Wireless Personal Communications manuscript No.
(will be inserted by the editor)

A Multi-Tier Data Prediction Mechanism for the Internet of Things Networks

Hassan Harb* · **Chady Abou Jaoude** · **David Laiymani** · **Abdallah Makhoul** · **Chamseddine Zaki** · **Layla Tannoury**
* *Corresponding author*

Received: date / Accepted: date

Abstract In this era, we need to make everything about us more smartly and communicating. Hence, the popularity of Internet of Thing (IoT) is increasing quickly across industries. In such networks, the sensors represent the eyes of the IoT that collect data about different environments and states, while the sink node forms the brain of the network that must analyze the collected data and take decisions. However, the big amount of data collected by the sensors leads, from one hand, to consume the limited energy of the sensor and, from another hand, to complicate the exploitation of the data at the sink for decision making. In this paper, we propose a multi-tier prediction mechanism in order to handle big data collected by sensor networks based on the clustering scheme. The prediction model uses the least squares approximation method which is applied at both tiers of each cluster: sensors and cluster-heads (CHs). At the first tier, each sensor applies the prediction model in order to send a reduced set of data to its appropriate CH; At the second tier, the CH combines data coming

H. HARB
Faculty of Sciences, Lebanese University, Beirut, Lebanon
E-mail: hassan.harb.1@ul.edu.lb

C. ABOU JAOUDE
TICKET lab, Faculty of Engineering, Antonine University, Baabda, Lebanon
E-mail: chady.aboujaoude@ua.edu.lb

D. LAIYMANI, A. MAKHOUL
FEMTO-ST Institute/CNRS, the DISC department, Univ. Bourgogne Franche-Comté, Belfort, France
E-mail: firstname.lastname@univ-fcomte.fr

C. ZAKI
College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait
E-mail: chamseddine.zaki@aum.ku

L. TANNOURY
Faculty of Business Administration and Economics, Lebanese University, Rashaya, Lebanon
E-mail: layla.tannoury@ul.edu.lb

from sensors and provides a predictive model to the sink, representing data for all sensors of the cluster. Extensive simulations on real sensor data collected from several applications demonstrated that our mechanism can efficiently reduce the data transmission and save the network energy, while maintaining an acceptable data accuracy level.

Keywords IoT · WSN · Data prediction · Least squares approximation · Cluster topology · Real sensor data

1 Introduction

The beginning of this decade have emerged an increasing number of connected devices accompanied with a vast improvements in communication technologies. This leads to the emergence of new sensing-based network applications called the Internet of Things (IoT). Such new technology founds its way quickly across industries affecting business and people lives [1]. From inside our homes to across nature, IoT represents a low-cost solution to sense surroundings and people behaviors with the opportunity to directly store data at the cloud. According to International Data Corporation (IDC) [2], it is estimated that more than 30 billion IoT devices will be deployed by 2020 with a technology spending about $1.2T. Applications dedicated to healthcare, security, retail, climate, etc. will be the main targets of IoT investments.

Generally, the IoT networks consist of a set of components which can be connected together to monitor desired zones. Wireless sensor network (WSN) constitutes the main component of the IoT networks; the sensors represent the eyes of the IoT that can be deployed at any place to monitor any environment. Whilst, the sink node looks like the brain that allows IoT to collect and analyze the data in order to take the right decision. However, data management is not an easy task in sensor networks and it represents a major challenge for decision makers as the amount of collected data is huge [3,4]. Furthermore, the sensors have a limited power energies, which is mostly not rechargeable, where the data transmission consumes lots of such available energy [5,6]. Hence, data reduction and prediction mechanisms are at the heart of research focuses as an efficient way to reduce data transmission and helping in taking decisions.

In this paper, we propose a multi-tier prediction mechanism dedicated to periodic large-scale sensor network applications. First, we consider a cluster-based network topology then we introduce a new prediction model that can be applied at both sensor and cluster-head (CH) nodes. Subsequently, the contribution of this study is described as follows:

- At the first tier, each sensor uses the least squares approximation method in order to send a predictive set to the CH instead of sending the whole periodic raw data.
- At the second tier, each CH at the second tier uses the same prediction method in order to provide a unique predictive set to the sink, representing the data of all sensors in that cluster.

– Our mechanism is evaluated upon a serie of simulations and based on real sensor data for various kinds of WSN applications. The results show the efficiency of our mechanism in terms of energy consumption, data latency and data accuracy.

The rest of paper is organized as follows. In Section 2, we briefly present literature on data prediction and reduction techniques. Section 3 presents the network design that is used in this paper. Section 4 and Section 5 detail the prediction model proposed at the sensors and CH levels respectively. The description of datasets used in our simulation are described in Section 6. Section 7 presents the simulation results with necessary explanation. Section 8 concludes this paper with some perspectives.

2 Related Work

Currently, researchers are focusing on data prediction approach as an efficient way to handle big data produced from WSN and save network energy [7–9]. The idea behind such approach is to build, based on the collected data, a predictive model in order to send to the sink which, in its turn, regenerates the raw data. Researchers on [10] have presented a review article about various data prediction mechanisms proposed at the literature for WSN, while comparing the difference between them.

The authors of [11] propose a similarity life prediction model of rolling bearing that works on two steps. First, a set of degradation features is extracted from the bearing vibration signals followed by a fusion mechanism to maintain the potential features based on the principle component analysis. Then, the life adjustment functions are constructed by calculating the comprehensive similarity while the life prediction of the monitoring bearings is corrected in real-time according to the PCA features. In [12], the authors propose a prediction model, called MooCare, that helps farmers in increasing their productivities of their daily cattle. After monitoring the animal feeding through IoT devices, MooCare uses the ARIMA prediction to forecast the milk production of each cow and, thus, allows farmers to design a suitable nutritional plan for each one. In [13], the authors propose a diagnostic prediction method for chronic kidney disease (CKD) that uses IoT networks. The proposed method aims to select the potential features from the huge amount of data collect about CKD then to predict the severity level of disease via several classification techniques such as random forest and logistic regression. The authors of [14] introduce a remaining useful life method for predicting data in sensor networks based on a data fusion model. According to the proposed method, the variation of the observed condition is expressed through a state transition function accompanied with Wiener process. Then, another function that uses multi-sensor signals has been adopted to inherent the system degradation followed by a selection algorithm to choose the list of active sensors while predicting the values of the sleep ones. In [15], the authors propose a

data prediction mechanism that investigates the data correlation among sensors with the objective to avoid transmitting unuseful information. Through a mathematical model, the prediction mechanism allows to study the variation between the readings collected by the sensors and eliminates the correlated ones in order to save the node energy. In [16], a similar data prediction technique that takes benefit of relationships among sensor data is proposed. The proposed technique introduces an enhanced version of linear regression that identifies the shape similarities in data curve in periodical data collection.

The authors in [17] propose a unsupervised machine learning algorithm, called kohonen, for predicting data generated by the sensors. Kohonen introduces a self organizing map based on a predictive temporal model that makes sensor in standby mode to reduce its transmission. In [18], the authors propose a mechanism that predicts future values based on the past one. The mechanism uses an autoregressive model of order p and allows to study the variation in sensed data along with the network lifetime. In [19], a derivative-based prediction (DBP) technique is proposed. DBP is dedicated to WSN applications requiring high data accuracy and it predicts the variation of data collected by a sensor node. The authors in [20] uses time series in order to predict temperature readings in WSN. First, the proposed model adapts the sensor sampling rate based on the variation on accuracy of data collected. Then, a stochastic process is used to analyze the temperature phenomena using time series model. In [21], an online data tracking and estimation (ODTE) is proposed in order to tracking poor data collected at the sink. ODTE is mainly based on two systems: Data prediction system (DPS) and distortion factor (DF). DPS is used at the sensor in order to reduce its transmission using a defined limit while DF estimates an optimal data collected at sink node. In [22], the authors introduce an efficient method to predict the occupancy rate in the smart buildings then to control the indoor conditions (air quality, heating, and ventilation). The proposed method is based on an artificial neural network trained on an indoor data and allows to trigger the ventilation rate control through an IoT communication protocol.

Finally, some data prediction techniques on sensor networks are based on aggregation and compression approaches. The idea behind such approaches is to reduce the amount of data transmitted from source nodes while regenerating the aggregated/compressed data at the sink. In [23], the authors propose a data aggregation mechanism that works at two levels: sensor and CH. At the first level, a similarity function that searches, then eliminates, the redundancies among raw data collected periodically by each sensor. At the second level, an in-network reduction mechanism, called prefix frequency filtering (PFF), is proposed. PFF allows the CH to remove the redundancies existing among data collected by neighboring sensor nodes, before sending them to the sink. The authors of [24] proposes a prediction technique based on a coding provenance scheme (CBP). Typically, CBP is characterized by its high data reduction rate due to its encoding and decoding operations that depends on the monitored condition. Lastly, an efficient prediction method based on a compression scheme, i.e. Sequential Lossless Entropy Compression (S-LEC), is proposed.

The idea behind S-LEC is to order the alphabet of integers into groups where each group is represented by two codes, i.e. entropy and binary; the entropy code indicates the group number while the binary code indicates the offset in the group. In [25], the authors propose a propose a multidimensional and multidirectional data aggregation (MMDA) technique in order to enhance the data communication and ensure the privacy of the data. MMDA allows each sensor device to organize the data into matrices then applying an aggregation process in two directions, e.g. rows and columns. The authors of [26] propose an entropy-driven data aggregation with a gradient distribution (EDAGD) technique that is relying on three algorithms. The first algorithm is called a multi-hop tree-based data aggregation and aims to reduce the transmission distance among sensors and the sink by minimizing the number of hops required to reach the destination. The second algorithm is a tree-based aggregation scheme that uses the entropy and the Choquet integral that allows to monitor and detect abnormal events based on the sleep/active nodes strategy. The last aggregation method is a gradient deployment algorithm which aims to deal with the energy hole problem in sensor applications.

Unfortunately, most of the proposed prediction techniques in sensor networks make a trade-off between data transmission ratio, latency, accuracy and energy consumption. From one hand, minimizing the data transmission often leads to save the sensor energies but it decreases the accuracy of the transmitted data. On the other hand, preserving the integrity of data may require complex techniques that mostly affects the latency of the transmitted information. In this paper, we propose a novel data reduction mechanism based on the prediction approach that ensures a trade-off among different metrics evaluated in sensor applications.

3 Cluster-Based Network Architecture

Network topology is one of the most key features that should be consider when deploying a sensor network. Although there are many topologies proposed in WSN [27], researchers are mainly focused on two architectures: clustering and tree. Indeed, tree-based WSN is more suitable for applications requiring a small size of sensors otherwise, e.g. number of sensors gets bigger, construction of the tree will be very complex. Such reconfiguration of the tree mostly requires high time processing and network energy consumed especially when a node is failed or its energy is depleted (particularly for those near to the sink). Hence, for less-complexity reason, most of the proposed techniques are dedicated to cluster-based topology in order to maintain the scalability of the network and save its energy. Subsequently, the authors in [27] study the various topologies of WSNs (tree, cluster, chain and flat) while comparing them according to many performance metrics like energy usage, network lifetime, scalability, latency, etc.

In this paper, we focus on cluster-based topology that has been widely used in IoT applications, particularly for data reduction study and scalability

purposes. Typically, a cluster is composed of a certain number of sensor nodes and has one cluster-head (CH) to manage the members. The main task of a sensor node is to sense the monitored field, detect events, perform quick local data processing, and then transmit the data to a specified CH. The CH acts as a gateway between its cluster members and the sink where it forwards data collected by the sensors after performing some processing operations. Figure 1 shows the cluster-based topology considered in this work in which a direct communication between sensors and their CH is established. Then, we consider that the data transmission between sensors-CHs and CHs-sink is performed in a periodic manner. Accordingly, we propose an energy-efficient mechanism which performs data prediction at sensor nodes as well as CHs.

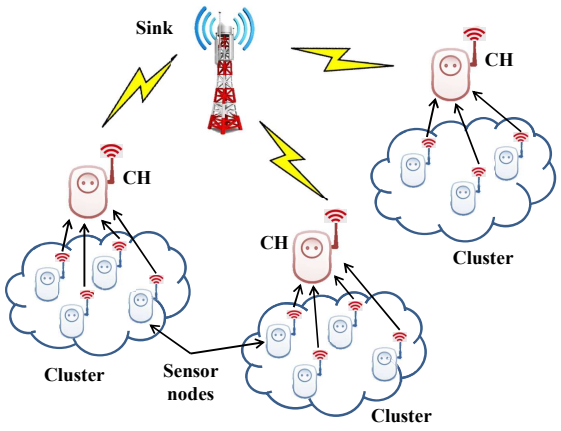


Fig. 1 Cluster-based topology for WSN.

Unfortunately, the formation of clusters is a critical task in sensing applications that imposes several challenges. Examples of such challenges include the CH selection, assigning members to clusters, cluster connectivity, on-/in-cluster data communication, etc. [28–36]. For the sake of simplicity, this paper considers a geographical cluster scheme in which each sensor member is assigned to nearest CH.

4 Data Prediction Model at Sensor Node

In almost WSN applications, data can be collected by the sensor nodes in three different ways: on-demand, event-based or periodic-based [35]. In the first collection model, the sensors collect the data according to a sink request. Such model is usually used to know about the status of the monitored zone at a time selected by the end user. In event-based collection, the nodes send data toward

the sink when a predefined event occurs. Such model is suitable to detect natural disasters (forest fire, flood, volcano activity, etc.) and enemy infiltration. In the periodic model, the target zone is constantly monitored where data are periodically sent to the sink node. This model is mostly used to monitor natural, human and animal behaviors like climate change, plant progress, animal movements, monitoring patient vital signs and tracking of elderly persons. Indeed, periodic collection model has a great number of applications nowadays but it introduces many challenges to WSN like data management and energy consumption. In this paper, we are interested in the periodic data collection model under the cluster-based WSN architecture. In this section, we propose a prediction model that uses the least squares approximation method in order to reduce the huge amount of data sent from each sensor to its appropriate CH.

4.1 Periodic WSN: Notations and Challenges

In periodic WSN, each sensor collects data for a period of time then it send them together to the CH, instead of directly sending each of one. Hence, each period is divided into a set of τ equal slots where a new reading is sensed at each slot. Therefore, each sensor node N_i will form a vector of τ data readings at the end of each period as follows: $D_i = [d_{i_1}, d_{i_2}, \ldots, d_{i_\tau}]$.

Indeed, the periodic collection model produces a high redundancy level among the collected data due to several reasons:

- First, the spatial correlation between the sensors which are mostly randomly dispersed over the target zone.
- Second, the temporal correlation between data collected by the sensors resulted from the huge amount of data required to collect in periodic applications.
- Third, the dynamic of the monitored condition which can slow down or speed up during the periods.
- Fourth, the sampling rate of the sensor which leads, in case of small slot time, to collected similar data.
- Fifth, the size of the entire period (i.e. τ) where small value of τ generated more redundancy among data collected in each period (i.e. D_i).

As a result for the redundancy, the periodic sensor networks face three major challenges:

- *Depleting overall network energy*: Data transmission consumes lots of the available energy in th sensors. Thus, sending redundant and useless data leads to overload the network with unnecessary transmission that consumes energy of the network devices (sensors and CHs) and minimizing its lifetime.
- *Complicating makers' decision*: Discovering knowledge and information from data received at the sink node is a fundamental operation in WSN in order to take decisions. However, the huge amount of data collected with

the existing redundancy makes such task a complicated mission for the makers' decision.
- *Delaying action response*: Data latency is one of the most critical operations in WSNs, especially for applications in healthcare and natural disaster where a fast response must be taken at the right time. However, processing time needed to eliminate redundancy among data collected can delay the time response required for an occurred event.

In the next section, we propose a data prediction model based on the least squares approximation method in order to reduce the amount of data sent from each sensor to the CH.

4.2 Integrating Least Squares Approximation (LSA) Method at Sensor Tier

In this section, we aim to reduce the amount of data periodically collected by each sensor node during each period, i.e. D_i, before sending to the CH. Thus, we propose to integrate the LSA method into the sensor processing in order to create a predictive model for the collected data to send later toward the CH. Indeed, LSA [37] is one of the most standard approaches used in statistical analysis that aims to determine the curve that best describes the relationship between expected and observed data sets by minimizing the sums of the squares of deviation between observed and expected values. Subsequently, each sensor finds the LSA polynomial that fits its data in D_i then, it send the LSA coefficient set toward the sink which, in its turn, it can regenerate all the raw data based on the received coefficient equation.

Indeed, the processing time needed to calculate the LSA polynomial of degree k will be huge, especially when the period size τ is high. Hence, in order to reduce the time complexity of LSA, we propose to select a subset of r readings, named as R_i, from D_i to find the corresponding polynomial. R_i can be formed based on the following equation:

$$R_i = \{(s_{1+j\times\lfloor\tau/(r-1)\rfloor}, d_{1+j\times\lfloor\tau/(r-1)\rfloor}), (s_\tau, d_\tau)\} \quad (1)$$

where $s_{1+j\times\lfloor\tau/(r-1)\rfloor}$ are all readings collected at slot numbers $s_{1+j\times\lfloor\tau/(r-1)\rfloor}$ (such that $j \in [0, \tau]$ and $1 + j \times \lfloor\tau/(r-1)\rfloor < \tau$) and d_τ is the last reading in D_i.

After selecting the readings, the sensor computes the LSA polynomial by resolving the equations mentioned in definition 1. Then, the sensor will send only the coefficient set of LSA polynomial $\mathcal{C}_i = \{a_0, a_1, \ldots, a_k\}$ which is necessary to recalculate the values of all raw readings.

Formally, Algorithm 1 describes the prediction method based on LSA applied at the node level. The algorithm takes data readings collected by each sensor at every period and then returns the LSA coefficient set that will send to its CH. First, the node selects the readings at indexes determined by equation 1 (lines 1-4). Then, the algorithm formulates and solves the system of equations determined by the Definition 1. Finally, the sensor sends the coefficient set of the polynomial to the CH (lines 8-9).

Algorithm 1 Prediction Node Algorithm.

Require: Node: N_i, Period size: τ, Data readings: $D_i = [d_{i_1}, d_{i_2}, \ldots, d_{i_\tau}]$, LSA degree: k, Subset size: r.
Ensure: Coefficient set: $\mathcal{C}_i$.
1: $R_i \leftarrow \emptyset$
2: **for** $j = 1$ to $r - 1$ **do**
3: $\quad R_i \leftarrow R_i \cup \{(s_{1+j\times\lfloor\tau/(r-1)\rfloor}, d_{1+j\times\lfloor\tau/(r-1)\rfloor})\}$
4: **end for**
5: $R_i \leftarrow R_i \cup \{(s_\tau, d_\tau)\}$
6: formulate LSA equations
7: solve the LSA equations
8: find the set $\mathcal{C}_i$ of coefficients a_0 to a_k
9: **return** $\mathcal{C}_i$

5 Data Prediction Model at CH Node

The CH will receive the sets of LSA coefficients sent from each sensor member, $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$, at the end of each period. Subsequently, the CH receives $n \times (k+1)$ data values at each period, where n is the number of sensor nodes and $k+1$ is the size of each coefficient set. This amount of data will be of huge size especially WSN where an enormous number of sensor nodes could be deployed in the network along with an increasing value of LSA degree k. In this section, we aim to reduce the size of data sets sent from each CH to the sink node by proposing a prediction pattern to all LSA sets, instead of sending each LSA coefficient set. Therefore, the available energy of CH will be saved while the raw data can be periodically reconstructed at the sink.

5.1 LSA Method at CH Node

As shown before, LSA can effectively reduce the amount of data sent from sensors to CH. In this section, we aim to adapt again the LSA method to be applied at the CH. Consider the following matrix that represents the LSA coefficient sets coming from all sensors at the end of each period:

$$\begin{array}{c} \\ \mathcal{C}_1 = \\ \mathcal{C}_2 = \\ \vdots \\ \mathcal{C}_n = \end{array} \begin{array}{c} \begin{array}{cccc} a_0 & a_1 & \ldots & a_k \end{array} \\ \begin{pmatrix} a_{1_0} & a_{1_1} & \ldots & a_{1_k} \\ a_{2_0} & a_{2_1} & \ldots & a_{2_k} \\ \vdots & \vdots & \ldots & \vdots \\ a_{n_0} & a_{n_1} & \ldots & a_{n_k} \end{pmatrix} \end{array} \tag{2}$$

In order to reduce the size of the matrix, we propose to apply LSA method at each column separately thus reducing the number of rows in the matrix.

Subsequently, we aim to find the LSA polynomial of degree k' for each column $\mathcal{A}_0 = \{a_{1_0}, a_{2_0}, \ldots, a_{n_0}\}$, $\mathcal{A}_1 = \{a_{1_1}, a_{2_1}, \ldots, a_{n_1}\}$ and $\mathcal{A}_k = \{a_{1_k}, a_{2_k}, \ldots, a_{n_k}\}$. The LSA polynomial for each column $\mathcal{A}_i$ can be found similarly to the process mentioned in Definition 1, where the $x-axis$ represents the sensor number and the y-axis represents the coefficient set values. Subsequently, the LSA polynomial for each column $\mathcal{A}_i$ can be represented as: $y_i = a'_{0_i} + a'_{1_i}x + a'_{2_i}x^2 + \cdots + a'_{k'_i}x^{k'}$. Furthermore, in order to maintain the accuracy of the polynomial, the value of k' should be greater than that selected for k.

Therefore, the CH will convert, at each period, the matrix of dimensions $(n,\ k+1)$ in equation 2 to a matrix of dimensions $(k+1,\ k'+1)$ as follows:

$$\begin{array}{c} \\ \mathcal{C}'_0 = \\ \mathcal{C}'_1 = \\ \vdots \\ \mathcal{C}'_k = \end{array} \begin{array}{c} \begin{array}{cccc} a'_0 & a'_1 & \ldots & a'_{k'} \end{array} \\ \begin{pmatrix} a'_{0_0} & a'_{1_0} & \ldots & a'_{k'_0} \\ a'_{0_1} & a'_{1_1} & \ldots & a'_{k'_1} \\ \vdots & \vdots & \ldots & \vdots \\ a'_{0_k} & a'_{1_k} & \ldots & a'_{k'_k} \end{pmatrix} \end{array} \tag{3}$$

Algorithm 2 describes the prediction model applied at the CH level. As input, the algorithm takes all set coefficients sent from the sensor nodes and returns, as output, a new and reduced set of coefficients to send to the sink. Briefly, the CH searches the coefficient values at the same index (lines 2-5) then, it calculates the LSA polynomial of degree k' based on the equations mentioned in Definition 1. Finally, the CH sends the new set of coefficients to the sink for raw data regeneration purpose (lines 6-10).

Algorithm 2 Prediction CH Algorithm.

Require: LSA Coefficient Sets: $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$, LSA degree: k'.
Ensure: Coefficient set: $\mathcal{C}' = \{\mathcal{C}'_1, \mathcal{C}'_2, \ldots, \mathcal{C}'_k\}$.

1: **for** $i = 0$ to k **do**
2: $T \leftarrow \emptyset$ // T is a temporary list
3: **for** $j = 1$ to n **do**
4: $T \leftarrow T \cup \{(j, \mathcal{C}_{j_i})\}$
5: **end for**
6: formulate equations of Definition 1 based on T
7: solve the equations of Definition 1 for T
8: find $\mathcal{C}'_i$ of coefficients a'_0 to $a'_{k'}$
9: $\mathcal{C}' \leftarrow \mathcal{C}' \cup \{\mathcal{C}'_i\}$
10: **return** $\mathcal{C}'$
11: **end for**

5.2 Raw Data Regeneration at Sink Node

Once the final coefficient sets are periodically received, the last process phase in our technique is starting and aims to reconstruct raw data at the sink

node. Obviously, the main mission is to find approximate reading values to those collected by the sensor rather than finding similar ones, which is not possible with regression and polynomial interpolation. Indeed, the accuracy of our model is highly dependent on the polynomial degree selected at the sensor node (e.g. k) and the CH level (e.g. k'); more the values of k and k' more the accuracy of regenerated data is.

The process of data regeneration can be done inversely to that used when searching the LSA polynomials at sensor and CH nodes. When the sink receives the reduced matrix mentioned in equation 3, it applies the following steps to retrieve raw data collected by all sensors:

- *Step 1*: Searches the coefficient set for each sensor. This can be done by formulating the equation for each row in the matrix of equation 3; the first row allows to find the first coefficient for the polynomial of each sensor, the second row finds the second coefficient and so on. Therefore, the following equations can be formulated according to equation 3:

$$\begin{cases} y'_0 = a'_{0_0} + a'_{1_0}x + \ldots + a'_{k'_0}x^{k'} \\ y'_1 = a'_{0_1} + a'_{1_1}x + \ldots + a'_{k'_1}x^{k'} \\ \vdots \quad = \ldots + \ldots \quad + \ldots + \ldots \\ y'_k = a'_{0_k} + a'_{1_k}x + \ldots + a'_{k'_k}x^{k'} \end{cases}$$

- *Step 2*: For each of the obtained equation, it computes the value of $y'_i(x)$ for all $x \in [1, n]$. This leads to find all set coefficients for all sensors determined in equation 2 as follows:

$$\begin{array}{c} \\ \mathcal{C}_1 = \\ \mathcal{C}_2 = \\ \vdots \\ \mathcal{C}_n = \end{array} \begin{array}{c} \begin{array}{cccc} a_0 & a_1 & \ldots & a_k \end{array} \\ \begin{pmatrix} y'_0(1) & y'_1(1) & \ldots & y'_k(1) \\ y'_0(2) & y'_1(2) & \ldots & y'_k(2) \\ \vdots & \vdots & \ldots & \vdots \\ y'_0(n) & y'_1(n) & \ldots & y'_k(n) \end{pmatrix} \end{array}$$

- *Step 3*: From the above matrix, finds the polynomial equation for row which represents the data collected by each sensor as follows:

$$\begin{cases} y_1 = y'_0(1) + y'_1(1)x + \ldots + y'_k(1)x^k \\ y_2 = y'_0(2) + y'_1(2)x + \ldots + y'_k(2)x^k \\ \vdots \quad = \ldots \quad + \ldots \quad + \ldots + \ldots \\ y_n = y'_0(n) + y'_1(n)x + \ldots + y'_k(n)x^k \end{cases}$$

- *Step 4*: For each of the above equation, it computes the value of $y_i(x)$ for all $x \in [1, \tau]$ in order to find the data set for D_i for each sensor as follows:

$$\begin{array}{l} D_1 = [\, y_1(1),\, y_1(2),\, \ldots,\, y_1(\tau)\,] \\ D_2 = [\, y_2(1),\, y_2(2),\, \ldots,\, y_2(\tau)\,] \\ \vdots \quad \vdots \qquad \vdots \qquad \ldots\ \vdots \\ D_n = [\, y_n(1),\, y_n(2),\, \ldots,\, y_n(\tau)\,] \end{array}$$

5.3 Illustrative Example

In this section, we show an illustrative example for how to construct the LSA polynomial at the CH as well as the reconstruction of raw data at the sink node (Figure 2). First, assume that 5 sensor nodes collect their reading sets, e.g. D_1 to D_5 respectively, during a period then each of them computes its LSA polynomial equation and sends its LSA coefficient set to the CH (see section IV.D). After receiving the sets of coefficients from the 5 sensors, the CH selects a LSA degree k' equals to 2 then it finds the LSA polynomial for each column in the received coefficient matrix; for instance, the equation y'_0 corresponds to the column $[5, 6, 3, 4, 4]$, y'_1 corresponds to the column $[3, 1, 4, 1, -1]$ and so on. After that, the CH sends the LSA coefficients for each equation to the sink, e.g. $\mathcal{C}'_0$ to $\mathcal{C}'_3$. Later, the sink receives the coefficient sets and follows the reverse process to reconstruct the raw data; first, it regenerates the polynomial equations based on the received coefficient sets, reconstructs the coefficient sets for each sensor by scanning x-values from 1 to 5, regenerates the LSA polynomial for each sensor and, finally, finds raw data for each sensor by computing y-values for all x-values from 1 to τ.

6 Simulation Data Description

WSNs support a huge number of applications ranging from weather to industrial and healthcare environments. In order to evaluate the relevance of our technique, we conducted extensive simulations based on real sensor data collected from various domains. The objective of these simulations is to test the performance of our technique against different types of data and application circumstances. In the next sections, we show the description of each sensor data along with the simulation setup.

6.1 Weather Data Collected at Intel Lab

The first kind of sensor data are picked up from sensors deployed in the Intel Berkeley Research lab [38]. In such network, 46 sensor nodes of type Mica2Dot with weather boards that collect temperature, humidity, light and voltage values once every 31 seconds. Sensor nodes are deployed for about 40 days starting at February 28^{th} where 2.3 reading values are approximately collected during this period. Furthermore, the dimensions of the monitored zone (e.g. lab) were 42×33 meters where the indoor sensing range for each sensor is 25 meters. In our simulation, we assume a CH located at the center of the lab with a sink distant at 50 meters from the CH. Finally, for the sake of simplicity, we are interested, in this paper, in the temperature readings field where other fields are treated in the same manner.

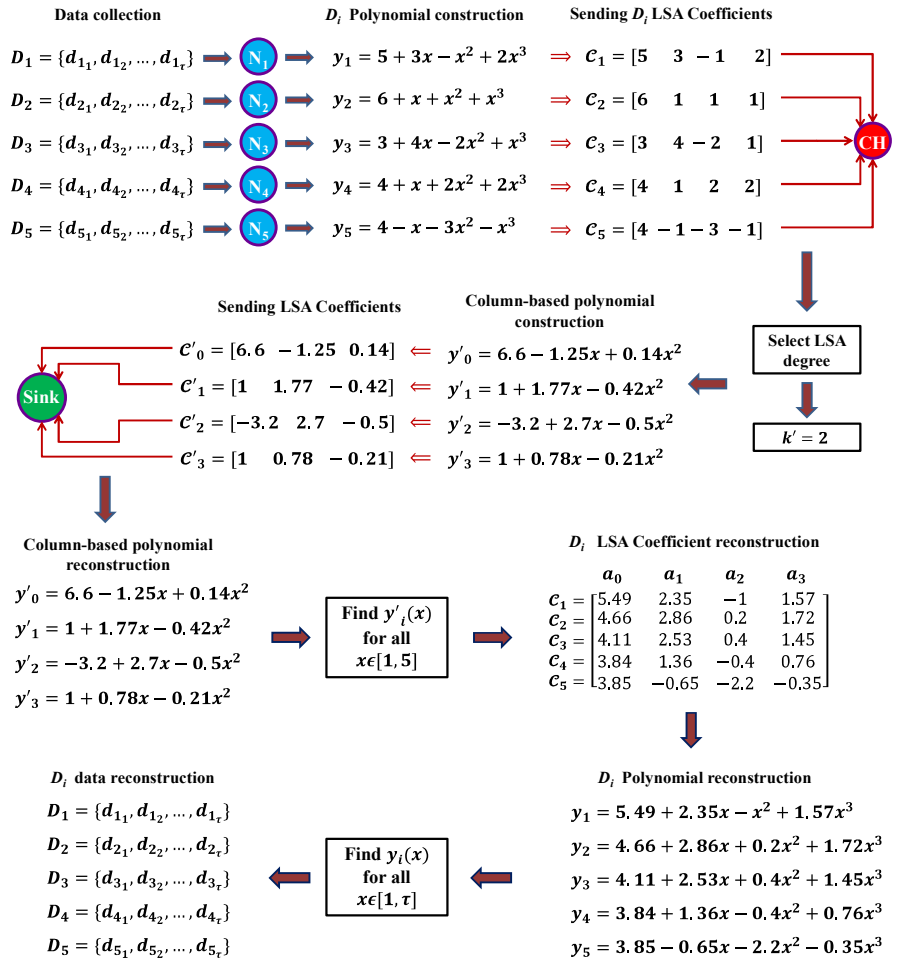


Fig. 2 Illustrative example of our technique at CH and sink nodes.

6.2 Underwater Data Collected by ARGO Project

The second type of data evaluated in our simulation are coming from sensors used in Argo project [39]. Argo is an underwater sensor network that deploys more than of 3800 free-drifting profiling nodes that measures temperature, salinity and velocity data of the upper 2000 m of the ocean. Float sensors collect data in a daily basis where collected data are sent at a period of 10 days to a float navigator which, in its turn, forwards them to the end centers through satellite communication. In our simulation, we focus on data sensed by 119 float sensors dispersed in the Indian ocean over an area of 5000 $\times$ 5000 m. About 1.5 million reading profiles are collected by the sensors during approximately 3 months of deployment in 2014. Data collected by the float sensors are sent

to the float navigator, located at the center of the ocean zone, by means of wireless acoustic links. Finally, in our simulation, we are interested in the salinity readings.

6.3 Healthcare Data Collected by MIMIC

In the third kind of simulation, we used real medical readings collected from the online MIMIC (Multiple Intelligent Monitoring in Intensive Care) database [40]. MIMIC has 72 patient records that contains data about vital signs obtained from bedside ICU monitors. Vital signs data include heart rate, blood pressure (mean, systolic, diastolic), respiration rate, oxygen saturation, etc. Every one second, the medical sensors capture new vital signs for a patient during his stay in hospital. The size of data collected exceeds 5 million medical readings for all patients with an average of 70000 readings for each one. In our simulation, we are interested in the heart rate vital sign. We assume that sensors send their data to a PDA located at the emergency staff center to detect and analyze urgent situations of patients.

6.4 Methane Gaz-based Industrial Sensor Data

The last type of data is dedicated to industrial sectors where readings about methane gas have been collected [41]. Mostly, the leakage of methane gas leads to a critical explosion of an industry thus, constantly monitoring this gas is an essential operation in most industries. A small SensorScope network contains 16 chemical sensors has been deployed at an industry located at the Grand-St-Bernard pass at 2400 meters between Switzerland and Italy. The industry area is about $140 \times 120\, m^2$. More than 4 million methane readings have been collected by each sensor at a sampling of 1 reading per minute. The collected data are sent to a central device (represent the CH) located at the center of the industry.

7 Simulation Results

In this section, we show the simulation results obtained of our technique over each of the described data. We compared the results to the PFF technique proposed in [23] and the S-LEC compression method in [42]. We varied the parameter variables as follows:

- The period size τ takes the values: 50, 100, 250 and 500 readings.
- The LSA polynomial degree at sensor (k) takes the values: 2, 4, 6 and 8.
- The LSA polynomial degree at CH (k') takes the values: 6, 8 and 10.
- The selected size of readings at the sensor r takes the values: 5 for $\tau = 50$, 8 for $\tau = 100$, 10 for $\tau = 250$ and 12 for $\tau = 500$.

The performance of our technique is tested according to the following metrics:

- Variation between raw and regenerated data at sensor and CH levels.
- Data ratio sent from each sensor to the CH and between CH and sink.
- Execution time at sensor and CH nodes.

7.1 Data Redundancy Study

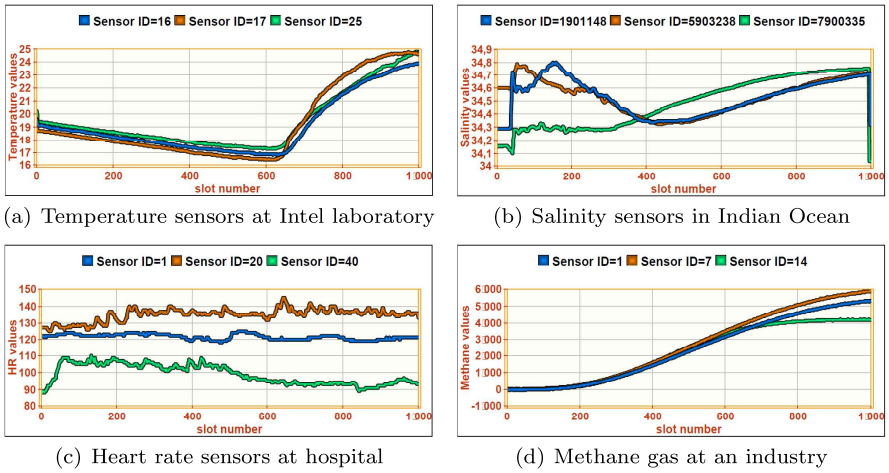
We aim first to study the redundancy between data collected by various types of sensors and for different types of applications. As mentioned in section IV.A, redundancy is mostly happened in WSNs for several reasons, especially between neighbouring nodes. Figure 3 shows real examples of redundancy generated inside each sensor or among nearest nodes, for the first 1000 readings collected by each sensor. In each subfigure, we selected three random sensors collecting temperature, salinity, heart rate and methane gas respectively then we study the variation of their collected data. The obtained results allow several observations: first, the successive readings collected by a sensor are almost similar (Figs 3(a) to 3(d)). Second, neighbouring sensors are probably generating redundant data like in Figure 3(a). Third, distant sensors can be also temporally correlated depending on the slow variation of the monitored environment. This can be confirmed by the 3 sensors collecting salinity condition which are very dispersed in the ocean however they collect very similar readings. Fourth, spatial correlation does not have any effect on redundancy between collected data in some critical application like healthcare; for instance, Figure 3(c) shows 3 neighboring medical sensors in which data collected is dependent on the situation of the patient (critical for patient IDs 1 and 20, and normal for patient ID 40). Fifth, sometimes, it may occur that sensors take similar data for a period of time then, later, they collect different readings (see Figure 3(d)). This happens when an event occurs nearest a sensor and far from others.

7.2 Discussion of Weather Data Results

In this section, we discuss the results of simulation conducted over the temperature data collected by the sensor nodes deployed at the Intel research Berkeley lab according to the following metrics:

7.2.1 Variation Between Raw and Regenerated Data at Sensor

Figure 4 shows the efficiency of the LSA prediction model at the sensor node level when fixing the period size to 250 and the LSA polynomial degree to 6. According to the obtained results, we can clearly observe that our prediction model ensures a high level of data accuracy where the regenerated data are



(a) Temperature sensors at Intel laboratory (b) Salinity sensors in Indian Ocean

(c) Heart rate sensors at hospital (d) Methane gas at an industry

Fig. 3 Redundancy among readings collected by various sensors, $\tau = 1000$.

much closer to those collected by the sensor. Subsequently, the worst scenario happens at the slot number 23 where an error of about 0.1 is noticed between raw and regenerated data. This error is almost negligible and it does not affect the accuracy of the temperature condition.

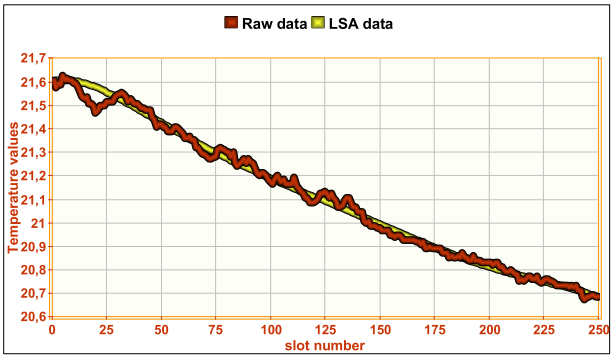
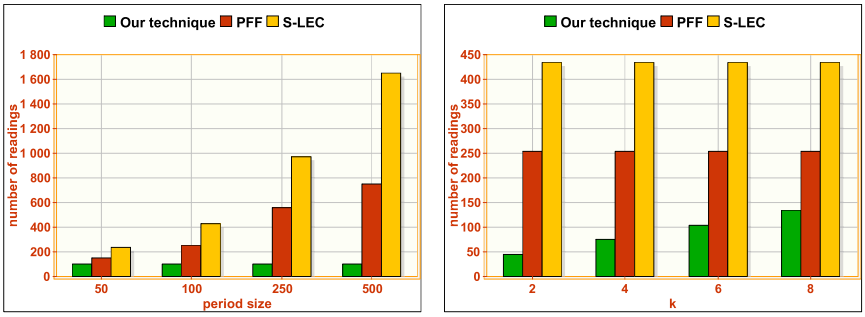


Fig. 4 Raw data vs regenerated data using LSA, $\tau = 250$, $k = 6$.

7.2.2 Data Transmission Ratio at Sensor

Figure 5 shows the data transmission ratio from each sensor to its CH, after applying the LSA method and by varying the value of the prediction degree and the period size. The obtained results show the efficiency of our technique in reducing the amount of sent data compared to other existing techniques.

Particularly, our technique reduces up to 86% and 93% compared to PFF and S-LEC when fixing k (Figure 5(a)), and up to 82% and 89% when fixing period size (Figure 5(b)). Such reduction is performed due to the reduced set of coefficients sent with LSA method while the aggregation and compression operations impose a minimum portion of data to be sent instead of the entire raw data. As a result, our technique will save the sensor energy more than that with PFF and S-LEC since energy consumption is highly dependent on the amount of transmitted data.



(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

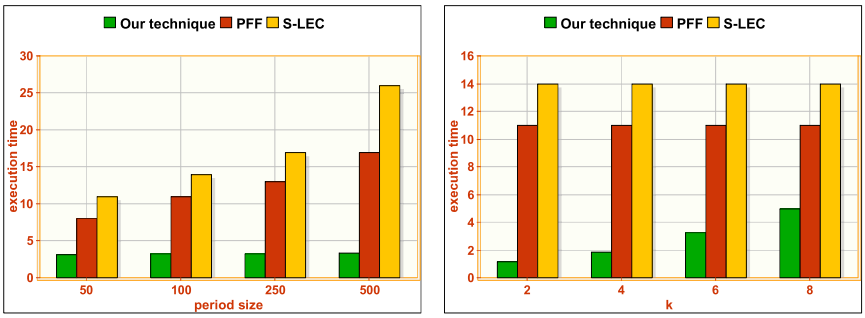
Fig. 5 Data transmission ratio from sensor to CH.

7.2.3 Execution Time at Sensor

Indeed, complexity takes a play role in IoT networks due to the limited resources of the sensors. In this paper, we study the complexity metric of each technique in terms of the processing time required to execute their algorithms at the sensor level (Figure 6). Thus, a proposed technique must minimize the execution time in order to deliver the packet to the sink as soon as possible. The results show that the execution time of our technique is highly optimized to those of PFF (i.e. 2 to 9 times of minimization) and S-LEC (i.e. 3 to 11 times of minimization). In addition, we show that the execution time of our technique is independent from the prediction degree and the period size unlike those of PFF and S-LEC that increase with the increase of the period size.

7.2.4 Variation Between Raw and Regenerated Data at CH

In this section, we aim to study the efficiency of the LSA prediction model at the CH level when fixing the LSA polynomial degree to 8. Figure 7 shows the coefficient value sent from each sensor to the CH (e.g. raw coefficient) and the regenerated coefficient value when applying the LSA method. Obviously, a small variation of the LSA coefficient can lead to a high change in the raw



(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

Fig. 6 Execution time required for each algorithm on the sensor.

data of the sensor. The obtained results show a high convergence between raw coefficient values and those regenerated by the CH before sending to the sink. Therefore, our prediction model canbe efficiently used at sensor and CH levels.

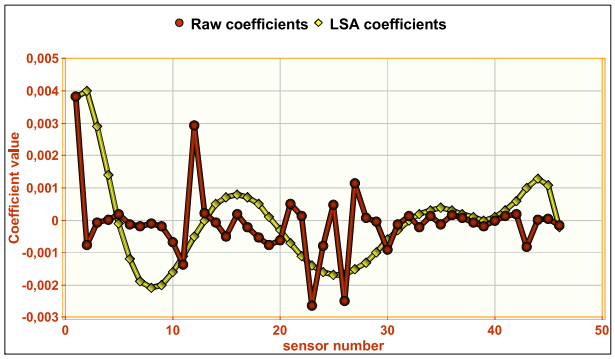


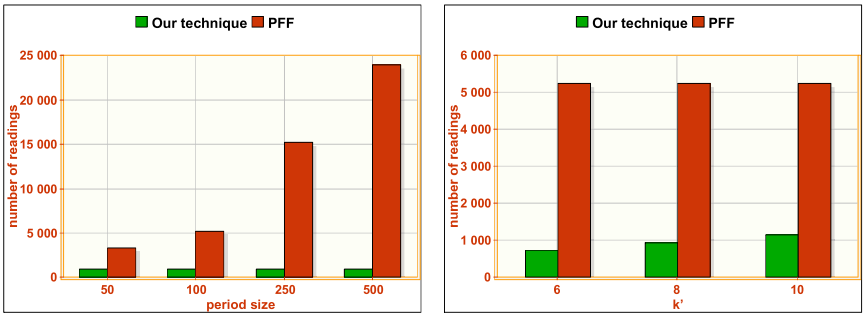
Fig. 7 Raw coefficients vs regenerated LSA coefficients using LSA, $\tau = 250$, $k = 6$, $k' = 8$.

7.2.5 Data Transmission Ratio at CH

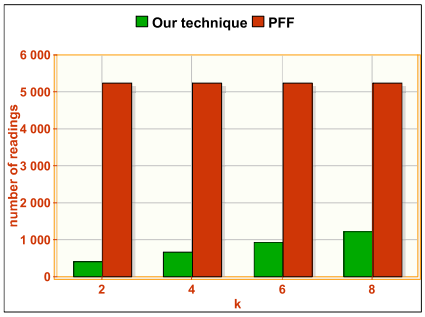
In this section, we show the efficiency of LSA prediction model at the CH level in terms of reducing the amount of data sent to the sink node (Figure 8). We studied the efficiency in terms of three variables: the period size (τ), the LSA degree at sensor (k) and at CH (k'). The obtained results reflect a huge difference between the data transmitted using our technique and PFF; our technique reduces from 71% to 96% of data sent toward the sink. Furthermore, the following observations can be noticed:

– by increasing the period size from 50 to 500, the number of sent readings using PFF increases while it still fixed using our technique. This is because,

the similarities between readings decrease in PFF when the period size increases while our technique does not dependent on the period size.
– by icnreasing the LSA degree at the CH, the data transmission from the CH increases because more coefficient values will sent to the sink.
– by increasing the LSA degree at the sensor, the CH sends more data to the sink.



(a) $k = 6$, $k' = 8$ (b) $k = 6, \tau = 100$

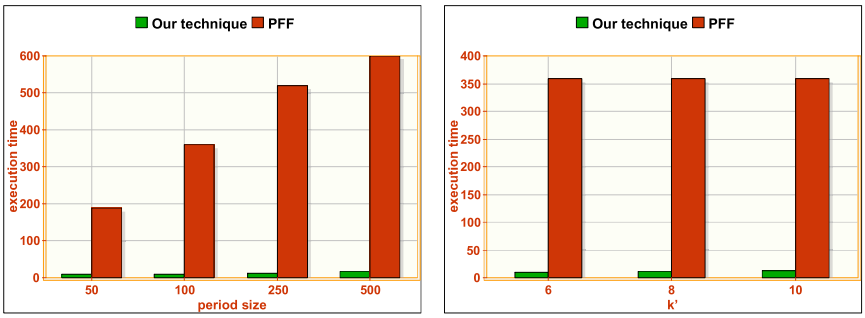


(c) $\tau = 100, k' = 8$

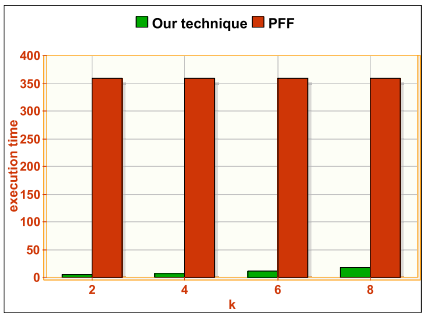
Fig. 8 Data transmission ratio from CH to the sink.

7.2.6 Execution Time at CH

Figure 9 shows the execution time required at the CH when applying our technique and PFF, and when varying τ, k and k'. As expected, our technique largely outperforms PFF in terms of execution time in all cases. Subsequently, our technique reduces 19 to 37 times the execution time when varying the period size (Figure 9(a)), from 27 to 36 when varying the LSA CH degree (Figure 9(b)) and from 20 to 64 when varying the LSA sensor degree (Figure 9(c)), compared to PFF. This is because the PFF works by searching the similarities between every pair of sets which takes a long time processing unlike our technique which computes the equation coefficients based on predefined formulas.



(a) $k = 6$, $k' = 8$ (b) $k = 6, \tau = 100$



(c) $\tau = 100, k' = 8$

Fig. 9 Execution time required for each algorithm on the CH.

7.3 Discussion of Underwater Data Results

In this section, we discuss the results of simulation conducted over the salinity data collected by the underwater sensor nodes deployed in the Indian ocean according to the following metrics:

7.3.1 Variation Between Raw and Regenerated Data at Sensor

Similarly to Figure 4, Figure 10 shows the difference between raw data and those generated by LSA model collected about salinity condition. Obviously, the salinity condition is varying very slowly compared to the temperature condition, where the salinity values changed in range $[34.5, 35.1]$ (a variation of 0.6 degree) for a period of 250 readings. Consequently, the LSA model produces a very small error of, at most, 0.05 compared to the raw data.

7.3.2 Data Transmission Ratio at Sensor

Similarly to Figure 5, Figure 11 shows the average number of salinity readings sent from each sensor to the CH. Indeed, using LSA model, the sensor sends

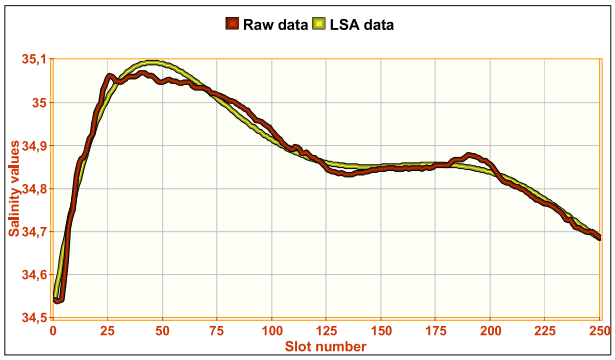
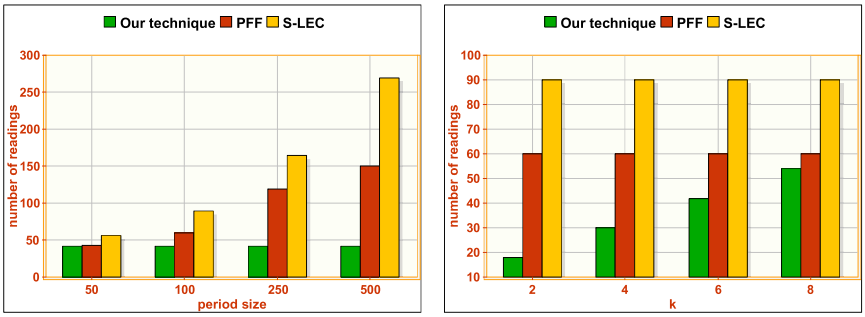


Fig. 10 Raw data vs regenerated data using LSA, $\tau = 250$, $k = 6$.

the same number of values (e.g. LSA coefficient set) which is dependent on the LSA degree and independent from the monitored condition. However, the amount of data sent using PFF and S-LEC are highly dependent on the monitored condition which can generate less or more redundant data. The obtained results show that our technique reduces up to 72% and 84% compared to PFF and S-LEC, when varying τ and k.



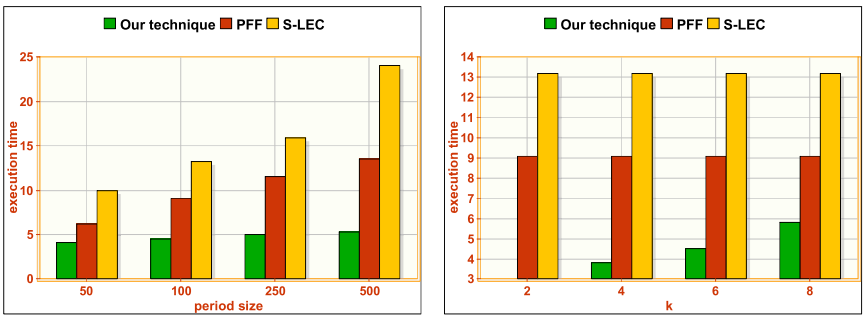
(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

Fig. 11 Data transmission ration from sensor to CH.

7.3.3 Execution Time at Sensor

Figure 12 shows the execution time required to apply each algorithm at underwater sensor node. The obtained results show that our technique accelerates time processing at the sensor from 2 to 3 times compared to PFF and from 2 to 5 compared to S-LEC. Compared to temperature results, our technique gives less performance in comparison with PFF and S-LEC; this is because the salinity condition produces more redundancy than temperature among

the collected data which increases the performance of PFF and S-LEC while that of our technique still fix.



(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

Fig. 12 Execution time required for each algorithm on the sensor.

7.3.4 Variation Between Raw and Regenerated Data at CH

Similarly to Figure 7, Figure 13 shows the difference between raw coefficients sent from the sensor and those regenerated by the LSA for the salinity condition. The obtained results reveal a well convergence between both curves except for few values produced at sensor IDs ranging from 105 to 120. This error is expected since when the coefficient value of any sensor diverges far from the other sensors (like the case of sensor IDs 105 and 110), then the regenerated curve using LSA with simultaneously diverge for the nearest sensors.

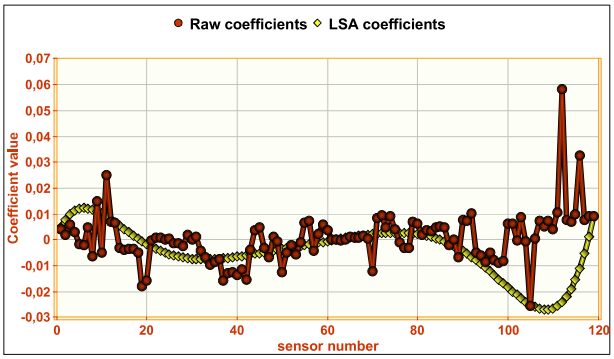
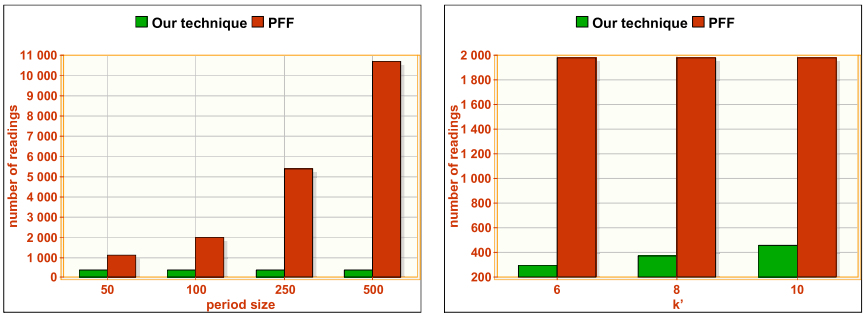


Fig. 13 Raw coefficients vs regenerated LSA coefficients using LSA, $\tau = 250$, $k = 6$, $k' = 8$.

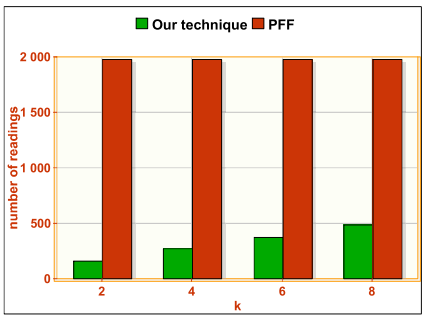
7.3.5 Data Transmission Ratio at CH

Figure 14 shows the number of salinity readings sent from each CH to the sink. As shown, we notice that the data transmission at CH is highly minimized using our technique compared to that sent with PFF; using our technique, the CH only sends, in the best case, 1.4% among the raw to the sink while PFF sends at least 5% of raw data. Hence, our technique allows to reduce from 66% to 96% of data transmission at CH compared to PFF.



(a) $k = 6$, $k' = 8$

(b) $k = 6, \tau = 100$

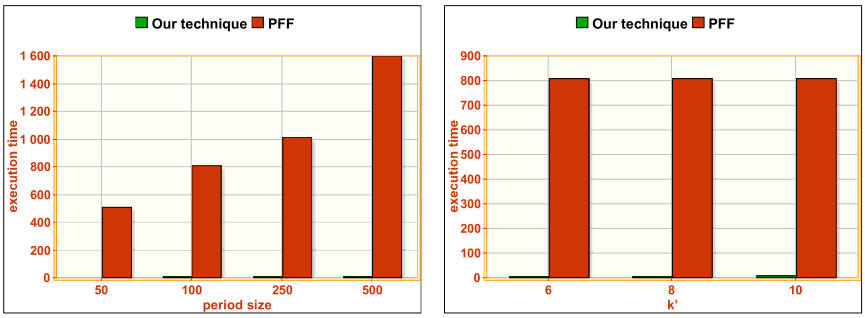


(c) $\tau = 100, k' = 8$

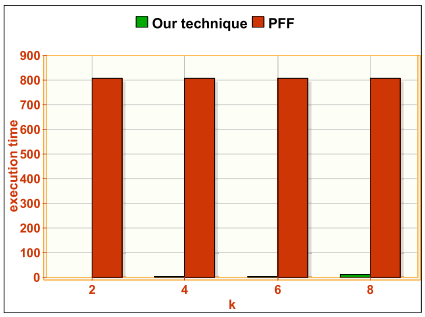
Fig. 14 Data transmission ratio from CH to the sink.

7.3.6 Execution Time at CH

Figure 15 shows the processing time needed for each algorithm, our technique and PFF, at the CH when varying τ, k and k'. Similar to temperature condition, our technique largely outperforms the execution time of prediction than that required for PFF; An enhancement of 98 times has been detected using our technique at the CH level compared to PFF.



(a) $k = 6$, $k' = 8$ (b) $k = 6, \tau = 100$



(c) $\tau = 100, k' = 8$

Fig. 15 Execution time required for each algorithm on the CH.

7.4 Discussion of Medical Data Results

In this section, we discuss the results of simulation conducted over the heart rate data collected by the MIMIC according to the following metrics:

7.4.1 Variation Between Raw and Regenerated Data at Sensor

Similarly to Figure 4 and Figure 10, Figure 16 studies the variation of between raw and regenerated data for heart rate medical sensor. Indeed, heart rate condition usually varies more quickly than temperature and salinity, where the heart rate of a patient can change from minute to minute according to its situation. The figure shows the raw data of a normal patient where his heart rate varies between 69 and 71 whilst, the regenerated data appear nearly to the raw data. In the worst case, an error of 1 degree is noticed which still the patient in his normal status. Therefore, our model can be efficiently used in critical application while keeping the status of the monitored condition as it is.

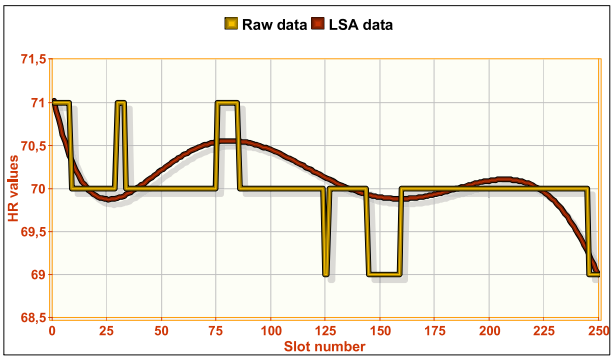
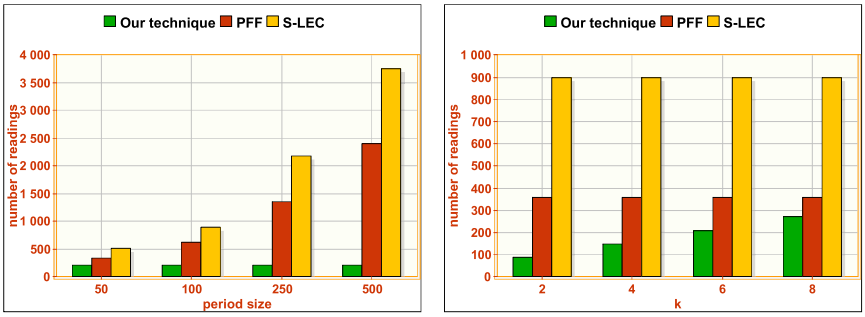


Fig. 16 Raw data vs regenerated data using LSA, $\tau = 250$, $k = 6$.

7.4.2 Data Transmission Ratio at Sensor

Figure 17 shows the average number of heart rate readings periodically sent from each sensor to the CH. As mentioned before (Figure 16), the variation of heart rate condition is more noticeable compared to temperature and salinity conditions thus, the performance of PFF and S-LEC will degrade. The obtained results show that our technique reduces up to 91% and 94% of readings sent from each sensor compared to PFF and S-LEC, when varying τ and k.



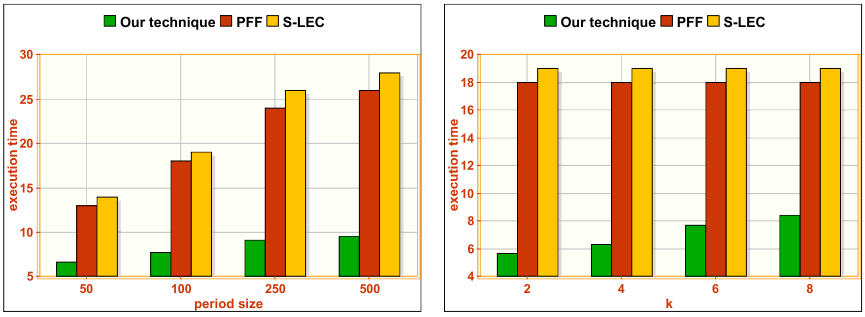
(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

Fig. 17 Data transmission ratio from sensor to CH.

7.4.3 Execution Time at Sensor

Figure 18 shows the execution time required to apply each algorithm at medical sensor nodes. The obtained results show that our technique accelerates time processing at the sensor about 3 times compared to PFF and S-LEC techniques. We can also notice that, due to the high variation between col-

lected data, PFF and S-LEC gives approximate execution time results when varying k and τ.



(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

Fig. 18 Execution time required for each algorithm on the sensor.

7.4.4 Variation Between Raw and Regenerated Data at CH

Figure 19 shows the raw coefficient values for heart rate sensors compared to coefficient values regenerated by LSA method. The obtained results are very similar to those obtained with salinity readings where the raw coefficient and LSA coefficient curves are very converged to each other except for some values generated for sensor IDs 40 to 50. However, it is important to notice that the maximum obtained difference between both curves does not exceed 0.12.

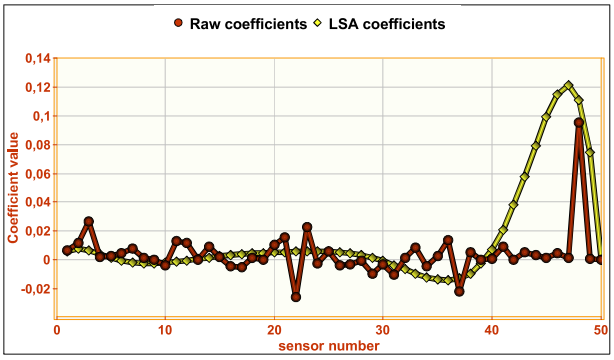
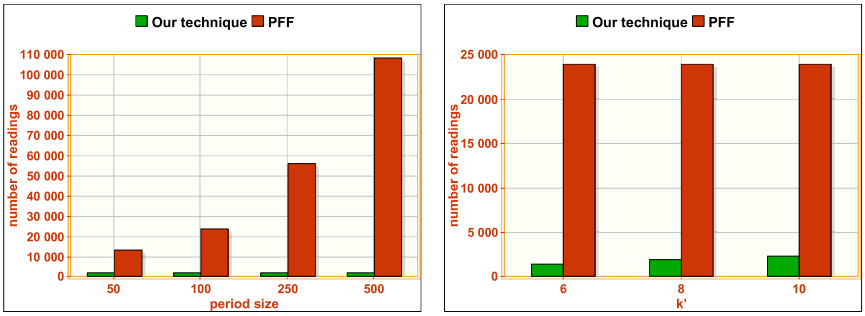


Fig. 19 Raw coefficients vs regenerated LSA coefficients using LSA, $\tau = 250$, $k = 6$, $k' = 8$.

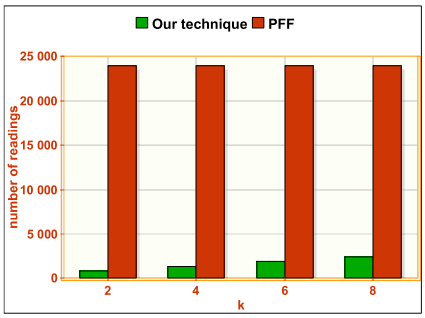
7.4.5 Data Transmission Ratio at CH

Figure 20 shows the number of heart rate readings sent from each CH to the sink. the obtained results shows that the data transmission at CH reduced by at least 86% when varying the period size, 90% when varying the LSA degree at the sensor and 89% when varying the LSA degree at the CH, compared to PFF technique.



(a) $k = 6$, $k' = 8$

(b) $k = 6, \tau = 100$

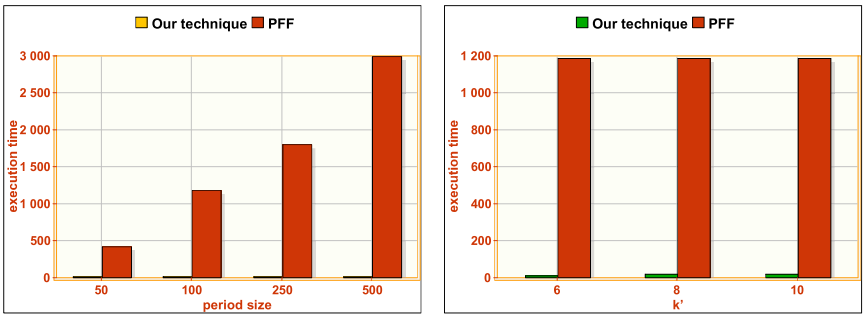


(c) $\tau = 100, k' = 8$

Fig. 20 Data transmission ratio from CH to the sink.

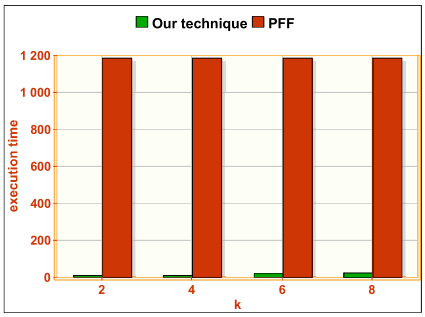
7.4.6 Execution Time at CH

Figure 21 shows the execution time required to apply both our technique and PFF over heart rate readings at the CH. Unlike temperature and salinity conditions, PFF gives less performance in terms of execution time because medical data varies quickly which decreases the redundancy among the collected data (thus increases time processing). Otherwise, our technique gives the same performance independent on the monitored conditions. Therefore, it accelerates the execution time from 23 to 142 times compared to PFF.



(a) $k = 6$, $k' = 8$

(b) $k = 6, \tau = 100$



(c) $\tau = 100, k' = 8$

Fig. 21 Execution time required for each algorithm on the CH.

7.5 Discussion of Industrial Data Results

In this section, we discuss the results of simulation conducted over the methane gas data collected by the chemical sensor nodes according to the same previous metrics:

7.5.1 Variation Between Raw and Regenerated Data at Sensor

Figure 22 shows the variation between the methane gas values compared between raw data and regenerated data using LSA model. Similarly to the results obtained with other conditions, LSA polynomial allows to regenerate raw data with high level accuracy. Hence, LSA prediction model proposed in our technique can be efficiently adapted to any type of sensor and applied in various domains.

7.5.2 Data Transmission Ratio at Sensor

Figure 23 shows the average number of methane gas readings sent from each chemical sensor to the CH. The variation of methane gas condition seems

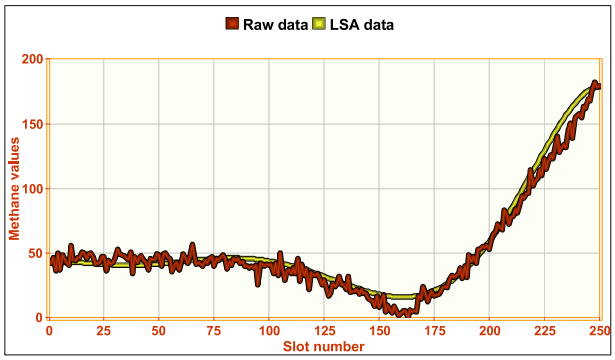
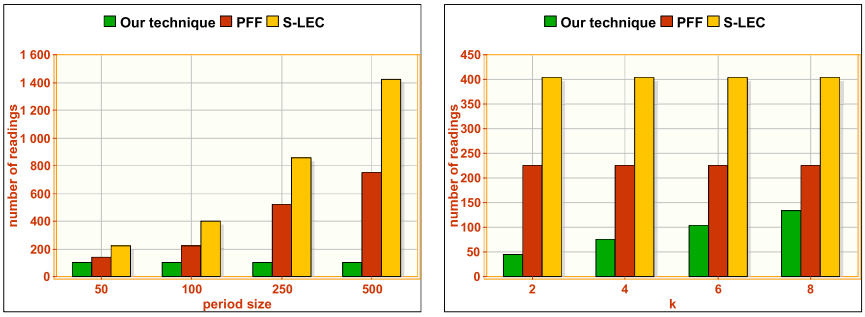


Fig. 22 Raw data vs regenerated data using LSA, $\tau = 250$, $k = 6$.

similar to the temperature condition at the Intel lab and far from the heart rate condition which varies quickly. The obtained results show that our technique reduces up to 86% and 92% of readings sent from each sensor compared to PFF and S-LEC, when varying τ and k.



(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

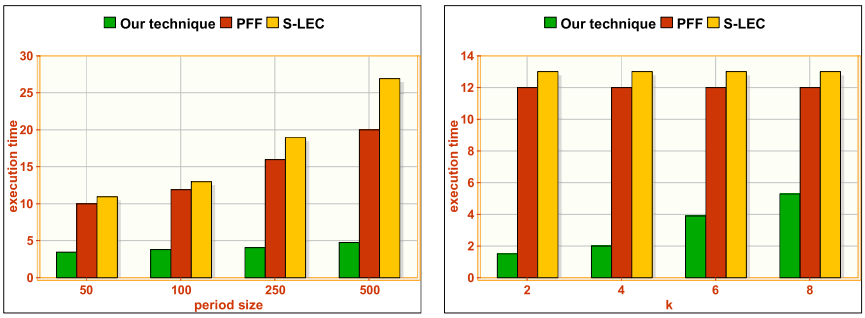
Fig. 23 Data transmission ratio from sensor to CH.

7.5.3 Execution Time at Sensor

Figure 24 shows the execution time required to apply each algorithm at chemical sensor nodes. The obtained results show that our technique accelerates time processing at the sensor up to 4 compared to PFF and up to 6 times compared S-LEC technique.

7.5.4 Variation Between Raw and Regenerated Data at CH

Figure 25 shows the variation between raw coefficient values and values regenerated by LSA at the CH level for the methane gas sensors. The results



(a) LSA sensor degree, $k = 6$ (b) period size, $\tau = 100$

Fig. 24 Execution time required for each algorithm on the sensor.

show that the LSA prediction model allows to regenerate very close coefficient values for each sensor where the error arrives to 0.001 in the worst case. Therefore, LSA model can be considered as an efficient prediction method for both sensors and CHs.

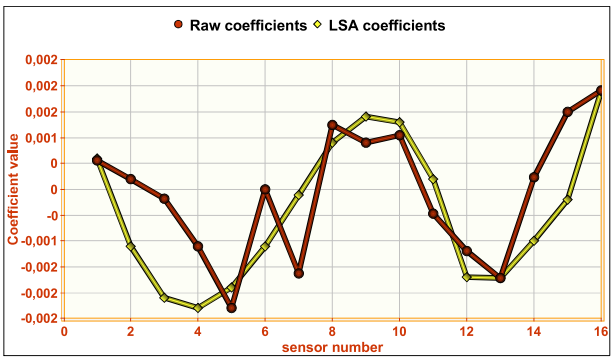
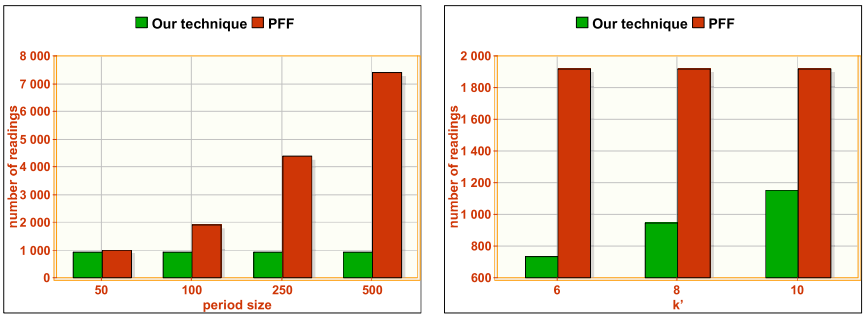


Fig. 25 Raw coefficients vs regenerated LSA coefficients using LSA, $\tau = 250$, $k = 6$, $k' = 8$.

7.5.5 Data Transmission Ratio at CH

Figure 26 shows the number of methane gas readings sent from each CH to the sink after applying our technique and PFF. The obtained results shows that the data transmission at the CH reduced by from 5% to 87% compared to PFF, when varying the period size, LSA degree at sensor and CH. We can also notice that our technique gives better results when the period size increases.



(a) $k = 6$, $k' = 8$

(b) $k = 6, \tau = 100$

(c) $\tau = 100, k' = 8$

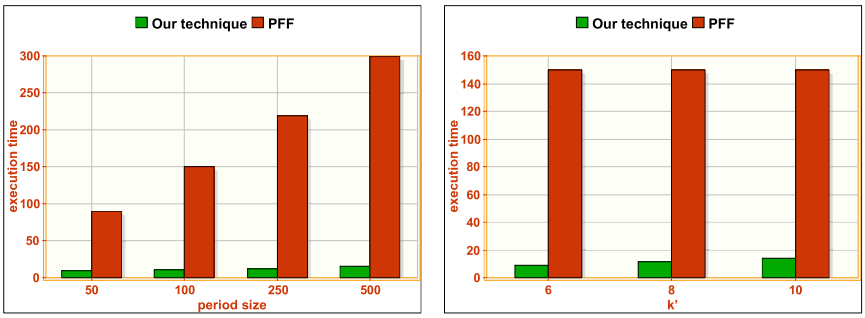
Fig. 26 Data transmission ratio from CH to the sink.

7.5.6 Execution Time at CH

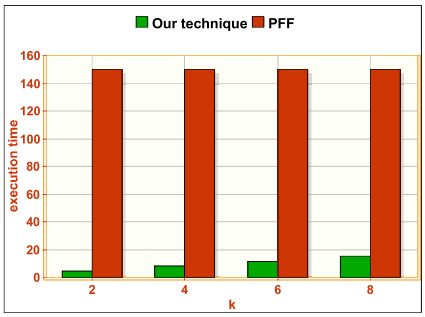
Figure 27 shows the execution time required to apply both our technique and PFF over methane gas readings at the CH. The obtained results that our technique can accelerate the prediction processing time at the CH from 9 to 18 times when varying the period size, from 10 to 16 times when varying k and from 9 to 30 times when varying k'.

8 Conclusion and Future Work

As the number of connected devices will continue to rise every day, the IoT will take more attention from both industries and governments. Thus, data reduction and prediction algorithms will remain at the heart of data management in IoT. In this paper, we have proposed an energy-efficient prediction mechanism dedicated to periodic large-scale sensing-applications. Our prediction model uses the least squares approximation method at sensors and CHs nodes in a cluster-based network architecture. Our model allows each sensor, at the first tier, to send a predictive set of data to the CH, while, at the second tier, it allows CH to send one predictive set for the whole cluster nodes toward the sink.



(a) $k = 6$, $k' = 8$ (b) $k = 6, \tau = 100$



(c) $\tau = 100, k' = 8$

Fig. 27 Execution time required for each algorithm on the CH.

We evaluated our mechanism through extensive simulation on real sensor data collected from various sensor applications (weather, underwater, health, and industrial). Compared to other existing techniques, the results demonstrated the efficiency of our mechanism in terms of optimizing the data regeneration error, the data transmission ratio, and the execution time at both sensors and CH levels (Table 1).

	Data regeneration error		Data transmission ratio		Execution time	
	Sensor	CH	Sensor	CH	Sensor	CH
Weather	0.1	0.008	93%	96%	11 times	37 times
Underwater	0.05	0.09	84%	96%	5 times	98 times
Health	1	0.12	94%	90%	3 times	142 times
ndustrial	20	0.001	92%	87%	6 times	30 times

Table 1 Summarizing of best results obtained with our mechanism.

Several directions can be pass through to improve our mechanism. First, a scheduling strategy can be applied inside each cluster in order to switch correlated sensors into sleep/active modes. Second, we plan to increase the accuracy of data received at the sink by adding some improvements to the

proposed prediction algorithm. Third, we seek to adapt our mechanism to heterogeneous sensor networks where each node can collect data about several conditions.

9 Data Availability Statements

The datasets generated during and/or analysed during the current study are available in the: Intel repository, https://www.kaggle.com/datasets/divyansh22/intel-berkeley-research-lab-sensor-data, ARGO repository, https://argo.ucsd.edu/, and MIMIC repository, https://archive.physionet.org/mimic2/.

10 Funding

The authors have no relevant financial or non-financial interests to disclose.

11 Conflicts of interest/Competing interests

The authors declare that they have no conflict of interest.

12 Code availability

NA.

References

1. H. Harb, A. K. Idrees, A. Jaber, A. Makhoul, O. Zahwe, and M. A. Taam, "Wireless sensor networks: a big data source in internet of things," *International Journal of Sensors Wireless Communications and Control*, vol. 7, no. 2, pp. 93–109, 2017.
2. M. Framingham, "Idc forecasts worldwide technology spending on the internet of things to reach 1.2 trillion in 2022," *International Data Corporation (IDC)*, 2018.
3. H. Harb, H. Baalbaki, C. A. Jaoude, and A. Jaber, "Orchestration-based mechanism for sampling adaptation in sensing-based applications," *IET Smart Cities*, vol. 3, no. 3, pp. 158–170, 2021.
4. H. Harb, A. Mansour, A. Nasser, E. M. Cruz, and I. de la Torre Diez, "A sensor-based data analytics for patient monitoring in connected healthcare applications," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 974–984, 2020.
5. A. El Sayed, H. Harb, M. Ruiz, and L. Velasco, "Zizo: A zoom-in zoom-out mechanism for minimizing redundancy and saving energy in wireless sensor networks," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3452–3462, 2020.
6. G. Saad, H. Harb, A. Abouaissa, L. Idoumghar, and N. Charara, "P2d: An efficient patient-to-doctor framework for real-time health monitoring and decision making," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 14 240–14 252, 2020.
7. N. Merabtine, D. Djenouri, D.-E. Zegour, B. Boumessaidia, and A. Boutahraoui, "Balanced clustering approach with energy prediction and round-time adaptation in wireless sensor networks," *International Journal of Communication Networks and Distributed Systems*, vol. 22, no. 3, pp. 245–274, 2019.

8. R. Ranjan Swain, P. Mohan Khilar, and T. Dash, "Fault diagnosis and its prediction in wireless sensor networks using regressional learning to achieve fault tolerance," *International journal of communication systems*, vol. 31, no. 14, pp. 1–17, 2018.
9. A. Idakwo Monday, I. Umoh, and S. Man-yahaya, "Real time wireless sensor network for environmental data prediction and monitoring," *International Journal of Scientific & Engineering Research*, vol. 8, no. 1, pp. 1522–1529, 2017.
10. G. M. Dias, B. Bellalta, and S. Oechsner, "A survey about prediction-based data reduction in wireless sensor networks," *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, p. 58, 2016.
11. H. Wang, G. Ni, J. Chen, and J. Qu, "Research on rolling bearing state health monitoring and life prediction based on pca and internet of things with multi-sensor," *Measurement*, vol. 157, p. 107657, 2020.
12. R. da Rosa Righi, G. Goldschmidt, R. Kunst, C. Deon, and C. A. da Costa, "Towards combining data prediction and internet of things to manage milk production on dairy cows," *Computers and Electronics in Agriculture*, vol. 169, p. 105156, 2020.
13. M. Hosseinzadeh, J. Koohpayehzadeh, A. O. Bali, P. Asghari, A. Souri, A. Mazaherinezhad, M. Bohlouli, and R. Rawassizadeh, "A diagnostic prediction model for chronic kidney disease in internet of things platform," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16 933–16 950, 2021.
14. N. Li, N. Gebraeel, Y. Lei, X. Fang, X. Cai, and T. Yan, "Remaining useful life prediction based on a multi-sensor data fusion model," *Reliability Engineering & System Safety*, vol. 208, p. 107249, 2021.
15. C. Salim and N. Mitton, "K-predictions based data reduction approach in wsn for smart agriculture," *Computing*, vol. 103, no. 3, pp. 509–532, 2021.
16. 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.
17. A. Russo, F. Verdier, and B. Miramond, "Energy saving in a wireless sensor network by data prediction by using self-organized maps," *Procedia computer science*, vol. 130, pp. 1090–1095, 2018.
18. G. Krishna, S. K. Singh, J. P. Singh, and P. Kumar, "Energy conservation through data prediction in wireless sensor networks," 2018.
19. 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.
20. S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, "Time series data analysis of wireless sensor network measurements of temperature," *Sensors*, vol. 17, no. 6, p. 1221, 2017.
21. J. Karjee and M. Kleinsteuber, "Data estimation with predictive switching mechanism in wireless sensor networks," *International Journal of Sensor Networks*, vol. 25, no. 3, pp. 184–197, 2017.
22. L. C. Tagliabue, F. R. Cecconi, S. Rinaldi, and A. L. C. Ciribini, "Data driven indoor air quality prediction in educational facilities based on iot network," *Energy and Buildings*, vol. 236, p. 110782, 2021.
23. J. M. Bahi, A. Makhoul, and M. Medlej, "A two tiers data aggregation scheme for periodic sensor networks." *Ad Hoc & Sensor Wireless Networks*, vol. 21, no. 1-2, pp. 77–100, 2014.
24. Q. Xu, R. Akhtar, X. Zhang, and C. Wang, "Cluster-based arithmetic coding for data provenance compression in wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.
25. P. Zeng, B. Pan, K.-K. R. Choo, and H. Liu, "Mmda: Multidimensional and multidirectional data aggregation for edge computing-enhanced iot," *Journal of Systems Architecture*, vol. 106, p. 101713, 2020.
26. J. Zhang, Z. Lin, P.-W. Tsai, and L. Xu, "Entropy-driven data aggregation method for energy-efficient wireless sensor networks," *Information Fusion*, vol. 56, pp. 103–113, 2020.
27. Q. Mamun, "A qualitative comparison of different logical topologies for wireless sensor networks," *Sensors*, vol. 12, no. 11, pp. 14 887–14 913, 2012.

28. T. M. Behera, S. K. Mohapatra, U. C. Samal, M. S. Khan, M. Daneshmand, and A. H. Gandomi, "Residual energy based cluster-head selection in wsns for iot application," *IEEE Internet of Things Journal*, 2019.
29. S. Biswas, J. Saha, T. Nag, C. Chowdhury, and S. Neogy, "A novel cluster head selection algorithm for energy-efficient routing in wireless sensor network," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. IEEE, 2016, pp. 588–593.
30. R. R. Priyadarshini and N. Sivakumar, "Cluster head selection based on minimum connected dominating set and bi-partite inspired methodology for energy conservation in wsns," *Journal of King Saud University-Computer and Information Sciences*, 2018.
31. Y. K. Yousif, R. Badlishah, N. Yaakob, and A. Amir, "An energy efficient and load balancing clustering scheme for wireless sensor network (wsn) based on distributed approach," in *Journal of Physics: Conference Series*, vol. 1019, no. 1. IOP Publishing, 2018, p. 012007.
32. S. Kang, "Energy optimization in cluster-based routing protocols for large-area wireless sensor networks," *Symmetry*, vol. 11, no. 1, p. 37, 2019.
33. G. P. Gupta, "Improved cuckoo search-based clustering protocol for wireless sensor networks," *Procedia Computer Science*, vol. 125, pp. 234–240, 2018.
34. A. Rais, K. Bouragba, and M. Ouzzif, "Routing and clustering of sensor nodes in the honeycomb architecture," *Journal of Computer Networks and Communications*, vol. 2019, 2019.
35. H. Harb, A. Makhoul, R. Tawil, and A. Jaber, "A suffix-based enhanced technique for data aggregation in periodic sensor networks," in *2014 international wireless communications and mobile computing conference (IWCMC)*. IEEE, 2014, pp. 494–499.
36. B. Raj, I. Ahmedy, M. Y. I. Idris *et al.*, "A survey on cluster head selection and cluster formation methods in wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.
37. O. Bretscher, *Linear algebra with applications*. NJ: Prentice Hall, 1995.
38. S. Madden, "Intel berkeley research lab data," 2004.
39. Argo, "Argo project," 2019.
40. PhysioNet, "The mimic and mimic ii databases on physionet," 2000.
41. Sensorscope, "Audiovisual communications lcav," 2007.
42. Y. Liang and Y. Li, "An efficient and robust data compression algorithm in wireless sensor networks," *IEEE Communications Letters*, vol. 18, no. 3, pp. 439–442, 2014.