IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 13, 2022, accepted May 10, 2022, date of publication May 16, 2022, date of current version May 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175522

A Combinational Data Prediction Model for Data Transmission Reduction in Wireless Sensor Networks

KHUSHBOO JAIN[1], ARUN AGARWAL[2], AND AJITH ABRAHAM[3,4], (Senior Member, IEEE)

[1]School of Computing, DIT University, Dehradun, Uttarakhand 248009, India
[2]Ramanujan College, University of Delhi, New Delhi 110019, India
[3]Machine Intelligence Research Laboratories, Auburn, WA 98071, USA
[4]Center for Artificial Intelligence, Innopolis University, 420500 Innopolis, Russia

Corresponding author: Khushboo Jain (khushboojain2806@gmail.com)

This work was supported by the Analytical Center for Government of the Russian Federation under Grant 70-2021-00143 dd. 01.11.2021 and Grant IGK000000D730321P5Q0002.

ABSTRACT **Background**: Data prediction methods in wireless sensor networks (WSN) has emerged as a significant way to reduce the redundant data transfers and in extending the overall network's lifetime. Nowadays, two types of data prediction algorithms are in use. The first focus on reassembling historical data and providing backward models, resulting in unmanageable delays. The second is concerned with future data forecasting and gives forward models, that involve increased data transmissions. **Method:**Here, we developed a Combinational Data Prediction Model (CDPM) that can build prior data to control delays as well as anticipate future data to reduce excessive data transmission. To implement this paradigm in WSN applications two algorithms are implemented. The first algorithm creates step-by-step optimal models for sensor nodes (SNs). The other predicts and regenerates readings of the sensed data by the base stations (BS). **Comparison**: To evaluate the performance of our proposed CDPM data-prediction method, a WSN-based real application is simulated using a real data set. The performance of CDPM is also compared with HLMS, ELR, and P-PDA algorithms. **Results:**The CDPM model displayed significant transmission suppression (16.49%, 19.51% and 20.57%%), reduced energy consumption (29.56%, 50.14%, 61.12%) and improved accuracy (15.38%, 21.42%, 31.25%) when compared with HLMS, ELR and P-PDA algorithms respectively. The delay caused by CDPM training is also controllable in data collection. **Conclusion:** Results advised the efficacy of the proposed CDPM over a single forward or backward model in terms of decreased data transmission, improved energy efficiency, and regulated latency.

INDEX TERMS Data prediction, energy efficiency, network lifetime, transmission suppression, wireless sensor networks.

I. INTRODUCTION

In WSN applications, SNs usually sense the environmental data at high frequencies. Continuous data transmissions cause SNs to consume a lot of energy. Since WSNs are battery-equipped, energy conservation becomes a key concern [1], [2]. Because radio communications require more energy at SNs than any other activity [3], [4], data reduction is becoming more popular as a means of conserving WSNs' limited energy resources [5], [6]. By minimizing duplicate data transfers, data prediction methods [7]–[9] will conserve constrained resources such as unnecessary communication overheads, energy consumption, etc [10]–[13]. Every SN in a prediction-based method trains prediction models based on sensed data values and forwards them to the BS. Then, the SN predicts and reconstructs sensed reading using the same model as the BS. If the prediction threshold is lower than the application in that case the data prediction model is not acceptable, as the total communication overhead of such models will be larger than the original application i.e., without data prediction [13], [15], [16]. Apart from these two kinds of techniques, there are a few more methods that are comparatively intricate in training data prediction models and whose feasibility has yet to be determined and is discussed

The associate editor coordinating the review of this manuscript and approving it for publication was Irfan Ahmed.

in detail in the work [17]. To summarize, data-prediction-based techniques face three challenges: unpredictable latency, increased transmission overheads, and difficulty in model training. To the best of our knowledge, solving all these issues and challenges is still is work in progress.

This research provides a combinational model that may be used to reconstruct historical data as well as predict future data. The number of data points used to train the model is adjusted to meet predetermined upper constraints on error and latency, ensuring that data quality and delay are tightly regulated. To eliminate unnecessary transmissions and enhance the energy efficiency, the model is used in both data regeneration of previous data and data prediction of future data. Two techniques are proposed in this work to implement the combinational model for real-world WSNs application. To generate optimal combinational models, a step-wise technique is developed for SNs. This approach can reduce the combinational model's computational load and increase its viability. For the BS, another data prediction and data regeneration technique are proposed. Extensive experiments are simulated on real-world WSN applications to evaluate the combinational model's performance. The model's energy efficiency is compared to three already existing techniques. Simulation findings demonstrate that the proposed method can effectively suppress data transmissions, reduce overall WSN energy consumption, and tightly limit the delay induced by training. The objectives of this proposed work are as follows:

- To eliminate unnecessary data transmissions by the data-prediction models which can the number of redundant data transfers and extend the overall network's lifetime.
- To enhance the energy efficiency, a combinational model is developed for both data regeneration of previous data and data prediction of future data.
- To provide excellent proficiency in terms of reduced data transmission, reduce energy consumption, and regulated latency by implementing and simulating the proposed combinational model.

The remainder of this proposed work is organized as follows. In Section II, we review the related research work. Section III offers the overall framework of the proposed combinational data prediction model (CDPM). Section IV proposes the CDPM in WSN applications and Section V discusses the implementation of CDPM. Section VI presents the energy model for the proposed work. Section VII includes the experimental setup, dataset, and performance metrics for the proposed work. In Section VIII, the experimental results and discussion is presented to demonstrate the effectiveness of the framework. Finally, Section IX presents the conclusions and further research.

II. RELATED WORKS

In many cases, transmitting all the sensed data is not a smart idea. Data transmission reduction is crucial to resolve some WSN issues, such as reducing energy consumption and eliminating redundant measurements. In this respect, this section presents the related work based on data prediction to reduce data transmission.

Zhao *et al.* [18] proposed a P-DPA algorithm that uses the valuable information of the potential law contained in periodicity as guidance to change the prediction values. P-DPA effectively improves the accuracy, reduces communication frequency, and prolongs WSN lifetime but it does not describe how to find attribute correlation, and control delay was not reduced. Tan *et al.* [19] proposed the predicting approach can predict the measured values both at the SN and at the BS. HLMS provides low energy consumption, reduced data transmission, and high data accuracy but only the temporal data prediction not spacial is predicted. The synchronization of the filters at SN and BS is unexplored in this work.

Makhoul *et al.* [20] proposed a data reduction model (KW) that allows SNs to adapt their sensing rates based on the data variance to eradicate similar readings from the vector by a similar function. A local aggregation algorithm was further introduced to reduce the size of the dataset before transmitting it to the BS. This model minimizes the data size for transmission over the WSN for energy conservation but does not apply correlation between the adjacent SNs. Al-Qurabat *et al.* [21] proposed an Adaptive data gathering Dimensionality reduction using the adaptive-piecewise constant-approximation (APCA) method, Sampling rate adaptation based on dynamic time warping (DTW) similarity, and Frequency reduction using symbolic aggregate approximation (SAX) method. APCA removes the redundant data and adapts the sampling rate following the environment conditions, conserves energy, and also prolongs network lifetime but has high complexity and requires more processing time.

Tayeh et. al. [22] proposed the Adaptive Sampling + Transmission Reduction (AS+TR) based data prediction technique which aims to reduce radio communication and data sensing by combining adaptive sampling and dual prediction mechanism techniques. AS+ TR reduces energy consumption and extends the overall network lifetime. The AS method does not compute the risk of data loss and replicated data. This work does not control delay also. Cheng et. al. [23] proposed a prediction model based on the two-directional Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) which is named as multi-node multifeatured (MNMF) prediction model. MNMF overcomes the issues of Spatial or Temporal correlations in the data collection method as the redundant data impose unnecessary burdens on both the SN and WSN. This method Reduces the energy consumption of SNs and extends the WSN lifetime with high prediction accuracy and reasonable prediction bias. Although it considers only homogeneous WSN applications and does not even control delay.

Jain *et al.* [24] proposed DA-AFM for reducing Correlated Spatial-Temporal Data, one at the SNs for determining Temporal Redundancies in data readings using both AFM and RD

TABLE 1. Comparison of various existing data transmission methods in WSN.

Techniques	References	Issue Addressed	Contributions	Limitations
Periodic Data Prediction Algorithm (P-PDA)	[18]	Energy consumption	1. Improve Accuracy 2. Reduces communication frequency 3. Prolongs WSN lifetime.	How to find attribute correlation is not discussed. Does not control delay
Hierarchical Least Mean Square (HLMS)	[19]	To schedule the data communications between SNs and the BS. Energy Consumption Network Lifetime	1. Energy conservation 2. Data transmission reduction 3. Data Accuracy	Only temporal data prediction not spacial is predicted. The synchronization of the filters at SN and BS is unexplored. Does not control latency.
Data reduction model based on Kruskal-Wallis's (KW) test	[20]	Energy saving Reduced Sensing Reduced Data Transmission	Minimizing the data size for transmission over the WSN for energy conservation.	Do not apply correlation between the adjacent SNs
Energy-Efficient Adaptive-Distributed Data-Collection (EADiDaC)	[21]	Continuous collection of large amounts of redundant sensed data.	Remove the redundant data and adapts the sampling rate following the environmental conditions. Conserves energy and Prolongs network lifetime	High complexity and required more processing.
Adaptive Sampling + Transmission Reduction (AS+TR)	[22]	Reduce radio communication and data sensing	Combination of the adaptive sampling and dual prediction mechanism techniques reduce energy consumption and extend the overall network lifetime.	Risk of data loss and replicated data Does not control delay
Multi-Node Multi-Feature (MNMF)	[23]	Spatial or Temporal redundancy in data collection. The redundant data impose unnecessary burdens on both the SN and WSN.	Reduces the energy consumption of SNs and extends the WSN lifetime. high prediction accuracy reasonable prediction bias.	Consider only homogeneous WSN application Does not control delay
Data Aggregation based Adaptive Frame Method (DA-AFM)	[24]	Data gathering overheads Energy Consumption	Exploits both spatial and temporal correlations. Reduce data transmissions and energy consumption Improves accuracy	Consider only homogeneous WSN application Does not control delay
Extended Cosine Regression (ECR)	[25]	Data gathering overheads is a bottleneck in scaling WSN applications Energy Consumption	Simple, structure-free and lightweight, and scalable data prediction model. Reduce data transmissions cycles and energy consumption Improves accuracy	Does not support cluster-based architecture Does not control delay
Data-gathering and aggregation with selective transmission (DGAST)	[26]	Periodic sensor networks lifetime Energy Preservation	Energy Preservation Extend periodic sensor network's lifetime	Complex computation and high memory usage Does not control delay
Extended Linear Regression (ELR) model	[27]	Data gathering overheads Energy Consumption	Reduce data transmission Energy Efficient Extends network lifetime	Does not support cluster-based architecture and scalability Does not control delay
Data Prediction Technique Based on Linear Regression Model (DP-LRM)	[28]	Data gathering overheads Energy Consumption Network Lifetime	Reduce transmission cost Maintains accuracy and integrity in reduced data Energy Efficient Extends network lifetime	High Algorithmic Complexity Does not support scalability and control delay
Data reduction approach based on Kalman filter	[29]	Reduce data transmission Energy Consumption	High data reliability Efficient and Effective data reduction Extends network lifetime	Does not support scalability and control delay Large computation overhead
HFQKLMS filter	[30]	Data Efficiency Energy consumption	Maintains accuracy in reduced data Energy Efficient Extends network lifetime	High Algorithmic complexity and computation. Does not control delay

TABLE 1. ***(Continued.)*** **Comparison of various existing data transmission methods in WSN.**

RCHST-IETSMP	[31]	Network's lifetime	Extend the WSN lifetime Reliable	No effective CH selection High Algorithmic complexity and computation. Does not consider scalability and controls delay
Data Transmission Reduction Method (DTRM)	[32]	Redundant data transmission leads to collisions, data loss, and energy consumption	Data Accuracy Reduced data transmissions. Low complexity costs Lightweight processing Limited memory footprint Robust and Effective	Single value comparison

and the other at the CHs for determining spatial redundancies using AFM and RD. This method Exploits both spatial and temporal correlations, reduces data transmissions and energy consumption, and improves accuracy but it considers only homogeneous WSN applications and does not control delay. Jain and Kumar [25] in 2020 proposed an ECR model which is based on a two-vector model to synchronize predicted data in the intra-cluster transmissions to evade cumulative error in continuous data predictions. In the initialization phase of the data collection cycle, it generates future data approximation and computes its prediction threshold error. ECR is a simple, structure-free and lightweight, and scalable data prediction model. It reduces data transmission cycles and energy consumption while maintaining accuracy but has high complexity.

Al-Qurabat and Idrees [26] proposed a DGAST that gathers sensor data periodically and divides the networks into rounds. Each round in DGAST is divided into four stages: data collection, data aggregation, selective transmissions, and modifying the frequency of samples obtained for SNs. DGAST preserves energy and extends the periodic sensor network's lifetime but has complex computation and high memory usage. Jain *et al.* [27] in 2021 proposed an ELR model which exempts the SN from the transmission of huge volumes of data for a specific duration during which the BS will predict the future data values and thus minimize the energy consumption of WSN. ELR is an energy-efficient model which reduces data transmission and extends network lifetime but it does not consider cluster-based topology, scalability, and control delay.

Agarwal *et al.* [28] proposed a DP-LRM model that reduces the data transmission of redundant data by developing a regression model on linear descriptors on continuous sensed data values and is built on top of any data aggregation model. It uses a buffer-based linear filter algorithm that compares all incoming values and establishes a correlation between them. DP-LRM is an energy-efficient model which successfully reduces the data transmission cost and maintains accuracy and integrity in reduced data but it is Complex computation and does not consider scalability. Wang *et al.* [29] proposed a data reduction approach based on the Kalman filter This method performs data reduction through two phases: data reduction phase and data prediction phase. This is an efficient and effective data reduction that is reliable, energy-efficient, and extends network lifetime but has large computation overhead and it does not consider cluster-based topology and network scalability.

Nels *et al.* [30] proposed the HFQKLMS filter was developed by integrating HFBLMS and QKLMS. The HFBLMS model is devised by integrating FC theory and the HLMS scheme. The prediction process is carried out using the HFQKLMS filter approach for data aggregation. This work is energy Efficient, maintains accuracy in reduced data, and extends network lifetime but has Complex computation and does not consider scalability. Famila et al [31] proposed the RCHST-IETSMP integrates two critical parameters that define energy and trust parameters via a Hyper—Erlang process for successful CH selection assisted by the benefits of Semi-Markoc prediction integrated with the Hyper Erlang distribution process. This work is reliable and extends the WSN lifetime but has complex computation and does not consider scalability and control delay.

Jain and Kumar [32] in 2021 proposed a DTRM is implemented on the CHs and can be used in combination with most data aggregation algorithms. This study eliminates temporal redundancies and correlations from data readings and allows the SN to transmit only a few data values, which increases data transmission effectiveness and reduces energy consumption. DTRM provides data accuracy, reduced data transmissions, low complexity costs, lightweight processing, limited memory footprint, robustness, and effectiveness but it is based on single value comparison. Table 1 compares all the above-discussed data transmission methods in WSNs with the well-known parameters.

Many methods have been proposed for data transmission reduction in WSNs, but the control delay is not yet introduced. In comparison to the methods and techniques discussed above, the strength of the proposed CDPM algorithm lies in its ability to control delay, and reduce energy consumption by achieving high data transmission suppression and reduced RMSE (improved data quality).

III. SYSTEM MODEL AND ASSUMPTION

This section presents the overall framework which includes the network model and assumption for the proposed combinational data prediction model (CDPM).

A. NETWORK MODEL

The WSN consists of N number of sensor nodes, $S = \{S_1, S_2, S_3, \ldots, \ldots, S_N\}$, and B as a base station (BS) positioned away from the sensing region. These SNs are deployed randomly such that each SN S_i senses and transmits the measurement to the BS in each time slot $t = \{t_1, t_2, t_3, \ldots, \ldots, t_i\}$. The Combinational Data-Prediction Model suggests that the correlated and duplicate sensed values of the SNs will be flushed and not transmitted to the BS. Thus, the sensed data which is deviated from the prediction error will only be sent to the BS. Then, the BS will predict the non-transmitted data. Thus, the task of the SNs is to sense the environment parameters and transmit them to the BS if it is outside the prediction budget and the task of the BS is to receive the communicated data and predict the non-transmitted values.

The data transmission protocol was not taken into account in this study. Rather, we presumed that the data transmission between the SN and the BS was device-to-device. As a result, data transfers between SN and BS are accomplished on time. So, at any time t_i if no data is obtained, it will be believed that it was discarded by the SN. Therefore, the CDPM will predict the non-transmitted data. It will be used for both data regeneration of previous data and data prediction of future data.

B. NETWORK ASSUMPTION

We have considered followed assumption for the CDPM model:

- The SNs are considered to be stationary and are randomly deployed in the sensing region
- The BS is positioned away from the sensing region.
- All SNs have fixed data sampling rate.
- The data transmission between the SN and the BS is considered device-to-device, which means, that the data transfers between SN and BS reach without any delay.
- Dissimilar to SNs, the BS has no power, memory, or processing constraints.

IV. COMBINATIONAL DATA-PREDICTION MODEL

In the case of slight variations in the data, recent research has indicated that linear Data- Prediction models outperform the others. In line with that, this work provides a linear Data- Prediction Technique based Combinational Data-Prediction Model. A generic version of the linear prediction model, as well as the proposed Combinational Data-Prediction Model (CDPM), are explained in this section.

A. GENERIC VERSION OF LINEAR DATA-PREDICTION MODEL

The frequently changing environmental data are represented as a function of time: $f(t) = d$ in a specific area of the physical world. Then the sensed readings of an SN can be denoted by the time-series reading as follow in Equation (1) below:

$$f(t_i) = SR_i, \quad i \in \{1, 2, 3, \ldots\ldots, \mathrm{N}\} \tag{1}$$

where $\{1, 2, 3, \ldots\ldots, \mathrm{N}\}$ is the epoch period in which the SN senses the environmental parameters.

The SN will send (t_i, SR_i) for N epochs without prediction. Environmental data is assumed to follow a short-term linear pattern in linear models. Then, as a linear function, sensor readings can be approximated as follow in Equation (2) below:

$$\widehat{SR} = \widehat{f(t)} = \begin{cases} m_1 t + c_1 & step_1 \leq t < step_2 \\ m_2 t + c_2 & step_2 \leq t < step_3 \\ & \ldots\ldots \end{cases} \tag{2}$$

We train to build the prediction model in each step function. Some methods generate backward data prediction models for past data re-construction at the end of each step, which means they generate m_{j-1} and t_{j-1} when $t = step_j$. After training, instead of using the original sensed reading in the previous step for data regeneration, the model's parameters should be uploaded to the BS, which causes delays. While some methods generate forward data prediction models at the start of each step for future data prediction, which means they generate m_j and t_j when $t = step_j$. After training, the model's parameters, as well as the original sensed reading, are uploaded to the BS, resulting in new transmissions.

B. PROPOSED COMBINATIONAL DATA-PREDICTION MODEL (CDPM)

We have considered followed assumption for the CDPM model: The CDPM algorithm updates data in every step which has two stages: the first stage is the training phase and the second stage is the data prediction phase. During the first phase, the proposed CDPM model is trained and developed on d data values and the CDPM model is communicated to the BS. In the second phase, the BS will predict the non-transmitted data. BS will re-construct sensor data of the first phase. If the prediction threshold is more than the predefined error, the CDPM model will be retrained, *i.e.*, the next step begins. The d_j represents the training data values in *the* j^{th} step. At least two data points are needed to develop a linear data-prediction model which implies $d_j \geq 2$. Thus, we have expressed the CDPM model as $(step_j, m_j, c_j)$. The CDPM model can be used to rebuild at least two data values, one of which has two points. In other words, three parameters of one model can represent at least four values of the sensed reading without requiring any further transmission.

In this model, the values obtained by regeneration are usually deferred due to the time required in CDPM's training phase. In real-time WSN-based applications, the SNs sensed the surroundings with a predefined frequency. The extreme delay that can be produced during the training phase is expressed in Equation (3). The delay in *the* j^{th} step is directly related to the number of data (d_j) in training data

$$delay_j = \left(d_j - 1\right) \Delta t \tag{3}$$

Here $delay_j$ represents the maximum delay in *the* j^{th} step and Δt represents the sensing period. The maximum delay can then be controlled by restricting the d_j in the training phase. Let us assume that the highest tolerable delay is $delay_c$ and the maximum values for the training phase are delimited as d_c. It is predefined by the following Equation (4) below:

$$d_c = \begin{cases} if\,(delay_c \geq \Delta t), & d_c = \left\lfloor \frac{(delay_c}{\Delta t} \right\rfloor + 1 \\ else, & d_c = 2 \end{cases} \tag{4}$$

V. IMPLEMENTATION OF COMBINATIONAL DATA-PREDICTION MODEL

The combinational data-prediction model is proposed for use in real-world WSNs in this section. For SN to train and update the CDPM model, we present a stepwise approach. Another technique for reconstructing and predicting the sensed readings is also proposed for the BS.

A. TRAINING OF COMBINATIONAL DATA-PREDICTION MODEL

We use the least square method (LMS) to reduce error to create the best precise linear prediction model in the training phase. We have calculated *AbsoluteError* (*AE*), which is the difference between the sensed reading and the predicted data as follows in Equation (5) below:

$$AE_i = SR_i - \widehat{SR_i} \tag{5}$$

Then we calculate the error in the training phase, we have evaluated the *residualSumofSquares* (*RSS*) as follows in Equation (6) below:

$$RSS = \sum (AE_i)^2 \tag{6}$$

The *RSS* will attain its minimum value, when the $\frac{\partial RSS}{\partial m} = 0$ and $\frac{\partial RSS}{\partial C} = 0$ as per the LSM. Thus, the values of m and c are computed as follow in Equation (6) and (7) below:

$$m = \frac{d_j \sum (t_i \times SR_i) - \sum (t_i) \times \sum (SR_i)}{d_j(\sum (t_i^2) - \sum (t_i)^2} \tag{7}$$

$$c = \frac{(\sum (t_i^2) \sum (SR_i) - \sum (t_i) \times \sum (t_i \times SR_i)}{d_j(\sum (t_i^2) - \sum (t_i)^2} \tag{8}$$

We can express least square method as a following function of basic operations expressed in Equation (9) below:

$$(m, c) = \{(t_1, SR_1), (t_2, SR_2), \ldots\ldots\} \tag{9}$$

Then, to decrease the error in the measurement by an SN, we have expressed *Root Mean Squared Error* (*RMSE*) as follows in Equation (10) below:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (AE_i)^2} \tag{10}$$

In the data-prediction phase, if the *AE* of the predicted data is greater than the predefined threshold ε then the combinational model will be reinstructed and updated. Since we have considered ε to be upper-bound of *RMSE*, the prediction error will always be within the threshold. The threshold value of $RMSE_j$ is calculated conferring the values of ε, where $RMSE_j$ represents the *RMSE* of the j^{th} training phase.

Lemma 1: *For* $\forall\, j$ *if* $(RSS_j \leq \epsilon^2 d_j)$ *then* $RMSE \leq \epsilon$

Since $RMSE = \sqrt{\frac{1}{N}(\sum_{i=1}^{N} RSS_j + \sum AE_k^2)}$; $\forall k$ and $AE_k \leq \epsilon$
therefore if $(RSS_j \leq \epsilon^2 d_j)$ *then* $RMSE \leq \epsilon$

Thus, according to *lemma-1*, the *RSS* threshold in the first (training) phase is directly related to the d_j. To create optimal combinational models, we present a forward stepwise method for SNs. Each training phase is divided into multiple steps using this algorithm. The LMS is used only whenever new data is sensed, to avoid a huge amount of concurrent computations. *Algorithm-1* states that: If the value of *RSS* is greater than the value of RSS_c, then it will return an earlier outcome. Whenever the value of d_j matches d_c, the *Algorithm-1* will return an up-to-date outcome.

Algorithm 1: CDPM Training Phase

Parameters: $\epsilon, t, d_c, \{SR_t, SR_{t+1}, SR_{t+2}, \ldots\ldots\}$

Procedure: (t, m, c)

```
ALGORITHM BEGIN
d = 1;
RSS_c = 2ε²;
while new sensor's reading is generated do
    d++;
    (m', c') = lsm(t, SR_t), ..., ..., (t + d − 1, SR_{t+d−1});
    RSS = Σ_{i=t}^{t+d−1} (SR_i − i × m' − c')²
    if (RSS > RSS_c)
        return (t, m, c);
    else
        m = m'; c = c'; RSS_c+ = ε²;
    end
    if d == d_c
        return (t, m, c);
    endif
end while
ALGORITHM END
```

The algorithmic complexity of the CDPM training phase is low. The worst-case complexity when only one reading is sensed will be $O(d_b)$, which is easy enough for SNs.

After the model is trained, it is then updated and forwarded to the BS; and later the trained model will be used for data prediction. Later, every predicted value will be compared with the newly sensed value to determine AE. In case AE exceeds ϵ, *Algorithm-1* will be iteratively called to retrain the model and the latest data prediction model will be updated.

B. DATA PREDICTION AND REGENERATION PHASE

When the BS obtains the trained values of the CDPM model from an arbitrary SN, the estimates of data in the first phase are regenerated. The BS then predicts the sensed data based on this. The *Algorithm-2* for Data Prediction & Regeneration phase is presented below.

Algorithm 2: CDPM Model for Data Prediction & Regeneration phase

Procedure: (t, m, c) , Δt

Parameters: $\{SR_1, SR_2, SR_3, \ldots\ldots\}$

```
ALGORITHM BEGIN
while new − model is obtained do
    SR_t = mt + c;
    for (i = t + 1; i ≤ present − epoch; i++)
        SR_t += m;
    endfor
    while no model is obtained do
            SR_{i++} += m;
            sleep(Δt);
    endwhile
end while
ALGORITHM END
```

Here t represents the epoch period of data sensing. The outcome of data-regeneration is expressed as $SR_1, SR_2, SR_3, \ldots\ldots\}$.

VI. ENERGY MODEL

The combinational data-prediction model is proposed for use in real-world WSNs in this section. For SN to train and update the CDPM model, we present a stepwise approach. Another technique for reconstructing and predicting the sensed readings is also proposed for the BS. In this section, we propose an energy model for CDPM: To calculate an SN's energy consumption, the energy consumed in each operation must be considered. As shown in Equation (11), the total energy consumption, in general, is related to four essential tasks:

i data sensing (E_{SEN}) which is the energy needed to sense one data value,
ii data transmission (E_{DT}) is the energy required by each SN per each communication round,
iii data aggregation (E_{DA}) is the energy needed to aggregate data, and
iv data prediction (E_{CDPM}) is the energy to perform data prediction by CDPM.

To estimate the total energy consumption of an SN (E_{Tot-SN}), we have used employed the model as discussed in the work [33].

$$E_{Tot-SN} = E_{SEN} + E_{DT} + E_{DA} + E_{CDPM} \tag{11}$$

Equation (12) evaluates the E_{SEN} which is the energy required to transform the physical data into digital one, where b is the number of bits in the sensed data, V is the supply voltage, I_S is the total current required in data sensing, and T_S is the total duration of data sensing.

$$E_{SEN} = bVI_ST_S \tag{12}$$

To evaluate the amount of energy dissipated by each SN per round of communication, the classical first-order radio energy model [33] has been employed for transmission and reception energy. The energy ingesting of SNs depends on the distance between the transmitter (SN) and the receptor (BS) in both free space (*fsp*) and multipath (*mph*). A threshold selects the channel, which is d^2 energy loss for a small distance and d^4 energy loss for a large distance. If a b-bit data has to be transmitted over a distance d, data transmission $E_{TX}(b, d)$ will be expressed by Equation (13).

$$E_{TX}(b, d) = \begin{cases} bE_{elec} + b\varepsilon_{fs}d^2, & ifd < d_0 \\ bE_{elec} + b\varepsilon_{mp}d^4, & ifd \geq d_0 \end{cases} \tag{13}$$

On receiving b-bit data, the energy ingesting will be computed by Equation (14).

$$E_{RX}(b) = bE_{elec} \tag{14}$$

E_{elec} is the energy used to send electronics for a transceiver which senses a single bit b. The coefficient of the free-space amplifier and multipath are ε_{fs} and ε_{mp} respectively. The threshold d_0 determines the energy consumption which is calculated as $d_0 = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}}$.

The energy dissipation to aggregate b bits is represented in Equation (15) as follows:

$$E_{AGG}(b) = bE_{DA} \tag{15}$$

Equation (16) presents the energy consumption in data prediction of CDPM (E_{CDPM}), where N_c is the number of data communication cycles, I_l is the leakage current, C_{avg} is the average capacitance switched per cycle, n_p is a constant value that depends on the SN capabilities, f is the frequency of data sensing and V_t is the thermal voltage.

$$E_{CDPM} = bVI_l\frac{N_c}{f}\frac{V}{n_p \times V_t} + bC_{avg}N_cV^2 \tag{16}$$

VII. PROFICIENCY ASSESSMENT

In this section, we present the simulation setup and proficiency metrics for the evaluation Combinational Data-Prediction Model (CDPM) in terms of transmission suppression, energy consumption, latency, and data quality. The performance of CDPM is also compared with P-PDA [15], HLMS [16], and ELR [22] algorithms.

TABLE 2. Parameters of energy model.

Symbol	Parameter	Value
b	Data packet	4496 bits
E_{TX}	Transmission energy	$150\ nJ/s for\ 1-bit, 10\ m$
E_{RX}	Reception energy	$50\ nJ/s\ for\ 1-bit$
E_{DA}	Aggregation energy	$5\ (nJ/bit)/s$
ε_{fs}	Free space amplifier energy	$10\ (pJ/bit)/m^2$
ε_{mp}	Multi-path fading amplifier energy	$0.0013\ (pJ/bit)/m^4$

TABLE 3. Simulation parameter.

Symbol	Parameter	Value
E_0	Initial energy	1 J
N	Number of SNs	54
-	Algorithms	CDPM, ELR, P-PDA, HLMS,
X	Network field length	500 m
Y	Network field breadth	500 m
-	Base station	Fixed coordinates $(X, Y/2)$
T	Simulation time interval	31 s

A. SIMULATION SETUP

We have implemented CDPM in network simulator NS-2.34 [34] along with ELR, P-PDA, and HLMS algorithms. NS-2.34 is an event-driven simulator that has aided in comprehending the dynamic nature of communication protocols. NS2 supports TCP, UDP, routing algorithms, network topologies, and multicast protocols on both wired and wireless networks [35]. NS2 is written in C++ and OTcl, which is an Object-oriented version of Tcl. The simulation parameters are shown in Table 3.

B. DATASET

The Intel Berkeley Research Laboratory (IBRL) [36] has approximately 2.3 million sensor measurements. Each SNs senses data after every 31 seconds. Several quantities are included in this dataset like temperature in degrees Celsius and humidity which ranges from 0-100%. The brightness is measured in Lux, and the voltage varies between 2 to -3 volts. The total readings in the dataset for temperature are 1048574, and for humidity is 104845. This simulation runs at each SN for five days to evaluate the performance of data prediction algorithms in humidity and temperature only. Linear interpolation is used to fill the missing values at different epochs.

C. PROFICIENCY METRIC

The proficiency of CDPM is evaluated by performing exhaustive experiments on the IBRL dataset and the following metrics are defined for them. Moreover, according to [27], [32], data transmission is the major issue for the energy depletion of such a network. Therefore, in the proposed CDPM model the energy consumption metric is estimated based on the number of data transmitted from SNs to the BS.

1) TRANSMISSION SUPPRESSION

It is the estimate of the ratio of the transmitted data by using any data prediction model with the actual sensed data without implementing any data prediction method.

$$\mathrm{TS}\% = \left(\frac{\mathrm{Transmitteddatabyusingpredictionalgorithm}}{\mathrm{actualsenseddata}}\right) \times 100 \tag{17}$$

2) ENERGY CONSUMPTION

The amount of energy consumed in a WSN is directly proportional to the number of radio communications carried out by the SNs. Reduced data delivered to the BS would considerably boost WSN lifespan. The greater the transmission suppression, the less data is transferred and the less energy spent. The energy model of this work is explained in detail in Section VI.

3) DATA QUALITY

Data quality is a critical element in defining excellence in the WSN. We have already expressed Root Mean Squared Error (RMSE) as a way to lessen the error of data sensed by any SN (RMSE).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(AE_i)^2} \tag{18}$$

where $AE_i = SR_i - \widehat{SR_i}$, SR_i is the senses reading of i^{th} SN and $\widehat{SR_i}$ is the predicted values of that SN.

4) LATENCY

The latency is the measure of the delay. In WSN, it is defined as the time taken by the data to transmit data from the SN and reach the BS. It has a key impact on the performance of any network.

5) ALGORITHMIC COMPLEXITY

An algorithm's complexity is defined as how the algorithm performs in different conditions. It is expressed numerically as a function of $T(n)$ time versus n input size [37]. Here we have estimated the algorithm's efficiency asymptotically. $T(n)$ time will be measured as the number of required "steps," given that each such step takes constant time.

VIII. RESULTS AND ANALYSIS

In this section, we present the simulation setup and performance metrics for the evaluation Combinational Data-Prediction Model (CDPM) in terms of transmission suppression, energy consumption, latency, data quality, and algorithmic complexity. The performance of CDPM is also compared with P-PDA [18], HLMS [19], and ELR [25] algorithms.

A. COMPARISON OF TRANSMISSION SUPPRESSION %

For experiments, CDPM, P-PDA, HLMS, and ELR algorithms are deployed to gather data for ten rounds of

TABLE 4. Transmission suppression % of average temperatures and average humidity of SNs.

S. No	Rounds	Threshold	Average temperatures of SNs				Average Humidity of SNs			
			CDPM	HLMS	ELR	P-PDA	CDPM	HLMS	ELR	P-PDA
1	ro_1	0.05	48.94	39.59	26.68	13.88	54.39	46.59	31.37	22.99
2	ro_2	0.10	65.35	47.48	34.35	25.68	67.66	53.78	38.16	33.55
3	ro_3	0.15	71.46	53.25	45.22	38.35	71.14	59.47	43.75	38.89
4	ro_4	0.20	76.77	58.39	54.55	49.66	72.66	60.69	58.18	45.77
5	ro_5	0.25	78.46	61.35	59.02	54.45	73.33	61.67	57.36	46.66
6	ro_6	0.30	79.13	63.13	61.75	57.35	74.61	62.36	56.65	47.44
7	ro_7	0.35	80.11	65.02	63.24	60.55	75.59	63.85	55.21	48.88
8	ro_8	0.40	81.02	67.91	65.51	63.33	76.25	64.58	54.77	49.69
9	ro_9	0.45	82.84	68.11	66.75	65.88	77.02	65.76	53.98	50.57
10	ro_{10}	0.50	83.68	69.88	67.35	66.46	77.85	66.17	52.02	49.15

communication. Each round has a varying threshold ε from 0.0 to 0.5 with a step function of 0.5. We determine the transmission suppression ($TS\%$) of four algorithms for average temperatures and average humidity of SNs as presented in Section VII in Equation (17). The larger the TS% will be, the fewer data will be transmitted and less energy will be consumed. The $TS\%$ of four algorithms for IBRL data for average temperatures and average humidity of SNs are visualized in Figures 1 and 2 respectively and are illustrated in Table 4. The Parameter settings of CDPM will be $delay_c$ is set to be 600 seconds. Since Δt of the IBRL dataset is 31 seconds and d_c is set to be 20. The $TS\%$ of CDPM is always higher than the $TS\%$ of P-PDA, HLMS, and ELR algorithms at any value of threshold in any round of communication. Furthermore, CDPM can guarantee that the $TS\%$ is always less than 100% which means that the additional transmissions are avoided. The network scales in IBRL applications are small enough that each SN can directly transfer data to the BS. Although, the default TCP packet size in NS2 is 12 packets which is a bottleneck in data transmission and scaling in such WSN applications. Thus, conferring to the message format of NS2, the message size of P-PDA, HLMS, and ELR algorithms are set to be 12 bytes each and for CDPM it is set to 10 bytes.

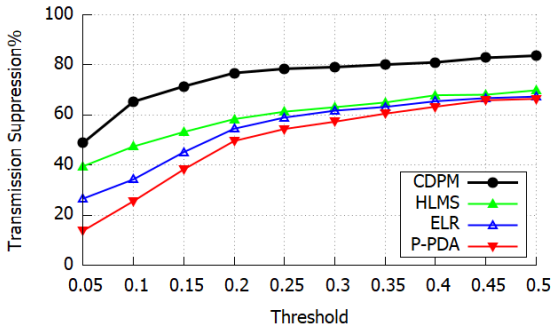

FIGURE 1. Transmission Suppression % (for Temperature) of CDPM, HLMS, ELR, and P-PDA while the threshold varies (0.05 to 0.50). TS% of CDPM is always high.

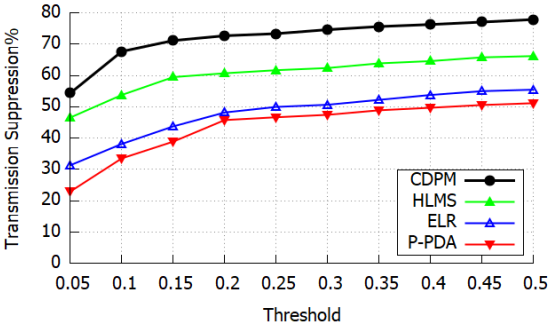

FIGURE 2. Transmission Suppression % (for Humidity) of CDPM, HLMS, ELR, and P-PDA while the threshold varies (0.05 to 0.50). TS% of CDPM is always high.

B. COMPARISON OF ENERGY CONSUMPTION

Most data prediction methods deliver reduced data transmission, so we also compare the energy consumption of CDPM with P-PDA, HLMS, and ELR algorithms. CDPM along with these three algorithms is deployed to gather data for ten rounds of communication where each round has a varying threshold ε from 0.0 to 0.5 with a step function of 0.5. We determine the energy consumption based on the energy model for both average temperatures and average humidity of SNs as described in section VI. The energy consumption of four algorithms for IBRL data for average temperatures and average humidity of SNs are illustrated in Table 5. The Parameter settings of CDPM will be $delay_c$ is set to be 10 minutes. Since Δt of the IBRL dataset is 31 seconds and d_c is set to be 20. It is observed that the energy consumption of CDPM is always higher than the energy consumption of P-PDA, HLMS, and ELR algorithms at any value of threshold in any round of communication. In IBRL applications, the network scales are small enough for every SN to transmit the data to the BS directly. The energy consumption of each SN for sending one-byte data is set to

TABLE 5. Energy efficiency of average temperatures and average humidity of SNs.

S. No	Rounds	Threshold	Average temperatures of SNs				Average Humidity of SNs			
			CDPM	HLMS	ELR	P-PDA	CDPM	HLMS	ELR	P-PDA
1	ro_1	0.05	12.23	16.45	25.65	40.87	5.34	8.85	10.71	15.14
2	ro_2	0.10	10.32	14.73	23.48	36.57	4.87	8.25	10.12	14.35
3	ro_3	0.15	9.65	13.16	21.65	34.83	4.42	7.92	9.65	13.73
4	ro_4	0.20	8.41	12.43	19.34	30.27	4.12	7.56	9.11	12.84
5	ro_5	0.25	7.93	11.65	17.75	27.93	3.98	6.98	8.79	11.53
6	ro_6	0.30	7.38	10.54	15.87	24.47	3.58	6.32	8.21	10.55
7	ro_7	0.35	6.82	9.66	13.23	21.84	3.27	5.87	7.72	9.81
8	ro_8	0.40	6.25	8.76	12.54	18.28	2.98	5.12	7.19	8.35
9	ro_9	0.45	5.79	7.98	11.45	15.85	2.63	4.76	6.53	7.65
10	ro_{10}	0.50	5.17	7.34	10.37	13.36	2.23	4.08	5.75	6.85

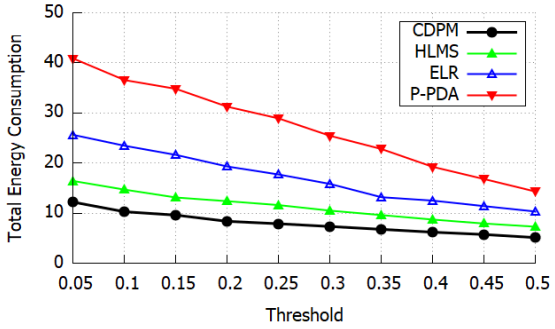

FIGURE 3. Energy consumption (for Temperature) of CDPM, HLMS, ELR, and P-PDA while the threshold varies (0.05 to 0.50). The energy consumption of CDPM is always low.

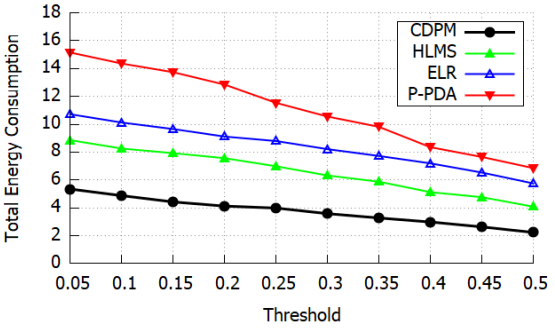

FIGURE 4. Energy consumption (for humidity) of CDPM, HLMS, ELR, and P-PDA while the threshold varies (0.05 to 0.50). The energy consumption of CDPM is always low.

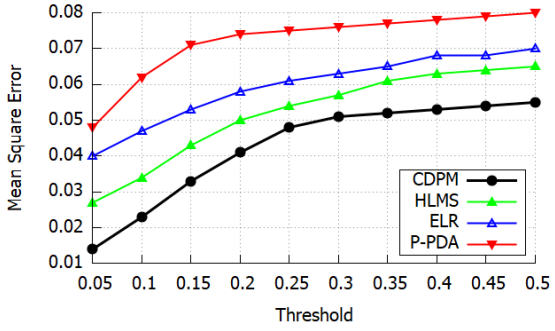

FIGURE 5. MSE (for Temperature) of CDPM, HLMS, ELR, and P-PDA while the threshold varies (0.05 to 0.50). Data Accuracy of CDPM is always high.

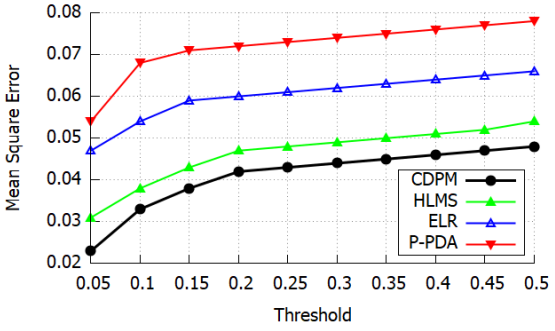

FIGURE 6. MSE (for Humidity) of CDPM, HLMS, ELR, and P-PDA while the threshold varies (0.05 to 0.50). Data Accuracy of CDPM is always high.

be 0.0144mJ and for receiving one-byte data is 0.0057mJ [38]. The cumulative energy consumption of algorithms after every round of communication for temperature and humidity are presented in Figures 3 and 4 respectively which illustrates that CDPM's energy consumption is much lower than other algorithms. These experiments have demonstrated that CDPM has greater data suppression rates and is more energy-efficient.

C. COMPARISON OF DATA QUALITY

The preceding experiments demonstrate that CDPM has greater energy efficiency and data suppression rates. Therefore, we also conduct experiments on the data quality by estimating the *RMSE* value as described in Section VII in Equation (18). For experiments, CDPM, P-PDA, HLMS, and ELR algorithms are deployed to gather data for ten rounds of communication where each round has a varying threshold ε from 0.0 to 0.5 with a step function of 0.5. The lower the

TABLE 6. Energy efficiency of average temperatures and average humidity of SNs.

S. No	Rounds	Threshold	Average temperatures of SNs				Average Humidity of SNs			
			CDPM	HLMS	ELR	P-PDA	CDPM	HLMS	ELR	P-PDA
1	ro_1	0.05	0.014	0.027	0.040	0.048	0.023	0.031	0.047	0.054
2	ro_2	0.1	0.023	0.034	0.047	0.062	0.033	0.038	0.054	0.068
3	ro_3	0.15	0.033	0.043	0.053	0.071	0.038	0.043	0.059	0.071
4	ro_4	0.2	0.041	0.050	0.058	0.074	0.042	0.047	0.060	0.072
5	ro_5	0.25	0.048	0.054	0.061	0.075	0.043	0.048	0.061	0.073
6	ro_6	0.3	0.051	0.057	0.063	0.076	0.044	0.049	0.062	0.074
7	ro_7	0.35	0.052	0.061	0.065	0.077	0.045	0.050	0.063	0.075
8	ro_8	0.4	0.053	0.063	0.068	0.078	0.046	0.051	0.064	0.076
9	ro_9	0.45	0.054	0.064	0.068	0.079	0.047	0.052	0.065	0.077
10	ro_{10}	0.5	0.055	0.065	0.070	0.080	0.048	0.054	0.066	0.078

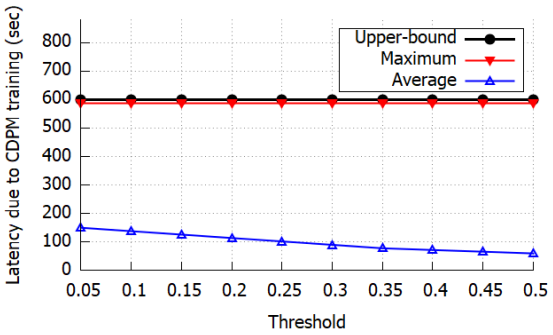

FIGURE 7. Latency caused by training in the IBRL dataset.

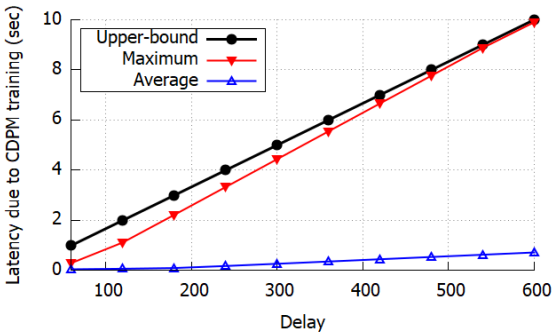

FIGURE 8. CDPM's latency is strictly within the upper bound.

RMSE score, the more accurate the predicted data will be. The *RMSE* of four algorithms for IBRL data for average temperatures and average humidity of SNs are illustrated in Table 6. The Parameter settings of CDPM will be $delay_c$ is set to be 600 seconds and d_c is set to be 20. It has been observed from Figures 5 and 6 that the RMSE of all four algorithms for temperature and humidity are low while the threshold varies (0.05 to 0.50) and thus provides high data accuracy. Although the Data Accuracy of CDPM is always higher as it has the lowest RMSE value for all thresholds. Thus, CDPM provides higher data suppression rates and energy efficiency while guaranteeing high data accuracy.

D. COMPARISON OF LATENCY

Two groups of experimentations are performed on IBRL data to determine the efficiency of CDPM in terms of latency. In the first set of experiments, ε varies from 0.05 to 0.5, and $delay_c$ is set to be 600 *seconds*. While in other sets of experiments, $delay_c$ varies from 60− 600 *seconds* and the ε is set to be 0.5. Figures 7 and 8 illustrate that the maximum delay created by CDPM's training is always inside the upper bound, while the mean value is much lower. These results indicate that if a WSN application collects data via CDPM, the delay caused by training is reasonable.

E. ALGORITHMIC COMPLEXITY OF CDPM

It is generally supposed that the greater the algorithm's complexity, the improved will be its performance. However, the algorithmic complexity of CDPM's training phase is low. The worst-case complexity when only one reading is sensed will be $O(d_b)$, which is easy to handle for SNs. After the model is trained, it is then updated and forwarded to the BS; and later the trained model will be used for data prediction. Then, every predicted value will be compared with a newly sensed value to determine AE. In case AE exceeds ϵ, *Algorithm-1* will be iteratively called to retrain the model and the latest data prediction model will be updated. The *Algorithm-2* for Data Prediction & Regeneration phase has a linear time complexity of $O(m)$. When no model adjustment is required, only one addition operation is required to predict the data value. When an adjustment is required, m number of additions are required. Hence, the proposed CDPM model has $O(d_b)$ for the model training phase and has a constant complexity of the order $O(m)$ for the data prediction and regeneration phase.

IX. CONCLUSION

This work presents a Combinational model for data prediction (CDPM) that can build prior data to control delays as well as predict future data to reduce excessive data transmission.

To eliminate unnecessary data transmission and to control delays, the proposed model is trained using an optimum current data value and then used to reconstruct previous values as well as anticipate future data. Two techniques are used to implement this paradigm in real-world WSN applications. The first technique generates step-by-step ideal models for SNs to prevent large concurrent computations and increase the model's feasibility. The other predicts and regenerates data readings of the sensed data is proposed for the BS. To evaluate the performance of our proposed CDPM data-prediction method, a WSN-based real application is simulated using a real data set. The performance of CDPM is also compared with ELR, P-PDA, and HLMS algorithms. The results demonstrated that the proposed model provides excellent proficiency in terms of reduced data suppression and data transmission, and improved energy efficiency as compared to the state-of-art algorithms. The delay caused by CDPM training is also controllable in data collection.

As future work, several improvements could be made to this work. To begin, we propose implementing the effect of transmission reduction in the real world by conducting experiments in a variety of application-based scenarios. Second, the reduction in data transmission affects bandwidth, energy consumption, latency, and data quality in WSNs. The impact of such methods determines the key performance indicators of one's interest in IoT applications. Third, the CDPM model can be used to influence other network protocols at different network layers, and thus it is critical to investigate the impact of these schemes on the various network layers.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

[1] A. Agarwal, A. Dev, and K. Jain, "Prolonging sensor network lifetime by using energy-efficient cluster-based scheduling," *Int. J. Sci. Technol. Res.*, vol. 9, no. 4, pp. 3410–3415, 2020.

[2] S. Rani, S. H. Ahmed, R. Talwar, and J. Malhotra, "Can sensors collect big data? An energy-efficient big data gathering algorithm for a WSN," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1961–1968, Aug. 2017.

[3] A. M. S. Saleh, B. M. Ali, M. F. A. Rasid, and A. Ismail, "A survey on energy awareness mechanisms in routing protocols for wireless sensor networks using optimization methods," *IEEE Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 12, pp. 1184–1207, Dec. 2014.

[4] K. Jain, A. Kumar, and C. K. Jha, "Probabilistic-based energy-efficient single-hop clustering technique for sensor networks," in *Proc. Int. Conf. Commun. Intell. Syst.* (Lecture Notes in Networks and Systems), vol. 120. Singapore: Springer, 2019, pp. 353–365.

[5] V. Pandiyaraju, R. Logambigai, S. Ganapathy, and A. Kannan, "An energy efficient routing algorithm for WSNs using intelligent fuzzy rules in precision agriculture," *Wireless Pers. Commun.*, vol. 112, no. 1, pp. 243–259, May 2020.

[6] J. Khushboo and A. Bhola, "An optimal cluster-head selection algorithm for wireless sensor networks," *WSEAS Trans. Commun.*, vol. 19, pp. 1–8, Feb. 2020.

[7] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1710–1732, 3rd Quart., 2018.

[8] K. Jain and A. Singh, "An empirical cluster head selection and data aggregation scheme for a fault-tolerant sensor network," *Int. J. Distrib. Syst. Technol.*, vol. 12, no. 3, pp. 27–47, Jul. 2021.

[9] A. Ali, Y. Zhu, and M. Zakarya, "A data aggregation based approach to exploit dynamic spatio-temporal correlations for citywide crowd flows prediction in fog computing," *Multimedia Tools Appl.*, vol. 80, no. 20, pp. 31401–31433, 2021.

[10] K. Jain and A. Kumar, "An optimal RSSI-based cluster-head selection for sensor networks," *Int. J. Adapt. Innov. Syst.*, vol. 2, no. 4, pp. 349–361, 2019.

[11] D. D. Olatinwo, A. M. Abu-Mahfouz, and G. P. Hancke, "Towards achieving efficient MAC protocols for WBAN-enabled IoT technology: A review," *EURASIP J. Wireless Commun. Netw.*, vol. 2021, no. 1, pp. 1–47, Dec. 2021.

[12] D. D. Olatinwo, A. Abu-Mahfouz, and G. P. Hancke, "A hybrid multi-class MAC protocol for IoT-enabled WBAN systems," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6761–6774, Mar. 2020.

[13] A. Agarwal, K. Jain, and A. Dev, "BFL: A buffer based linear filtration method for data aggregation in wireless sensor networks," *Int. J. Inf. Technol.*, vol. 14, pp. 1445–1454, Feb. 2022.

[14] K. Jain, A. Kumar, and V. Vyas, "A resilient steady clustering technique for sensor networks," *Int. J. Appl. Evol. Comput.*, vol. 11, no. 4, pp. 1–12, Oct. 2020.

[15] D. Fernandes, A. G. Ferreira, R. Abrishambaf, J. Mendes, and J. Cabral, "A machine learning-based dynamic link power control in wearable sensing devices," *Wireless Netw.*, vol. 27, no. 3, pp. 1835–1848, Apr. 2021.

[16] K. Jain, P. S. Mehra, A. K. Dwivedi, and A. Agarwal, "SCADA: Scalable cluster-based data aggregation technique for improving network lifetime of wireless sensor networks," *J. Supercomput.*, pp. 1–29, Mar. 2022, doi: 10.1007/s11227-022-04419-1.

[17] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco, "Practical data prediction for real-world wireless sensor networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2231–2244, Aug. 2015.

[18] J. Zhao, H. Liu, Z. Li, and W. Li, "Periodic data prediction algorithm in wireless sensor networks," in *Proc. China Conf. Wireless Sensor Netw.* (Communications in Computer and Information Science), vol. 334. Berlin, Germany: Springer, 2012, pp. 695–701.

[19] L. Tan and M. Wu, "Data reduction in wireless sensor networks: A hierarchical LMS prediction approach," *IEEE Sensors J.*, vol. 16, no. 6, pp. 1708–1715, Mar. 2015.

[20] A. Makhoul and H. Harb, "Data reduction in sensor networks: Performance evaluation in a real environment," *IEEE Embedded Syst. Lett.*, vol. 9, no. 4, pp. 101–104, Dec. 2017.

[21] 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.

[22] G. B. Tayeh, A. Makhoul, D. Laiymani, and J. Demerjian, "A distributed real-time data prediction and adaptive sensing approach for wireless sensor networks," *Pervasive Mobile Comput.*, vol. 49, pp. 62–75, Sep. 2018.

[23] H. Cheng, Z. Xie, Y. Shi, and N. Xiong, "Multi-step data prediction in wireless sensor networks based on one-dimensional CNN and bidirectional LSTM," *IEEE Access*, vol. 7, pp. 117883–117896, 2019.

[24] K. Jain and A. Kumar, "Energy-efficient data-aggregation technique for correlated spatial and temporal data in cluster-based sensor networks," *Int. J. Bus. Data Commun. Netw.*, vol. 16, no. 2, pp. 53–68, Jul. 2020.

[25] K. Jain and A. Kumar, "An energy-efficient prediction model for data aggregation in sensor network," *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 11, pp. 5205–5216, Nov. 2020.

[26] A. K. M. Al-Qurabat and A. K. Idrees, "Data gathering and aggregation with selective transmission technique to optimize the lifetime of Internet of Things networks," *Int. J. Commun. Syst.*, vol. 33, no. 11, p. e4408, Jul. 2020.

[27] K. Jain, A. Agarwal, and A. Kumar, "A novel data prediction technique based on correlation for data reduction in sensor networks," in *Proc. Int. Conf. Artif. Intell. Appl.* (Advances in Intelligent Systems and Computing), vol. 1164. Singapore: Springer, 2021, pp. 595–606.

[28] A. Agarwal, K. Jain, and A. Dev, "Modeling and analysis of data prediction technique based on linear regression model (DP-LRM) for cluster-based sensor networks," *Int. J. Ambient Comput. Intell.*, vol. 12, no. 4, pp. 98–117, Oct. 2021.

[29] H. Wang, Z. Yemeni, W. M. Ismael, A. Hawbani, and S. H. Alsamhi, "A reliable and energy efficient dual prediction data reduction approach for WSNs based on Kalman filter," *IET Commun.*, vol. 15, no. 18, pp. 2285–2299, Nov. 2021.

[30] S. N. Nels and J. A. P. Singh, "Hierarchical fractional quantized kernel least mean square filter in wireless sensor network for data aggregation," *Wireless Pers. Commun.*, vol. 120, no. 2, pp. 1171–1192, Sep. 2021.
[31] S. Famila, A. Jawahar, S. L. S. Vimalraj, and J. Lydia, "Integrated energy and trust-based semi-Markov prediction for lifetime maximization in wireless sensor networks," *Wireless Pers. Commun.*, vol. 118, no. 1, pp. 505–522, May 2021.
[32] K. Jain and A. Kumar, "A lightweight data transmission reduction method based on a dual prediction technique for sensor networks," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 11, Nov. 2021, Art. no. e4345.
[33] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 660–670, Oct. 2002.
[34] K. Fall and K. Varadhan, "The NS manual, the VINT project," UC Berkeley, USC/ISI, LBL, Xerox PARC, Tech. Rep., 2012, p. 6–9.
[35] T. Issariyakul and E. Hossain, *Introduction to Network Simulator NS2*. New York, NY, USA: Springer, 2009, pp. 1–18.
[36] S. Madden *et al.*, "Intel lab data," Intel, Santa Clara, CA, USA, Tech. Rep., 2004. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html
[37] K. Jain and A. Singh, "A two vector data-prediction model for energy-efficient data aggregation in wireless sensor network," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 11, p. e6898, May 2022. [Online]. Available: https://onlinelibrary.wiley.com/toc/15320634/2022/34/11
[38] J. Li and S. Cheng, "(ε, δ)-approximate aggregation algorithms in dynamic sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 3, pp. 385–396, Mar. 2012.

KHUSHBOO JAIN received the B.Tech. degree in information technology from MIT, Moradabad, and the M.Tech. degree in software engineering from Banasthali Vidyapith, Tonk. She is currently pursuing the Ph.D. degree in computer science and engineering. She joined the School of Computing as an Assistant Professor, in February 2021. She has published more than 30 research papers in international journals and conferences indexed in SCI and Scopus. Her research interests include wireless sensor networks, machine learning, data mining, data prediction, and software engineering. She is an editorial board member and a reviewer of many international journals.

ARUN AGARWAL received the B.Tech. degree from Uttar Pradesh Technical University, Lucknow, in 2008, and the M.Tech. and Ph.D. degrees from Guru Gobind Singh Indraprastha University, Delhi, in 2013 and 2021, respectively.

He joined as an Assistant Professor with the Department of Computer Science, Ramanujan College, University of Delhi, in 2018. His total academic experience is more than 13 years. He has published more than 30 papers in international journals and conferences. He is also the author of two book chapters. He has organized several faculty development programs and refresher courses as a Convener/a Coordinator. He has delivered key note addresses to various conferences and others in India and abroad, including Dubai. He has been a technical committee member and a reviewer in several international journals and conferences. His research interests include sensor networks, algorithms, data sciences, and software reliability modeling.

AJITH ABRAHAM (Senior Member, IEEE) received the Master of Science degree from Nanyang Technological University, Singapore, in 1998, and the Ph.D. degree in computer science from Monash University, Melbourne, Australia, in 2001. He is currently the Director of the Machine Intelligence Research Laboratories (MIR Laboratories), a Not-for-Profit Scientific Network for Innovation and Research Excellence Connecting Industry and Academia. The Network with HQ in Seattle, USA, is currently more than 1,500 scientific members from over 105 countries. As an Investigator/a Co-Investigator, he has won research grants worth over more than 100 Million U.S.$. Currently, he holds two university professorial appointments. He works as a Professor in artificial intelligence at Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair in Artificial Intelligence at UCSI, Malaysia. He works in a multi-disciplinary environment. He has authored/coauthored more than 1,400 research publications out of which there are more than 100 books covering various aspects of computer science. One of his books was translated into Japanese and a few other articles were translated into Russian and Chinese. He has more than 46,000 academic citations (H-index of more than 102 as Per Google Scholar). He has given more than 150 plenary lectures and conference tutorials (in more than 20 countries).

He was the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over more than 200 members), from 2008 to 2021, and served as a Distinguished Lecturer of IEEE Computer Society representing Europe (2011–2013). He was the Editor-in-Chief of *Engineering Applications of Artificial Intelligence* (EAAI), from 2016 to 2021, and serves/served on the editorial board for over 15 international journals indexed by Thomson ISI.