CDPM: A Combinational Data Prediction Model for Data Transmission Reduction in WSN

Khushboo Jain¹, Arun Agarwal², and Ajith Abraham³, ⁴

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

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

ABSTRACT

Background: Data prediction methods in wireless sensor networks (WSN) has been 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 involves increased data transmissions. Method: Here, we develop 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%, and 61.12%) and improved accuracy (15.38%, 21.42%, and 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 that 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 senses the environmental data at high frequencies. Continuous data transmissions cause SNs to consume a lot of energy. Because 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][8][9] will conserve the constrained resources such as unnecessary communication overheads, energy consumption etc [10][11][12][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 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 model and whose feasibility has yet to be determined and is discussed in detail in the work [17]. To summarize, data-prediction based techniques faces 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 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 application 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 follow:

- To eliminate unnecessary data transmissions by the data-prediction models which can the number of redundant data transfers and extending the overall network’s lifetime.
- To enhance 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 organised 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 of CDPM in WSN applications and Section V discuss 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 WORK

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 which uses the valuable information of the potential law contained in periodicity as a guidance to change the prediction values. P-DPA effectively improve 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 is able to predict the measured values both at the SN and at the BS. HLMS provide 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 reading from the vector by similar function. A local aggregation algorithm was further introduced to reduce the size of the dataset before transmitting it to the BS. This model minimises 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 adaptive-piecewise constant-approximation (APCA) method, Sampling rate adaptation based dynamic time warping (DTW) similarity and Frequency reduction using symbolic aggregate approximation (SAX) method. APCA removes the redundant data and adapts the sampling rate in accordance with 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 of adaptive sampling and dual prediction mechanism techniques. AS+ TR reduce energy consumption and extend 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 Spatial or Temporal corelations in data collection method as the redundant data impose unnecessary burdens on both the SN and WSN. This method Reduces the energy consumption of SNs, extends the WSN lifetime with high prediction accuracy and reasonable prediction bias. Although it considers only homogeneous WSN application and does not even control delay.

Jain et al. [24] in 2020 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 and the other at the CHs for determining spatial redundancies using AFM and RD. This method Exploits both spatial and temporal correlations, reduce data transmissions and energy consumption and improves accuracy but it considers only homogeneous WSN application and does not control delay. Jain and Kumar [25] in 2020 proposed a ECR model which is based on a two-vector model to synchronize predicted data in the intracluster transmissions to evade cumulative error in continuous data predictions. In the initialization phase of data collection cycle, it generates future data approximation and computes its prediction threshold.
error. ECR is simple, structure free and lightweight and scalable data prediction model. It reduces data transmissions cycles and energy consumption which maintaining the accuracy but has high complexity.

Al-Qurabat and Idrees [26] proposed a DGAST 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 periodic sensor networks lifetime but has complex computation and high memory usage. Jain et al. [27] in 2021 proposed a ELR model which exempt the SN from transmission 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 energy efficient model which reduces data transmission and extends network lifetime but it does not consider cluster-based topology, scalability and control’s delay.

Agarwal et al. [28] proposed a DP-LRM model 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 which compares all incoming values and establishes a correlation between them. DP-LRM is an energy efficient model which successfully reduces the data transmission cost, 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 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 which is reliable, energy efficient and extends network lifetime but has large computation overhead and it does not consider cluster-based topology and network’s 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 at 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-Markov prediction integrated with the Hyper Erlang distribution process. This work is reliable and extend the WSN lifetime but has complex computation and does not consider scalability and controls 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 few data values, which increases data transmission effectiveness and reduces the energy consumption. DTRM provides data accuracy, reduced data transmissions, low complexity costs, lightweight processing, limited memory footprint, robust and effective 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.

### TABLE I

| Techniques                          | References | Issue Addressed                        | Contributions                                      | Limitations                                      |
|-------------------------------------|------------|----------------------------------------|----------------------------------------------------|--------------------------------------------------|
| Periodic Data Prediction Algorithm (P-PDA) | [18]       | Energy consumption                      | 1. Improve Accuracy                                | How to find attribute correlation is not discussed.  |
|                                      |            |                                        | 2. Reduces communication frequency                 | Does not control delay                           |
|                                      |            |                                        | 3. Prolongs WSN lifetime.                          |                                                  |
| Hierarchical Least Mean Square (HLMS) | [19]       | To schedule the data communications between SNs and the BS.  | 1. Energy conservation                              | Only temporal data prediction is not spatially predicted. |
|                                      |            |                                        | 2. Data transmission reduction                     | The synchronization of the correlations filters at SN and BS is unexplored. |
|                                      |            |                                        | 3. Data Accuracy                                   | Does not control latency.                        |
| Data reduction model based on Kruskal-Wallis’s (KW) test | [20]       | Minimizing the data size for transmission over the WSN for energy conservation. |                                                     | Do not apply correlation between the adjacent SNs |
|                                      |            |                                        |                                                    |                                                  |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
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 in accordance with the environment 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
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
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 Efficient Extends network lifetime
Complex computation and high memory usage
Does not control delay

Extended Linear Regression (ELR) model [27]
Data gathering overheads
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
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
High data reliability Efficient and Effective data reduction Extends network lifetime
Does not support scalability and control delay
Large computation overhead
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 technique discussed above, the strength of the proposed CDPM algorithm lies in its ability to control delay, reduce energy consumption by achieving high data transmission suppression and reduced RMSE (improved data quality).

### III. SYSTEM MODEL AND ASSUMPTION

#### A. NETWORK MODEL

The proposed network model of CDPM is demonstrated in Figure 1. The WSN consists of \( N \) number of sensor nodes, \( S = \{ S_1, S_2, S_3, \ldots, S_N \} \) 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 \( t = \{ t_1, t_2, t_3, \ldots, t_j \} \). The Combinational Data-Prediction Model suggest that the correlatd 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 in 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 in a timely manner. 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 CDPM model:

- The SN 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 as device-to-device, which means, that the data transfers between SN and BS reach without any delay.
- Dissimilar to SNs, the BS have 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 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 a SN can be denoted by the time-series reading as follow in Equation (1) below:

\[
f(t_i) = SR_{t_i}, \quad i \in \{1, 2, 3, \ldots, N\}
\]

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

The SN will send \( (t_i, SR_{t_i}) \) 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:

\[ SR = f(t) = \begin{cases} m_1t + c_1 & \text{step}_1 \leq t < \text{step}_2 \\ m_2t + c_2 & \text{step}_2 \leq t < \text{step}_3 \\ \text{....} \end{cases} \]  

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 \( t = \text{step}_j \). 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 \( t = \text{step}_j \). 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 CDPM model: The CDPM algorithm updates data in every step which has two stages: first stage is the training phase and the second stage is the data prediction phase. During 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 reconstruct sensor data of the first phase. If the prediction threshold is more than the predefined error, CDPM model will be retrained, i.e., the next step begins. The \( d_j \) represents the training data values in \( j^{th} \) 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 \((\text{step}_j, m_j, c_j)\). The CDPM model can be used to rebuild at least two data points, 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 \( j^{th} \) step is directly related to the number of data \((d_j)\) in training data

\[ \text{delay}_j = (d_j - 1) \Delta t \]  

Here \( \text{delay}_j \) represents the maximum-delay in \( j^{th} \) step and \( \Delta t \) 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 \( \text{delay}_c \) and the maximum values for training phase is delimit as \( d_c \). It is predefined by the following Equation (4) below:

\[ d_c = \begin{cases} \text{if } (\text{delay}_c \geq \Delta t) , d_c = \left\lceil \frac{\text{delay}_c}{\Delta t} \right\rceil + 1 \\ \text{else } , d_c = 2 \end{cases} \]  

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 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 in order to create the best precise linear prediction model in training phase. We have calculated Absolute Error \((AE)\), which is the difference between the sensed reading and the predicted data as follow in Equation (5) below:

\[ AE_i = SR_i - SR_i \]  

Then we calculate the error in training phase, we have evaluated the residual Sum of Squares \((RSS)\) as follow in Equation (6) below:

\[ RSS = \Sigma(AE_i)^2 \]  

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

\[ m = \frac{\text{d}_j \Sigma (t_i \times SR_i) - \Sigma (t_i) \times \Sigma (SR_i)}{\text{d}_j \Sigma (t_i^2) - \Sigma (t_i)^2} \]  

\[ c = \frac{\Sigma (t_i^2) \Sigma (SR_i) - \Sigma (t_i) \times \Sigma (t_i \times SR_i)}{\text{d}_j \Sigma (t_i^2) - \Sigma (t_i)^2} \]  

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

\[ (m, c) = \{(t_1, SR_1), (t_2, SR_2), \ldots \} \]  

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

\[ RMSE = \sqrt{\frac{1}{N} \Sigma_{i=1}^{N} (AE_i)^2} \]
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 restructured and updated. Since we have considered ε to be upper-bound of RMSE, the prediction error will always within the threshold. The threshold value of \( RMSE_j \) is calculated conferring the values of \( \epsilon \), where \( RMSE_j \) represents the RMSE of \( j^{th} \) training phase.

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

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

Thus, according 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 Least square method is used only whenever a new data is sensed, in order to avoid a huge amount of concurrent computations. The Algorithm-1 states that: If the value of RSS is greater than value of \( RSS_e \), then it will return an earlier outcome. Whenever the value of \( d_j \) matches \( d_e \), the Algorithm-1 will return an up-to-date outcome.

**Algorithm 1:** CDPM Training Phase

| Parameters: \( \epsilon, t, d, (SR_t, SR_{t+1}, SR_{t+2}, \ldots) \) |

1. **Algorithm BEGIN**
2. \( d = 1; \)
3. \( RSS_e = 2\epsilon^2; \)
4. **while** new sensor’s reading is generated **do**
5. \( d += 1; \)
6. \( (m', c') = lsm((t, SR_t), \ldots, (t+d-1, SR_{t+d-1}); \)
7. \( RSS = \sum_{i=0}^{d-1} (SR_i - i \times m' - c')^2; \)
8. **if** \( (RSS > RSS_e) \)
9. \( return (t, m, c); \)
10. **else**
11. \( m = m'; c = c'; RSS_e += \epsilon^2; \)
12. **endif**
13. **if** \( d == d_e \)
14. \( return (t, m, c); \)
15. **endif**
16. **endwhile**
17. **ALGORITHM END**

The algorithmic complexity of CDPM training phase is low. The worst-case complexity when only one reading is sensed will be \( O(d) \), which is easy enough for SNs. After 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 newly sensed value to determine AE. In case \( AE \) exceeds \( \epsilon \), 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: \( (t, m, c), \Delta t \) |

| Parameters: \( \{SR_1, SR_2, SR_3, \ldots\} \) |

1. **ALGORITHM BEGIN**
2. **while** new-model is obtained **do**
3. \( SR_i = mt + c; \)
4. **for** \( (i = t + 1; i \leq present - epoch; i +) \)
5. \( SR_i += m; \)
6. **endfor**
7. **while** no model is obtained **do**
8. \( SR_{i+} += m; \)
9. \( sleep(\Delta t); \)
10. **endif**
11. **endwhile**
12. **ALGORITHM END**

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

**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 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 a 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.

$$E_{SEN} = bV I_S T_S \quad (12)$$

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 (f sp) and multipath (mph). A threshold selects the channel, which is $d^2$ energy loss for small distance and $d^4$ energy loss for 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).

$$E_{TX}(b,d) = \begin{cases} \frac{bE_{elec}}{d^4}, & \text{if } d < d_0 \\ \frac{bE_{elec}}{d^2}, & \text{if } d \geq d_0 \end{cases} \quad (13)$$

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

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

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

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

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

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.

$$E_{CDPM} = b V I_l \frac{N_c}{f} \frac{V}{n_p \times V_t} + b C_{avg} N_c V^2 \quad (16)$$

### 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.

#### 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.

| 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 | 31 s |

#### B. DATASET

The Intel Berkeley Research Laboratory (IBRL) [36] has approximately 2.3 million sensor measurements. Each SN 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 is 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 IBRL dataset and following metrics are defined for them. Moreover, according to [27][32], data transmission is the major issue for energy depletion of such network. Therefore, the 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.

\[
\text{TS\%} = \left( \frac{\text{Transmitted data by using prediction algorithm}}{\text{Actual sensed data}} \right) \times 100
\]  

(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 TS\% will be, the less data will be transmitted and the less energy will be consumed.

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.

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).

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (AE_i)^2}
\]  

(18)

Where \( AE_i = SR_i - \hat{SR}_i \), \( SR_i \) is the senses reading of \( i^{th} \) SN and \( \hat{SR}_i \) 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 communication. Each round has a varying threshold \( \varepsilon \) 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 less 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 visualised in Figure 1 and 2 respectively and are illustrated in Table 4. The Parameter settings of CDPM will be \( \text{delay}_c \) is set to be 600 seconds. Since \( \Delta t \) of the IBRL dataset is 31 seconds and \( d_c \) is set to be 20. The TS\% of CDPM is always less 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 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.

| S. No | Rounds | Threshold | Average temperatures of SNs | Average Humidity of SNs |
|-------|--------|-----------|-----------------------------|-------------------------|
|       |        |           | CDPM | HLMS | ELR | P-PDA | CDPM | HLMS | ELR | P-PDA |
| 1     | \( r_{o_1} \) | 0.05 | 48.94 | 39.59 | 26.68 | 13.88 | 54.39 | 46.59 | 31.37 | 22.99 |
| 2     | \( r_{o_2} \) | 0.10 | 65.35 | 47.48 | 34.35 | 25.68 | 67.66 | 53.78 | 38.16 | 33.55 |
| 3     | \( r_{o_3} \) | 0.15 | 71.46 | 53.25 | 45.22 | 38.35 | 71.14 | 59.47 | 43.75 | 38.89 |
| 4     | \( r_{o_4} \) | 0.20 | 76.77 | 58.39 | 54.55 | 49.66 | 72.66 | 60.69 | 58.18 | 45.77 |
| 5     | \( r_{o_5} \) | 0.25 | 78.46 | 61.35 | 59.02 | 54.45 | 73.33 | 61.67 | 57.36 | 46.66 |
| 6     | \( r_{o_6} \) | 0.30 | 79.13 | 63.13 | 61.75 | 57.35 | 74.61 | 62.36 | 56.65 | 47.44 |
| 7     | \( r_{o_7} \) | 0.35 | 80.11 | 65.02 | 63.24 | 60.55 | 75.59 | 63.85 | 55.21 | 48.88 |
| 8     | \( r_{o_8} \) | 0.40 | 81.02 | 67.91 | 65.51 | 63.33 | 76.25 | 64.58 | 54.77 | 49.69 |
| 9     | \( r_{o_9} \) | 0.45 | 82.84 | 68.11 | 66.75 | 65.88 | 77.02 | 65.76 | 53.98 | 50.57 |
| 10    | \( r_{o_{10}} \) | 0.50 | 83.68 | 69.88 | 67.35 | 66.46 | 77.85 | 66.17 | 52.02 | 49.15 |
B. COMPARISON OF ENERGY CONSUMPTION

Most of 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 $\varepsilon$ 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 $d_c$ is set to be 10 minutes. Since $\Delta t$ 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 be 0.0144 mJ and for receiving one byte data is 0.0057 mJ [38]. The cumulative energy consumption of algorithms after every round of communication for temperature and humidity are presented in Figure 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.

![Figure 1: Transmission Suppression for Temperature](image1)

![Figure 2: Transmission Suppression for Humidity](image2)

**TABLE V**

| S. No | Rounds | Threshold | Average temperatures of SNs |        | Average Humidity of SNs |        |
|-------|--------|-----------|----------------------------|--------|-------------------------|--------|
|       |        |           | CDPM | HLMS | ELR | P-PDA | CDPM | HLMS | ELR | P-PDA |
| 1     | $r_{o_1}$ | 0.05 | 12.23 | 16.45 | 25.65 | 40.87 | 5.34 | 8.85 | 10.71 | 15.14 |
| 2     | $r_{o_2}$ | 0.10 | 10.32 | 14.73 | 23.48 | 36.57 | 4.87 | 8.25 | 10.12 | 14.35 |
| 3     | $r_{o_3}$ | 0.15 | 9.65 | 13.16 | 21.65 | 34.83 | 4.42 | 7.92 | 9.65 | 13.73 |
| 4     | $r_{o_4}$ | 0.20 | 8.41 | 12.43 | 19.34 | 30.27 | 4.12 | 7.56 | 9.11 | 12.84 |
| 5     | $r_{o_5}$ | 0.25 | 7.93 | 11.65 | 17.75 | 27.93 | 3.98 | 6.98 | 8.79 | 11.53 |
| 6     | $r_{o_6}$ | 0.30 | 7.38 | 10.54 | 15.87 | 24.47 | 3.58 | 6.32 | 8.21 | 10.55 |
| 7     | $r_{o_7}$ | 0.35 | 6.82 | 9.66 | 13.23 | 21.84 | 3.27 | 5.87 | 7.72 | 9.81 |
| 8     | $r_{o_8}$ | 0.40 | 6.25 | 8.76 | 12.54 | 18.28 | 2.98 | 5.12 | 7.19 | 8.35 |
| 9     | $r_{o_9}$ | 0.45 | 5.79 | 7.98 | 11.45 | 15.85 | 2.63 | 4.76 | 6.53 | 7.65 |
| 10    | $r_{o_{10}}$ | 0.50 | 5.17 | 7.34 | 10.37 | 13.36 | 2.23 | 4.08 | 5.75 | 6.85 |
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 $\varepsilon$ from 0.0 to 0.5 with a step function of 0.5. The lower the 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 $d_{\text{delay}}$ is set to be 600 seconds and $d_c$ is set to be 20. It has observed from Figure 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 the 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.

| S. No | Rounds | Threshold | Average temperatures of SNs | Average Humidity of SNs |
|-------|--------|-----------|-----------------------------|-------------------------|
|       |        |           | CDPM | HLMS | ELR | P-PDA | CDPM | HLMS | ELR | P-PDA |
| 1     | $r_{o_1}$ | 0.05 | 0.014 | 0.027 | 0.040 | 0.048 | 0.023 | 0.031 | 0.047 | 0.054 |
| 2     | $r_{o_2}$ | 0.1 | 0.023 | 0.034 | 0.047 | 0.062 | 0.033 | 0.038 | 0.054 | 0.068 |
| 3     | $r_{o_3}$ | 0.15 | 0.033 | 0.043 | 0.053 | 0.071 | 0.038 | 0.043 | 0.059 | 0.071 |
| 4     | $r_{o_4}$ | 0.2 | 0.041 | 0.050 | 0.058 | 0.074 | 0.042 | 0.047 | 0.060 | 0.072 |
| 5     | $r_{o_5}$ | 0.25 | 0.048 | 0.054 | 0.061 | 0.075 | 0.043 | 0.048 | 0.061 | 0.073 |
| 6     | $r_{o_6}$ | 0.3 | 0.051 | 0.057 | 0.063 | 0.076 | 0.044 | 0.049 | 0.062 | 0.074 |
| 7     | $r_{o_7}$ | 0.35 | 0.052 | 0.061 | 0.065 | 0.077 | 0.045 | 0.050 | 0.063 | 0.075 |
| 8     | $r_{o_8}$ | 0.4 | 0.053 | 0.063 | 0.068 | 0.078 | 0.046 | 0.051 | 0.064 | 0.076 |
| 9     | $r_{o_9}$ | 0.45 | 0.054 | 0.064 | 0.068 | 0.079 | 0.047 | 0.052 | 0.065 | 0.077 |
| 10    | $r_{o_{10}}$ | 0.5 | 0.055 | 0.065 | 0.070 | 0.080 | 0.048 | 0.054 | 0.066 | 0.078 |
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 the 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 the high.

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, \( \epsilon \) varies from 0.05 to 0.5 and \( \text{delay}_c \) is set to be 600 seconds. While in other set of experiments, \( \text{delay}_c \) varies from 60–600 seconds and the \( \epsilon \) is set to be 0.5. Figure 7 and 8 illustrates that the maximum delay created by CDPM’s training which 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.

FIGURE 7. Latency caused by training in the IBRL dataset

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 \( O(d_p) \), which is easy to handle for SNs. After model is trained, it is then updated and forwarded to the BS; and later the trained model will used for data-prediction. Then, every predicted value will be compared with newly sensed value to determine \( AE \). In case \( AE \) exceeds \( \epsilon \), 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 \( O(d_p) \) for model training phase and has a constant complexity of the order \( O(m) \) for 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 in order 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 proposed model provides excellent proficiency in terms of reduced data suppression and data transmission, improved energy efficiency as compared to the state-of-art algorithms. The delay caused by CDPM training is also controllable in data collection.
As a 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 for 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.

ACKNOWLEDGMENT

This research has been financially supported by The Analytical Center for the Government of the Russian Federation (Agreement No. 70-2021-00143 dd. 01.11.2021, IGK0000000D730321PSQ0002).

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 Transactions on Data Industrial Informatics, vol. 13, no. 4, pp. 1961–1968, 2017.

[3] A. M. Shamsan Salah, 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,” Transactions on Emerging Telecommunications Technologies, vol. 25, no. 12, pp. 1184–1207, 2014.

[4] K. Jain, A. Kumar, and C. K. Jha, “Probabilistic-based energy-efficient single-hop clustering technique for sensor networks,” in International Conference on Communication and Intelligent Systems. 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 Personal Communications, vol. 112, no. 1, pp. 243–259, 2020.

[6] K. Jain and A. Bholia, “An optimal cluster-head selection algorithm for wireless sensor networks,” WSEAS Transactions on Communications, ISSN/E-ISSN, pp. 1109–2742.

[7] L. Li, G. Zhao, and R. S. Blum, “A survey of caching techniques in cellular networks: Research issues and challenging content placement and delivery strategies,” IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 1710–1732, 2018.

[8] K. Jain and A. Singh, “An empirical cluster head selection and data aggregation scheme for a fault-tolerant sensor network,” International Journal of Distributed Systems and Technologies (IJDST), vol. 12, no. 3, pp. 27–47, 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 and Applications, vol. 80, no. 20, pp. 31 401–31 433,2021.

[10] K. Jain and A. Kumar, “An optimal rssi-based cluster-head selection for sensor networks,” International Journal of Adaptive & Innovative Systems, 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 was-enabled iot technology: a review,” EURASIP Journal on Wireless Communications and Networking, vol. 2021, no. 1, pp. 1–47, 2021.

[12] D. D. Olatinwo, A. M. Abu-Mahfouz, and G. P. Hancke, “A hybrid multi-class mac protocol for iot-enabled was systems,” IEEE Sensors Journal, vol. 21, no. 5, pp. 6671–6674, 2020.

[13] Agarwal, K. Jain, and A. Dev, “Bft: a buffer based linear filtration method for data aggregation in wireless sensor networks,” International Journal of Information Technology, pp. 1–10, 2022.

[14] K. Jain, A. Kumar, and V. Vyas, “A resilient steady clustering technique for sensor networks,” International Journal of Applied Evolutionary Computing (IJAEC), vol. 11, no. 4, pp. 1–12, 2020.

[15] D. Fernandes, A. G. Ferreira, R. Abrishambah, J. Mendes, and J. Cabral, “A machine learning-based dynamic link power control in wearable sensing devices,” Wireless Networks, vol. 27, no. 3, pp. 1833–1848, 2021.

[16] K. Jain, PS. Mehra, AK Dwivedi, and A. Agarwal, SCADA: scalable cluster-based data aggregation technique for improving network lifetime of wireless sensor networks. The Journal of Supercomputing, 1–29, 2022.

[17] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco, “Practical data prediction for real-world wireless sensor networks,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, pp. 2231–2244, 2015.

[18] J. Zhao, H. Liu, Z. Li, and W. Li, “Periodic data prediction algorithm in wireless sensor networks,” in China Conference on Wireless Sensor Networks. Springer, 2012, pp. 695–701.

[19] L. Tan and M. Wu, “Data reduction in wireless sensor networks: A hierarchical lms prediction approach,” IEEE Sensors Journal, vol. 16, no. 6, pp. 1708–1715, 2015.

[20] A. Makhoul and H. Harb, “Data reduction in sensor networks: Performance evaluation in a real environment,” IEEE Embedded Systems Letters, vol. 9, no. 4, pp. 101–104, 2017.

[21] A. K. M. Al-Qurabat and A. K. Idrees, “Energy-efficient adaptive distributed data collection method for periodic sensor networks,” International Journal of Internet Technology and Secured Transactions, vol. 8, no. 3, pp. 297–335, 2018.

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

[23] H. Cheng, Z. Xie, Y. Shi, and N. Xiong, “Multi-step data prediction in wireless sensor networks based on one-dimensional cma and bidirectional lstm,” IEEE Access, vol. 7, pp. 117 883–117 896, 2019.

[24] K. Jain and A. Kumar, “Energy-efficient data-aggregation technique for correlated spatial and temporal data in cluster-based sensor networks,” International Journal of Business Data Communications and Networking (IJBDCN), vol. 16, no. 2, pp. 53–68, 2020.

[25] K. Jain and A. Kumar, “An energy-efficiency prediction model for data aggregation in sensor network,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 11, pp. 5205–5216, 2020.

[26] A. K. M. Al-Qurabat and A. Kadhum Idrees, “Data gathering and aggregation with selective transmission technique to optimize the lifetime of internet of things networks,” International Journal of Communication Systems, vol. 33, no. 11, p. e4408, 2020.

[27] K. Jain, A. Agarwal, and A. Kumar, “A novel data prediction technique based on correlation for data reduction in sensor networks,” in Proceedings of international conference on artificial intelligence and applications. 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,”
[29] H. Wang, Z. Yemeni, W. M. Ismael, A. Hawbani, and S. H. Alsamhi, “A reliable and energy efficient dual prediction data reduction approach to wsns based on kalman filter,” *IET Communications*, 2021.

[30] S. Ninisha Nels and J. Amar Pratap Singh, “Hierarchical fractional quantized kernel least mean square filter in wireless sensor network for data aggregation,” *Wireless Personal Communications*, vol. 120, no. 2, pp. 1171–1192, 2021.

[31] S. Famila, A. Jawahar, S. Vimalraj, and J. Lydia, “Integrated energy and trust-based semi-markov prediction for lifetime maximization in wireless sensor networks,” *Wireless Personal Communications*, vol. 118, no. 1, pp. 505–522, 2021.

[32] K. Jain and A. Kumar, “A lightweight data transmission reduction method based on a dual prediction technique for sensor networks,” *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 11, p. e4345, 2021.

[33] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, “An application-specific protocol architecture for wireless microsensor networks,” *IEEE Transactions on wireless communications*, vol. 1, no. 4, pp. 660–670, 2002.

[34] K. Fall and K. Varadhan, “The ns manual, the vint project, (2008),” *Ng y nh'b i*, pp. 06–9, 2012.

[35] T. Issariyakul and E. Hossain, “Introduction to network simulator 2 (ns2),” in *Introduction to network simulator NS2*. Springer, 2009, pp. 1–18.

[36] S. Madden et al., “Intel lab data,” *Web page, Intel*, 2004.

[37] K. Jain and A. Singh, “A Two Vector Data-Prediction Model for Energy-Efficient Data Aggregation in Wireless Sensor Network,” *Concurrency and Computation: Practice and Experience*, 2022.

[38] J. Li and S. Cheng, “(ε, δ)-approximate aggregation algorithms in dynamic sensor networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 3, pp. 385–396, 2011.
KHUSHBOO JAIN joined School of Computing as an Assistant Professor in February 2021. She is pursuing PhD in Computer Science & Engineering and completed her M. Tech in Software Engineering from Banasthali Vidyapith, Tonk and B. Tech in Information Technology from MIT, Moradabad. Her research interests include Wireless Sensor Networks, Machine Learning, Data Mining, Data Prediction & Software Engineering. She is Editorial Board Member and Reviewer of many International Journals. She has published has more than 30 research papers in international journals and conferences indexed in SCI and Scopus. Before joining DITU, she was an Assistant Professor in the Noida Institute of Engineering & Technology, Greater Noida, India.

ARUN AGARWAL joined as Assistant Professor, Department of Computer Science at Ramanujan College, University of Delhi in 2018. He completed his PhD Degree from Guru Gobind Singh Indraprastha University, Delhi in 2021. He did M. Tech. from Guru Gobind Singh Indraprastha University, Delhi in 2013 and B. Tech from Uttar Pradesh Technical University, Lucknow in 2008. His total academic experience is more than 13 years. His areas of interests include Sensor Networks, Algorithms, Data Sciences and Software Reliability Modeling. 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 Course as Convener/Coordinator. He has delivered Key Note Addresses in various Conferences and others in India and abroad including Dubai. He has been a technical committee member and reviewer in several international journal and conferences.

AJITH ABRAHAM works in a multi-disciplinary environment and he has authored / coauthored more than 1,400+ research publications out of which there are 100+ books covering various aspects of Computer Science. One of his books was translated to Japanese and few other articles were translated to Russian and Chinese. Dr. Abraham has more than 46,000+ academic citations (h-index of 100+ as per google scholar). He has given more than 150 plenary lectures and conference tutorials (in 20+ countries). Since 2008, Dr. Abraham was the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over 200+ members) during 2008-2021 and served as a Distinguished Lecturer of IEEE Computer Society representing Europe (2011-2013). Currently Dr. Abraham was the editor-in-chief of Engineering Applications of Artificial Intelligence (EAAI) during 2016-2021 and serves/served the editorial board of over 15 International Journals indexed by Thomson ISI. Dr. Abraham received Ph.D. degree in Computer Science from Monash University, Melbourne, Australia (2001) and a Master of Science Degree from Nanyang Technological University, Singapore (1998).