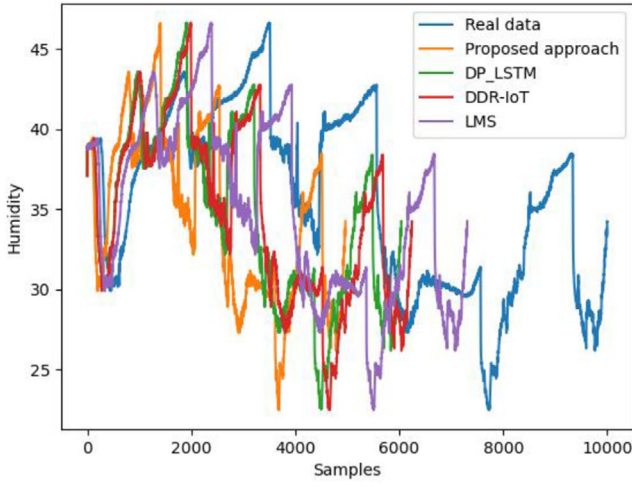
to DP_LSTM, DDR-IoT, and LMS for both sensors 2 and 3, respectively. It can be seen that the proposed approach reduction percentage ranges between 23.42%–72.16% and 24.99%–75.75%, respectively. The data reduction percentage of DP_LSTM goes between 16.4% and 50.27%, as shown in Table 5, and 6.72% and 53.75% in Table 6. Regarding DDR-IoT, the data reduction percentage ranges between 13.41%–48.57% and 13.51%–51.91% for Tables 5 and 6, respectively. Finally, the LMS data reduction percentage is the worst overall approach, ranging between 3.33% and 39.77% in Table 5 and 5.27% and 45.34% in Table 6.

Figure 6a,6c plot the obtained data reduction results shown in Tables 4–6. It is clear that the achieved data reduction percentage of the proposed approach is the highest compared with DP_LSTM, DDR-IoT, and LMS, while DP_LSTM achieves a higher data reduction percentage than DDR-IoT. In contrast, LMS achieved a worst data reduction percentage than DDR-IoT.

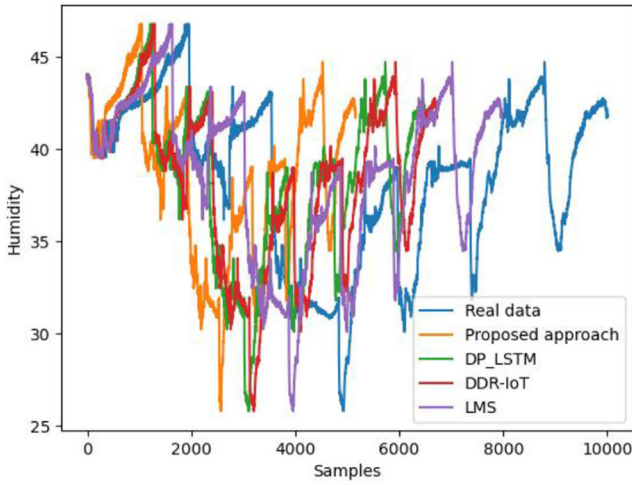
Once it comes to the data reduction percentage obtained with faulty data, Figure 7a,b, show that the highest reduction percentage is achieved by the proposed approach since fault

TABLE 3 Comparison of the number of transmissions of the proposed approach, DP_LSTM, DDR-IoT, and least-mean-square (LMS) and for sensor 3

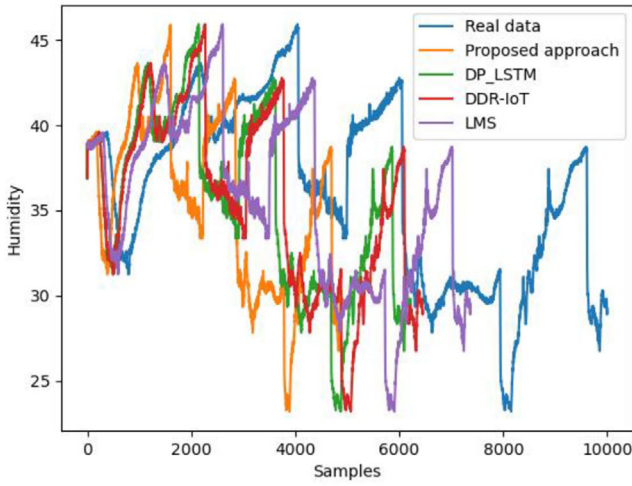
e_{\max}	No. of readings	Normal data				Faulty data			
		Proposed approach	DP_LSTM	DDR-IoT	LMS	Proposed approach	DP_LSTM	DDR-IoT	LMS
0.03	10,000	2499	1672	1351	527	2116	579	474	300
0.05	10,000	5074	3778	3548	2631	4415	1371	1278	1718
0.07	10,000	6695	4652	4455	3707	5924	1736	1629	2325
0.09	10,000	7575	5375	5191	4534	6681	2200	2115	2918



(a)

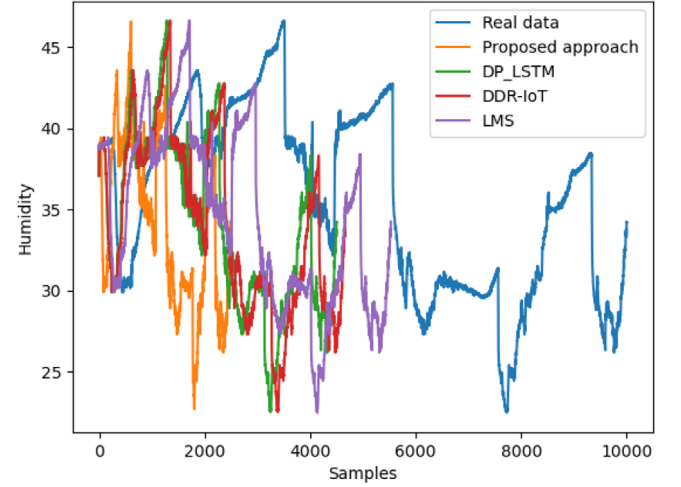


(b)

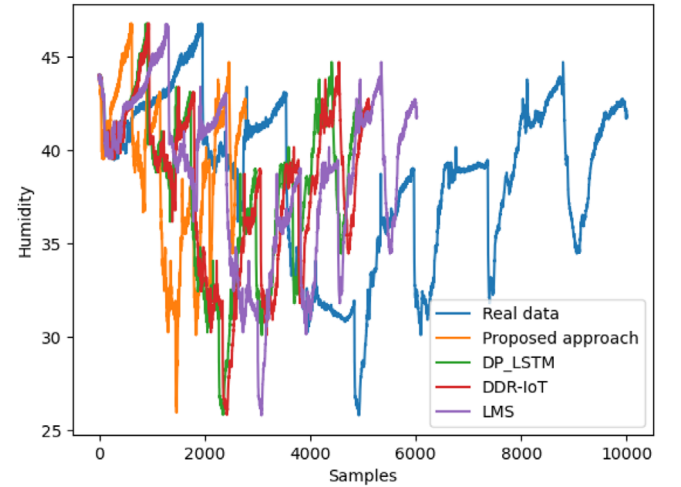


(c)

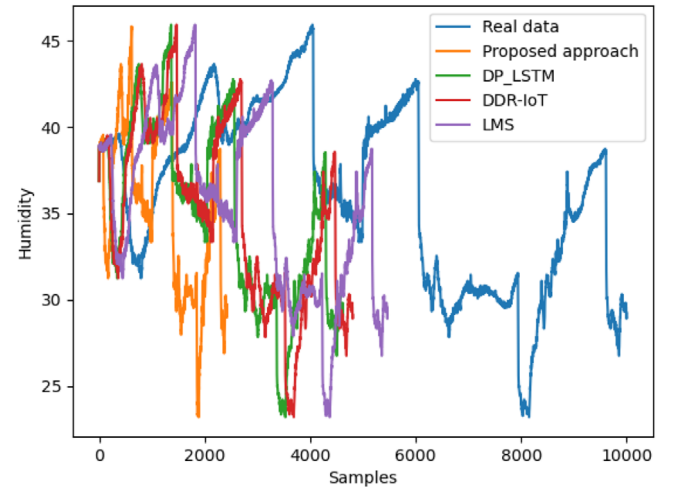
FIGURE 4 Comparison of the proposed approach, DP_LSTM, DDR-IoT, and LMS for sensors 1, 2, and 3 with $e_{\max} = (0.05)$ respectively. (a) Data reduction percentage of sensor 1 with $e_{\max} = 0.05$, (b) data reduction percentage of sensor 2 with $e_{\max} = 0.05$, (c) data reduction percentage of sensor 3 with $e_{\max} = 0.05$



(a)



(b)



(c)

FIGURE 5 Comparison of proposed approach, DP_LSTM, DDR-IoT, and LMS for sensors 1, 2, and 3 with $e_{\max} = (0.05)$ respectively. (a) Data reduction percentage of sensor 1 with $e_{\max} = 0.09$, (b) data reduction percentage of sensor 2 with $e_{\max} = 0.09$, (c) data reduction percentage of sensor 3 with $e_{\max} = 0.09$

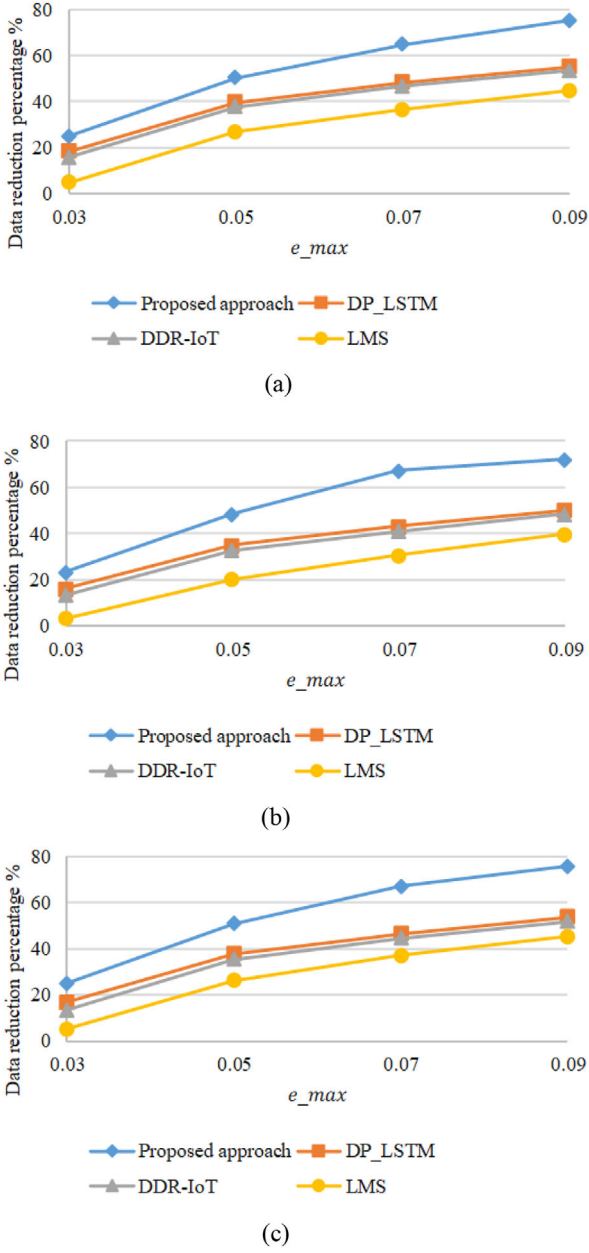


FIGURE 6 Data reduction percentage comparison of the proposed approach, DP_LSTM, DDR-IoT, and LMS for sensors 1, 2, and 3 with different e_{\max} values. (a) Data reduction percentage of sensor 1 without faulty readings, (b) data reduction percentage of sensor 2 without faulty readings, (c) data reduction percentage of sensor 3 without faulty readings

detection is one of the proposed approach's contributions. On the other hand, the LMS reduction percentage is higher than DP_LSTM and DDR-IoT since it decreased the number of transmissions based on the least square mean between the readings. In other words, the proposed approach shows a significant impact on data reduction with faulty data up to (66.81%) with e_{\max} 0.09. Besides all DP_LSTM, DDR-IoT, and LMS, the reduction percentage decreased dramatically when the collected data was injected with 10% faulty readings.

As can be seen in Table 4, the proposed approach realizes a data reduction percentage in the range between 24.92% and

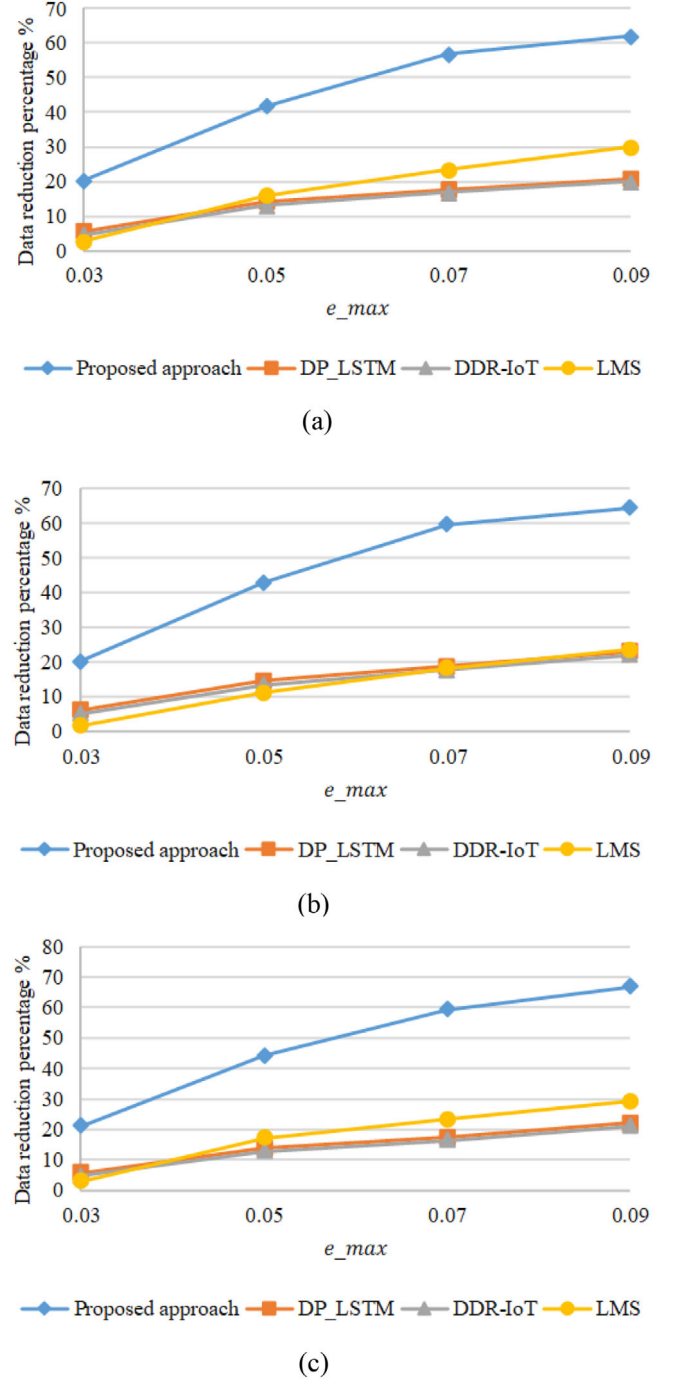


FIGURE 7 Data reduction percentage comparison of proposed approach, DP_LSTM, DDR-IoT, and LMS for sensors 1, 2, and 3 with different e_{\max} values. (a) Data reduction percentage of sensor 1 with 10% faulty readings, (b) data reduction percentage of sensor 2 with 10% faulty readings, (c) data reduction percentage of sensor 3 with 10% faulty readings

75.14% for the values of e_{\max} (0.03, 0.05, 0.07, 0.09). Concerning DP_LSTM, the achieved data reduction percentage is in the range from 18.4% to 54.99% using the same e_{\max} values of (0.03, 0.05, 0.07, 0.09). In contrast, the data reduction percentage achieved by DDR-IoT, ranges between 15.61% and 53.39 %. With regards to LMS, it shows a data reduction percentage ranging between 4.92% and 44.71%. In the

TABLE 4 Comparison of proposed approach, DP_LSTM, DDR-IoT, and least-mean-square (LMS) for sensor 1

ϵ_{\max}	Data reduction percentage %				Data Accuracy %			
	Proposed approach	DP_LSTM	DDR-IoT	LMS	Proposed approach	DP_LSTM	DDR-IoT	LMS
0.03	24.92	18.4	15.61	4.92	98.8	74.12	86.83	88.42
0.05	50.34	39.61	37.55	26.92	97.66	64.23	82.77	76.39
0.07	64.67	48.39	46.54	36.49	96.07	56.81	77.12	74.51
0.09	75.14	54.99	53.39	44.71	94.56	48.26	69.41	71.92

TABLE 5 Comparison of proposed approach, DP_LSTM, DDR-IoT, and least-mean-square (LMS) for sensor 2

ϵ_{\max}	Data reduction percentage (%)				Data accuracy (%)			
	Proposed approach	DP_LSTM	DDR-IoT	LMS	Proposed approach	DP_LSTM	DDR-IoT	LMS
0.03	23.42	16.04	13.41	3.33	98.71	79.18	78.7	87.6
0.05	48.4	34.98	32.71	20.31	96.23	63.45	75.22	75.67
0.07	67.34	43.26	41.08	30.61	95.16	54.38	70.24	74.16
0.09	72.16	50.27	48.57	39.77	94.5	44.71	62.6	71.54

same way, Tables 5 and 6 illustrate the percentage of data reduction achieved by the proposed approach compared to DP_LSTM, DDR-IoT, and LMS for both sensors 2 and 3, respectively. It can be seen that the proposed approach's reduction percentage ranges between 23.42%–72.16% and 24.99%–75.75%, respectively. The DP_LSTM data reduction percentage ranges between 16.4%–50.27% in Table 5 and 16.72%–53.75% in Table 6. Regarding DDR-IoT, the data reduction percentage ranges between 13.41%–48.57% and 13.51%–51.91% for Tables 5 and 6, respectively. Finally, the LMS data reduction percentage is the worst overall approach, ranges between 3.33%–39.77% in Table 5 and 5.27%–45.34% in Table 6.

Figure 6a,6c show the data reduction results obtained, which are presented in Tables 4–6. The proposed approach achieves a higher data reduction percentage than DP_LSTM, DDR-IoT, and LMS, while DP_LSTM achieves a higher data reduction percentage than DDR-IoT. In contrast, LMS achieved a worst data reduction percentage than DDR-IoT. Once it comes to the data reduction percentage obtained with faulty data, Figure 7a,7c show that the proposed approach achieved a higher reduction percentage since fault detection is one of the proposed approach's contributions. On the other hand,

LMS achieved a better reduction percentage than DP_LSTM and DDR-IoT since it decreased the number of transmissions based on the least square mean between the readings. In other words, the proposed approach shows a significant impact on data reduction with faulty data up to (66.81%) with ϵ_{\max} 0.09. Besides all DP_LSTM, DDR-IoT, and LMS, the reduction percentage decreased dramatically when the collected data was injected with 10% faulty readings.

5.3.2 | Data accuracy

Data accuracy is the similarity between the collected readings and the total transmitted readings, including predicted data. Thus, after applying the proposed approaches, the deviation between the obtained results and the input data. The data reduction accuracy percentage is calculated based on Equations (9) and (10):

$$DD = \left| \frac{\sum TCR - (\sum TTR + \sum IER)}{\sum TCR} \right| * 100 \quad (11)$$

TABLE 6 Comparison of proposed approach, DP_LSTM, DDR-IoT, and least-mean-square (LMS) for sensor 3

ϵ_{\max}	Data reduction percentage (%)				Data accuracy (%)			
	Proposed approach	DP_LSTM	DDR-IoT	LMS	Proposed approach	DP_LSTM	DDR-IoT	LMS
0.03	24.99	16.72	13.51	5.27	99.06	76.79	86.13	88.37
0.05	50.74	37.78	35.48	26.31	98.52	67.31	82.36	77.09
0.07	66.95	46.52	44.55	37.07	97.85	55.09	77.43	75.72
0.09	75.75	53.75	51.91	45.34	97.18	45.75	68.56	72.49

$$DA = (1 - DD) * 100 \quad (12)$$

where DD is the data deviation, TER is the total estimated readings of the sensor s_i and DA is data accuracy. TTR denotes the total transmitted readings, and TCR is the total collected readings.

According to Tables 4, 5, and 6, the proposed approach achieves a data accuracy percentage ranges between 98.80%–94.56%, 98.71%–94.50%, and 99.06%–97.18%, respectively, for the values of e_{max} (0.03, 0.05, 0.07, 0.09). Regarding DP_LSTM, it achieves the worst data accuracy percentage ranges between 74.12%–48.26%, 79.18%–44.71%, and 76.79%–45.75% as shown in Tables 4, 5, and 6. With regard to DDR-IoT, the data accuracy percentage ranges between 86.83%–69.41%, 78.70%–62.60%, and 86.63%–68.56%. While LMS shows a data accuracy percentage ranges between 88.42%–71.92%, 87.60%–71.54%, and 88.37%–72.49%.

Figures 8a, 8b, and 8c plot the obtained results shown in Tables 4, 5, and 6 graphically. When it comes to data accuracy, it is clear that the data accuracy percentage of the proposed approach is higher than DP_LSTM, DDR-IoT, and LMS. In contrast, LMS achieves a higher data accuracy percentage than DDR-IoT. In comparison, DP_LSTM achieved the worst data accuracy percentage than DDR-IoT. Figures 8a, 8b, and 8c show an illustrative demonstration of the obtained data accuracy results presented in Tables 4, 5, and 6.

Once it comes to accuracy, the proposed approach outperforms DP_LSTM, DDR-IoT, and LMS for the following reasons. First, the proposed approach considers that the transmitted readings may be affected by packet drops and loss, which can trouble the data prediction since the sink node cannot merely determine the reason for not receiving packets. Second, unlike the other approaches, the proposed approach discarded the transmission of the faulty data. Finally, there is no data loss in the proposed approach, while data comparison and fusion affect data loss.

5.3.3 | Energy consumption

Energy consumption plays an essential role in WSNs due to sensor node resource limitations. We evaluate the proposed approach's energy consumption against DP_LSTM, DDR-IoT, and LMS based on Equation (13) where TTR is the total transmitted readings and I_{tx} is the transmission current needed for one reading:

$$Energy\ consumption = TTR * I_{tx} \quad (13)$$

Figures 9a,c represent the proposed approach's energy consumption compared to DP_LSTM, DDR-IoT, and LMS in case of fault-free data. According to the results depicted in Figures 9a,c, the energy consumption achieved by the proposed approach is between 50.01 and 151.06 MJ for sensor 1, 56.07 and 154.01 MJ for sensor 2, and 93.05 and 167.55 MJ for sensor 3. DP_LSTM achieves energy consumption from

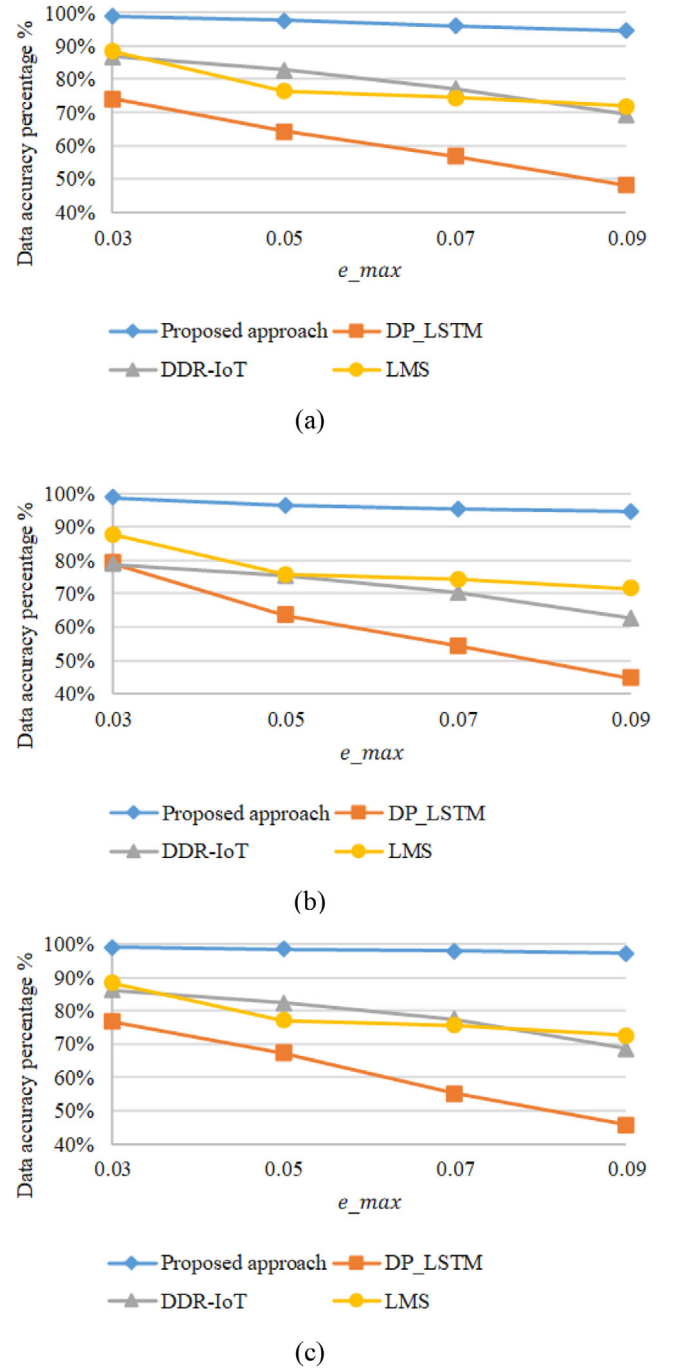


FIGURE 8 Data accuracy percentage comparison of proposed approach, DP_LSTM, DDR-IoT, and LMS for sensors 1, 2, and 3 with different e_{max} values. (a) Data accuracy percentage of sensor 1, (b) data accuracy percentage of sensor 2, (c) data accuracy percentage of sensor 3

93.77 to 169.76 MJ for sensor 1, 100.05 to 168.92 MJ for sensor 2, and 93.05 MJ to 167.55 MJ for Sensors3. In contrast, the energy consumption achieved by DDR-IoT ranges from 93.77 to 169.79 MJ for sensor 1, 103.47 to 174.21 MJ for sensor 2, and 96.75 to 174.01 MJ for sensor 3. Moreover, the energy consumption achieved by LMS is ranges between 111.24 and 191.30 MJ for sensor 1, 121.18 and 194.5 MJ for sensor 2, and 109.97 and 190.590 MJ for sensor 3.

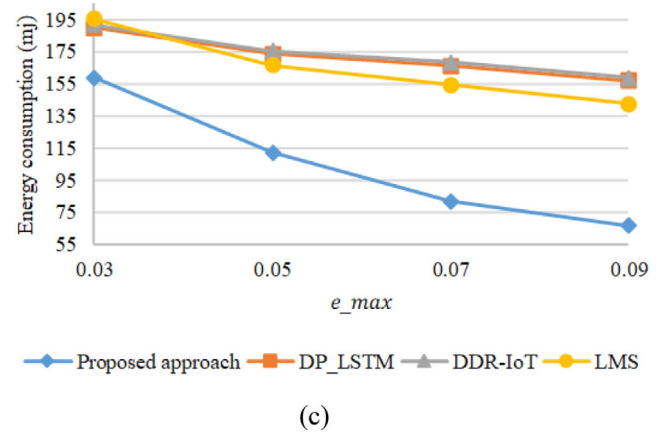
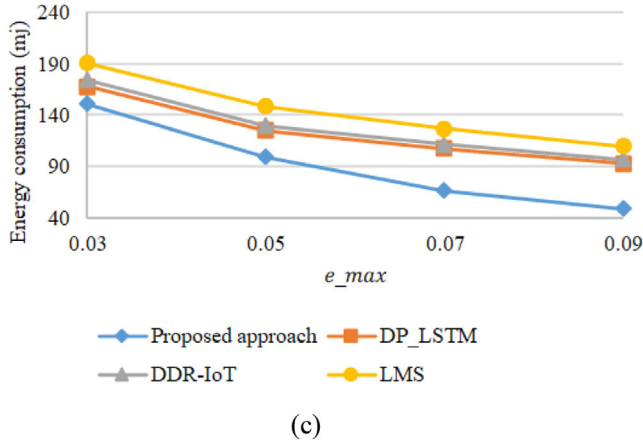
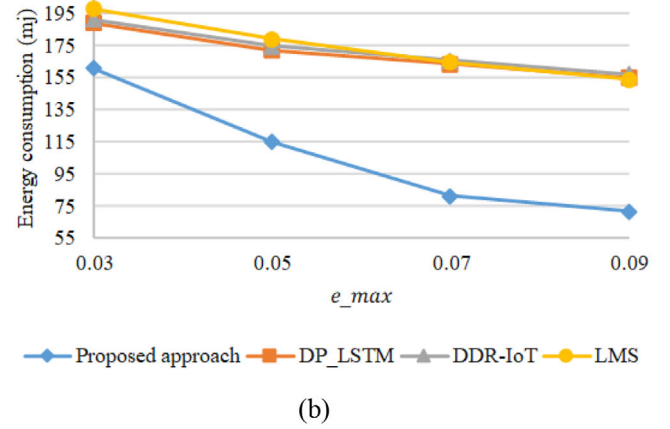
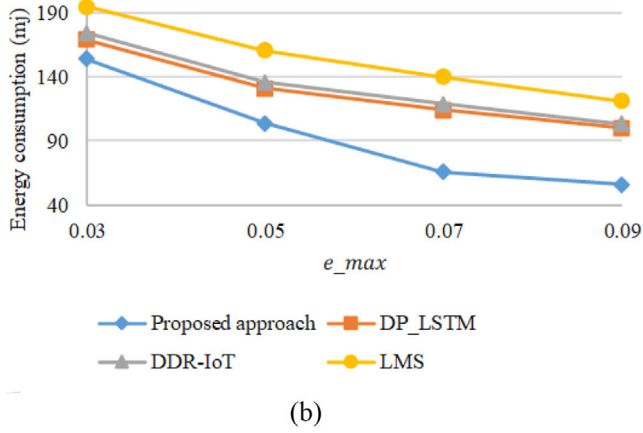
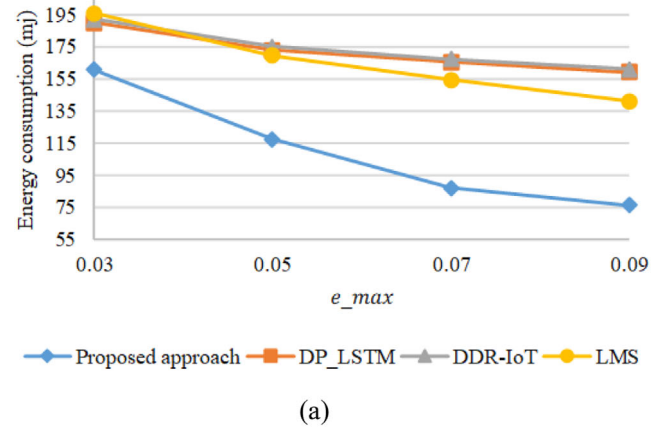
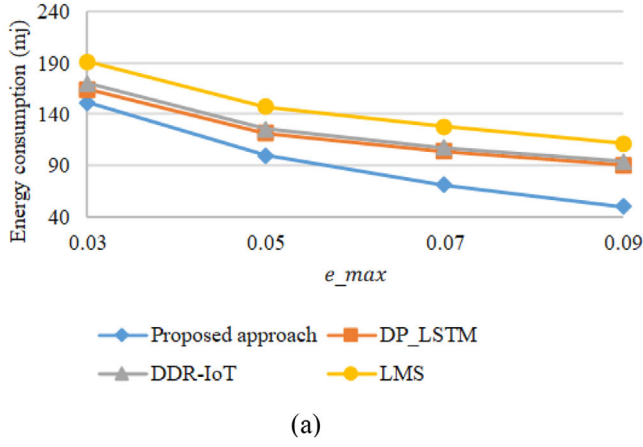


FIGURE 9 Energy consumption percentage comparison of proposed approach, DP_LSTM, DDR-IoT, and LMS for sensors 1, 2, and 3 with different e_{max} values. (a) Energy consumption percentage of sensor 1 without faulty readings, (b) energy consumption percentage of sensor 2 without faulty readings, (c) energy consumption percentage of sensor 3 without faulty readings

In contrast, Figures 10a, represent the proposed approach's energy consumption compared to DP_LSTM, DDR-IoT, and LMS if 10% of faulty readings are injected into the collected readings. According to the results in Figures 10a, the energy consumption achieved by the proposed approach ranges between 76.53 and 160.45 MJ for sensor 1, ranging between

FIGURE 10 Energy consumption percentage comparison of proposed approach, DP_LSTM, DDR-IoT, and LMS for sensors 1, 2, and 3 with different e_{max} values. (a) Energy consumption percentage of sensor 1 with 10% faulty readings, (b) energy consumption percentage of sensor 2 with 10% faulty readings, (c) energy consumption percentage of sensor 3 with 10% faulty readings

71.76 and 160.61 MJ for sensor 2, and ranging between 66.77 MJ and 158.62 MJ for sensor 3. The DP_LSTM energy consumption ranges between 159.23 and 189.76 MJ for sensor 1, range between 154.80 and 188.90 MJ for sensor 2, and ranges between 156.93 and 189.55 MJ for sensor 3. DDR-IoT shows energy consumption between 160.98 and 191.80 MJ for

sensor 1, a range between 156.99 and 190.72 MJ for sensor 2, and a range between 158.64 and 191.66 MJ for sensor 3. Moreover, LMS achieved energy consumption between 141.08 and 195.62 MJ for sensor 1, between 153.717 and 197.719 MJ for sensor 2, and between 142.49 and 195.16 MJ for sensor 3.

From Figures 9 and 10, we observed that the proposed approach achieved better energy consumption than the other approaches. It is also seen that increasing the value of ℓ_{\max} will decrease energy consumption and vice versa.

6 | CONCLUSION AND FUTURE WORK

In this paper, a dual prediction data reduction approach is proposed for WSNs. The proposed data reduction approach relies on two phases. The first phase is devoted to data reduction based on three techniques: Data equality and data deviation computation and faulty data detection. The second phase is based on the Kalman filter to improve the data reliability by estimating the filtered out data of the sensor nodes. The main objective of the proposed approach is to reduce the transmissions while balancing the data accuracy and reliability.

The obtained results showed that the proposed approach could reduce the data transmission by up to 75.75% while maintaining data reliability. In addition to data reduction, the proposed approach detects and eliminates faulty data. The proposed approach is compared with three different data prediction-based data reduction approaches, namely DP_LSTM and DDR-IoT, and LMS using 10,000 humidity real-world collected data. The proposed approach has achieved the highest efficiency in terms of data reduction, data accuracy, and energy consumption based on the obtained results. The future work will be extended to reconstruct the missing data that may occur due to the network's failure.

ACKNOWLEDGEMENT

This work is in part supported by the Fundamental Research Funds for the Central Universities (B200202216) and in part supported by Innovation Foundation of Radiation Application, China Institute of Atomic Energy (KFZC2020010401).

ORCID

Ammar Hawbani  <https://orcid.org/0000-0002-1069-3993>

REFERENCES

1. Tayeh, G.B., et al.: A spatial-temporal correlation approach for data reduction in cluster-based sensor networks. *IEEE Access* 7, 50669–50680 (2019)
2. Hawbani, A., et al.: Novel architecture and heuristic algorithms for software-defined wireless sensor networks. *IEEE/ACM Trans. Networking* 28(6), 2809–2822 (2020)
3. Wu, M., Tan, L., Xiong, N.: Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications. *Inf. Sci.* 329, 800–818 (2016)
4. Yemeni, Z., et al.: A DBN approach to predict the link in opportunistic networks. In: *Recent Developments in Intelligent Computing, Communication and Devices*. Springer, Berlin, pp. 575–587, (2019)
5. Lu, Y., et al.: Benefits of data aggregation on energy consumption in wireless sensor networks. *IET Commun.* 11(8), 1216–1223 (2017)
6. Jarwan, A., Sabbah, A., Ibnkahla, M.: Data transmission reduction schemes in WSNs for efficient IoT systems. *IEEE J. Selected Areas Communi.* 37(6), 1307–1324 (2019)
7. Yemeni, Z., et al.: Reliable spatial and temporal data redundancy reduction approach for WSN. *Comput. Networks* 185, 107701 (2021)
8. Tan, L., Wu, M.: Data reduction in wireless sensor networks: A hierarchical LMS prediction approach. *IEEE Sens. J.* 16(6), 1708–1715 (2015)
9. Mohamed, M.F., Ahmed, M.A., Nassar, H.: Lightweight energy-efficient frame-work for sensor real-time communications. *IET Commun.* 13(15), 2362–2368 (2019)
10. Salim, C., Mitton, N.: Machine learning based data reduction in WSN for smart agriculture. In: *International Conference on Advanced Information Networking and Applications*. vol. 1151, pp. 127–138. Springer, Berlin (2020)
11. Arbi, I.B., Derbel, F., Strakosch, F.: Forecasting methods to reduce energy consumption in WSN. In: *2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. Turin, Italy, 22–25 May 2017
12. Kulkarni, P.K.H., Jesudason, P.M.: Multipath data transmission in WSN using exponential cat swarm and fuzzy optimisation. *IET Commun.* 13(11), 1685–1695 (2019)
13. Singh, N.K., Kasana, A., Sachan, V.K.: Enhancement in lifetime of sensor node using data reduction technique in wireless sensor network. *Int. J. Comp. App.* 145(11), 1–5 (2016)
14. Bahi, J.M., Makhoul, A., Medlej, M.: A two tiers data aggregation scheme for periodic sensor networks. *Adhoc Sens. Wirel. Netw.* 21(1), (2014)
15. Du, T., et al.: A high efficient and real time data aggregation scheme for WSNs. *Int. J. Distrib. Sens. Netw.* 11(6), 261381 (2015)
16. Harb, H., et al.: A distance-based data aggregation technique for periodic sensor networks. *ACM Trans. Sens. Netw.* 13(4), 1–40 (2017)
17. Wu, H., et al.: A holistic approach to reconstruct data in ocean sensor network using compression sensing. *IEEE Access* 6, 280–286 (2017)
18. Basheer, A., Sha, K.: Cluster-based quality-aware adaptive data compression for streaming data. *J. Data Info. Qual.* 9(1), 1–33 (2017)
19. Silva, J.M.C., et al.: Litesense: An adaptive sensing scheme for WSNs. In: *2017 IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, Greece, 3–6 July 2017
20. Jon, Y.: Adaptive sampling in wireless sensor networks for air monitoring system. Dissertation. Retrieved from <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-295995>. (2016)
21. Makhoul, A., Harb, H.: Data reduction in sensor network performance evaluation in a real environment. *IEEE Embedded Sys. Lett.* 9(4), 101–104 (2017)
22. Makhoul, A., Harb, H., Laiymani, D.: Residual energy-based adaptive data collection approach for periodic sensor networks. *Ad Hoc Netw.* 35, 149–160 (2015)
23. Laiymani, D., Makhoul, A.: Adaptive data collection approach for periodic sensor networks. In: *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Sardinia, Italy, 1–5 July 2013
24. Raza, U., et al.: Practical data prediction for real-world wireless sensor networks. *IEEE Trans. Knowl. Data Eng.* 27(8), 2231–2244 (2015)
25. Li, S., Da Xu, L., Wang, X.: Compressed sensing signal and data acquisition in wireless sensor networks and internet of things. *IEEE Trans. Ind. Inf.* 9(4), 2177–2186 (2012)
26. Alam, M., et al.: Error-aware data clustering for in-network data reduction in wireless sensor networks. *Sensors* 20(4), 1011 (2020)
27. Ismael, W.M., et al.: An in-networking double-layered data reduction for internet of things (IoT). *Sensors* 19(4), 795–801 (2019)
28. Fathy, Y., Barnaghi, P., Tafazolli, R.: An adaptive method for data reduction in the internet of things. In: *2018 IEEE 4th World Forum on Internet of things (WF-IoT)*, IEEE, pp. 729–735, Singapore, 5–8 February 2018
29. Liu, X.Y., et al.: CDC: Compressive data collection for wireless sensor networks. *IEEE Trans. Parallel Distrib. Syst.* 26(8), 2188–2197 (2014)
30. Tayeh, G.B., et al.: A distributed real-time data prediction and adaptive sensing approach for wireless sensor networks. *Pervasive Mob. Comput.* 49, 62–75 (2018)

31. Jain, K., Agarwal, A., Kumar, A.: A novel data prediction technique based on correlation for data reduction in sensor networks. In: *Proceedings of International Conference on Artificial Intelligence and Applications*. pp. 595–606 Springer, Singapore (2021)
32. Singh, A.P., Chaudhari, S.: Embedded machine learning-based data reduction in application-specific constrained IoT networks. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. New York, 30 March–3 April 2020
33. Mohanty, S.N., et al.: Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks. *Phys. Commun.* 40, 101097 (2020)
34. Radhika, S., Rangarajan, P.: On improving the lifespan of wireless sensor networks with fuzzy based clustering and machine learning based data reduction. *Appl. Soft Comput.* 83, 105610 (2019)
35. Shaker, B.N., Hazar, M.J., Alzaidi, E.R.: Machine learning based for reducing energy conserving in WSN. *J. Phys.: Conf. Ser.* 1530, 012100 (2020)
36. Elsayed, W.M., El Bakry, H.M., El Sayed, S.M.: Data reduction using integrated adaptive filters for energy-efficient in the clusters of wireless sensor networks. *IEEE Embedded Sys. Lett.* 11(4), 119–122 (2019)
37. Feng, L., Kortoçi, P., Liu, Y.: A multi-tier data reduction mechanism for IoT sensors. In *Proceedings of the Seventh International Conference on the Internet of Things*, pp. 1–8. Linz, Austria, 22–25 October 2017
38. Bodik, P., et al.: Intel lab data. Online dataset (2004)

How to cite this article: Wang, H., et al.: A reliable and energy efficient dual prediction data reduction approach for WSNs based on Kalman filter. *IET Commun.* 1–15 (2021) <https://doi.org/10.1049/cmu2.12262>.