Science & Technology Development Journal 2023, 26(4):3132-3143

Open Access Full Text Article

Engineering and Technology: Research Article

Sensor Rotational Measurement with Virtual Sensors Applying Pearson Generative Adversarial Imputation Nets

Nguyen Thanh Quan[1,2,*], Nguyen Quang Hung[1,2], Nam Thoai[1,2]

ABSTRACT

Recent advances in sensor technology have increased the ability of humans to measure a wide range of phenomena and events. Nevertheless, in some cases, due to a variety of limitations, only a few sensors can be deployed at a given site. Consequently, setting up enough sensors at the right places to provide uniform monitoring can be difficult. In addition, virtual sensing, which is a set of strategies for replacing a portion of physical sensors with virtual sensors, has recently been developed. Therefore, this work leverages the imputation capability of PGAIN-VS to develop a black-box data-driven virtual sensing approach named sensor rotational measurement for the purpose of reducing the number of physical devices to be used in reality while still ensuring monitoring accuracy. The approach takes advantage of the PGAIN-VS and Borda voting methods to determine the subset of real sensors that can take turns observing information within an interval of time. The approach is seen as a black-box objective optimization problem with constraints that is solved by the OpenBox tool. We evaluated our method on several real-world datasets and achieved promising results, with the overall number of physical sensors reduced by up to 20%.

Key words: IoT, machine learning, virtual sensing, missing data imputation, optimal sensor placement

[1]*High Performance Computing Laboratory, Faculty of Computer Science and Engineering, Advanced Institute of Interdisciplinary Science and Technology, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Viet Nam;*

[2]*Viet Nam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Viet Nam*

Correspondence

Nguyen Thanh Quan, High Performance Computing Laboratory, Faculty of Computer Science and Engineering, Advanced Institute of Interdisciplinary Science and Technology, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Viet Nam;

Viet Nam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Viet Nam

Email: ntquan.sdh20@hcmut.edu.vn

History

- Received: 2023-04-27
- Accepted: 2023-11-14
- Published Online: 2023-12-31

DOI :

https://doi.org/10.32508/stdj.v26i4.4080

INTRODUCTION

Currently, the terms Internet of Things (IoT) and edge computing have appeared in most areas and provided many benefits for businesses as well as people's lives because they have created many smart industries, namely, smart villages, smart homes, smart farming, and more. However, why is this the case? The answer is because the utilization of numerous different physical sensor types in real-time data measurement and analytics has provided a solution. Several market researchers [1] estimate that there are more than 20 billion connected devices at present, and by 2025, there will be more than 70 billion IoT devices worldwide.

In general, sensors can be divided into two main categories: (1) physical sensors that can observe external phenomena on their own and (2) virtual sensors that are designed for processing data from various observed sources to determine values approximated by proper models [2]. To estimate the difficulty or expense of measuring quantity, virtual sensing usually relies on real physical sensor data collected and the use of appropriate algorithms to integrate them. There may be restrictions on the placement of data-collected devices, and the cause may be either natural or economical. As a result, recent efforts to construct and apply virtual sensing solutions that make use of indirect measurements [3] have proven effective.

This paper proposes a novel data-driven virtual sensing approach, named sensor rotational measurement (SRM), which leverages the missing-data imputation strength of PGAIN-VS [4] and the impact weighting of the Borda method [5] to find the best subset from an original set of physical sensors.

Basically, PGAIN-VS is a deep learning virtual sensor technique that can be used to generate data in place of missing or incomplete data. It is based on generative adversarial networks (GANs) [6] and inspired by generative adversarial imputation networks (GAINs) [7]. These networks consist of two neural networks: a generator that produces new data and a discriminator that distinguishes between observed and imputed data. In PGAIN-VS, the generator is trained to generate data that are consistent with the observed data based on the correlations among physical devices, while the discriminator is trained to distinguish between real and generated data. To apply the PGAIN-VS virtual sensors to the SRM, we first collected the data from the physical sensors. We then trained PGAIN-VS using these data to generate imputed data that followed the distribution of the observed data. These generated data can be used to estimate the missing data from physical sensors.

The Borda voting method involves assigning points to different options based on their rank. In Borda voting,

Cite this article : Quan N T, Hung N Q, Thoai N. **Sensor Rotational Measurement with Virtual Sensors Applying Pearson Generative Adversarial Imputation Nets**. *Sci. Tech. Dev. J.* 2023; 26(4):3132-3143.

VNUHCM PRESS

each voter ranks his or her choices in order of preference, and the choice with the highest rank receives the most points. The points are then summed, and the choice with the highest total number of points is declared the winner. The preference in SRM is the prediction error when particular sensors are present in a group of predictors to estimate missing or incomplete data. The smaller the error is, the greater the chance that the sensors in the predictor group can be voted on.

Combining the PGAIN-VS and Borda voting method is a novel useful approach for selecting a subset of real sensors that can take turns observing information within an interval of time. After that, PGAIN-VS itself continues to be applied to impute the missing data for the positions where physical sensors are no longer present. To validate the accuracy of the virtual sensor measurements, we compare them to the actual measurements obtained from the physical sensors through the root mean square error (RMSE).

The SRM is considered a black-box objective optimization problem with constraints and is solved by the OpenBox tool[8], which is based on a Bayesian optimization algorithm. The proposed approach is evaluated on real-world energy, temperature and vehicle speed datasets, and the results demonstrate that the SRM is able to achieve high accuracy in predicting the target variable and outperforms state-of-the-art virtual sensing approaches. Additionally, an SRM is able to identify the most informative physical sensors for capturing the underlying dynamics of the system, which can help reduce the cost of hardware installation and maintenance. Overall, the proposed approach provides a promising solution for virtual sensing in various industrial and engineering applications in situations where physical sensors may be unavailable or unreliable.

This work has two important contributions:

- Practical contribution: This paper provides a useful virtual sensing solution for replacing physical sensors with virtual sensors when there may be limitations from an environmental or economical nature.
- Academic contribution: This paper introduces a new dynamic measurement approach for collecting data via sensors. In addition, this work proves the ability of the proposed method to address the multiple missing-data dimensions problem via the imputation of PGAIN-VS.

This paper has six sections. The related work is discussed in Section 2. In Section 3, the materials, methods, and algorithms used to address the SRM are presented. Section 4 describes the experiments and results. Additionally, we discuss the experimental results further in Section 5. Section 6 discusses the drawbacks of our work and proposes some future directions to improve these issues.

RELATED WORK

Virtual sensing has received a great deal of attention in recent years in a variety of fields, including robotics, automation, transportation, agriculture, etc. Virtual sensing modeling techniques can be categorized into three groups based on the mechanism of creation and data consumption: (1) the first principle, where virtual sensors are mathematically created using the foundational laws of physics and extensive domain knowledge; (2) the black box, where empirical relations and correlation are present in the data are used to construct virtual sensors; and (3) the gray box, which is a combination of the two.

Virtual sensors were developed in[9] for an optimization problem. By selecting the most valuable subset of sensors to keep in a particular indoor environment and replacing unneeded sensors with virtual sensors, the authors hoped to reduce the number of actual sensors used in an indoor setting. Regarding the work in[10], for the example of heat exchanger systems, the authors presented an ideal sensor selection and fusion method using the minimum redundancy maximum relevance (mRMR) method. Another sensor selection method for the best fault detection and isolation (FDI) tests in complicated systems was also discussed in the study in[11].

Regarding mobile sensing techniques, a method was presented in[12] based on reverse combinatorial auctions, which required the concept of social welfare, information quality, and the cost incurred by each individual user to provide an observation. At first glance, our proposal looks similar to this mobile technique, as physical sensors must be moved among some locations, but the positions in our solution are static. Instead, in sensor scheduling problems, one or more sensors must be chosen at each stage, as stated in[13], was considered our motivation for this research. As another instance of a sensor selection technique, the authors in[14] used a weighted function of the error covariances associated with the state estimates to trim the search tree of all potential sensor schedules to achieve their objective. The authors in[15] proposed an algorithm to minimize the expected error covariance for stochastic sensor selection by using Kalman

filters with an underlying hidden Markov model. In addition, modern techniques in machine learning and sparse sampling were leveraged to design optimal sensor locations for signal reconstruction, as presented in [16].

Sensor selection approaches are thought to be important in the field of wireless sensor networks (WSNs). WSNs can be briefly described as networks connected to sensors and extending across large geographical areas to collect and transmit data. To illustrate this point, the research in [17] showed that by choosing an appropriate event-triggering threshold, a sensor data scheduler for linear systems was developed to achieve the ideal balance between the communication rate and estimation quality. A new method for selecting suitable sensors for WSNs was developed in [18], in which Kalman filters, multiple cost functions, network condition sets and an assumed time horizon were determined in advance. Additionally, in regard to WSNs, Kalman filters and interacting multiple model (IMM) filters were applied in [19] to minimize energy usage. Furthermore, the researchers in [20] proposed a solution to address the limitations of WSNs, including memory, energy and processing capabilities, by discrete cosine transform (DCT) and discrete wavelet transform (DWT) image compression techniques because they can be used on sensor nodes and allow for an effective trade-off between the compression ratio and energy consumption. In [21], a framework for virtual sensing with optimal sensor placement (OSP) was introduced. The framework is based on information and utility theory, employs the model expansion technique, and accounts for uncertainty. The framework was created with the intention of handling virtual sensing with output-only vibration measurements.

In contrast to the aforementioned solutions, the SRM, which is a novel combination variant of sensor selection and optimal sensor placement, simultaneously answers two major questions: (1) how many sensors should be used and (2) how can missing data be handled at positions where sensors are not present?

MATERIALS AND METHODS

Suppose the space where sensors are placed in an area creates a wireless sensor network (WSN) with a d-dimensional space $\mathbf{S} = S_1$ x ... x S_d. Let $\mathbf{Y} = (Y_1 ..., Y_d)$ be an observed variable of sensor (continuous due to environmental sensors) having values in $\mathbf{S}$. P($\mathbf{Y}$) is the denotation for the distribution of the variable $\mathbf{Y}$.

As specified in the PGAIN-VS architecture, there is a random variable with values of $\{0,1\}^d$, $\mathbf{M} = (M_1 ...,M_d)$, derived from **Y,** which is the data vector observed by sensor , and **M** is the mask vector.

We define a new space $(\widetilde{\mathbf{S}}_1) = S_i \cup \{*\}$ where * is simply a point that does not belong to any **S** *i*. In other words, it is a value at the period of time that a specific location is not equipped with sensor S, with i ∈ {1..., *d*}. Let $\widetilde{\mathbf{S}} = \widetilde{S}_1 \times ... \times \widetilde{S}_d$. $\widetilde{\mathbf{Y}} = \left(\widetilde{Y}_1, ..., \widetilde{Y}_d\right) \in \widetilde{\mathbf{S}}$ is a new random variable and is defined in the following way Eq. (1):

$$\widetilde{Y}_i = \begin{cases} Y_i & if\ observed \\ * & if\ unobserved \end{cases} \quad (1)$$

Obviously, **M** can be recovered from $\widetilde{\mathbf{Y}}$.

The generator, **G**, in PGAIN-VS takes the three variables $\widetilde{\mathbf{Y}}$ with the Pearson correlation (Pc) arrangement, **M** and noise **N** as inputs and produces the prediction $\bar{\mathbf{Y}}$.

Basically, the random variables are defined $\bar{\mathbf{Y}}$ and $\widehat{\mathbf{Y}}$ as shown in Eq. (2) and Eq. (3):

$$\bar{\mathbf{Y}} = G\left(\widetilde{\mathbf{Y}} | \mathbf{Pc}, \mathbf{M}, (1-\mathbf{M}) \odot \mathbf{N}\right) \quad (2)$$

$$\widehat{\mathbf{Y}} = \mathbf{M} \odot \mathbf{N} + (1-\mathbf{M}) \odot \bar{\mathbf{Y}} \quad (3)$$

where $\odot$ is the elementwise multiplication. $\bar{\mathbf{Y}}$ is the imputed value vector for which the data points are predicted by PGAIN-VS. $\widehat{\mathbf{Y}}$ is the complete data vector obtained by combining the observed data vector in a period of time extracted from the combination of complete and incomplete observation $\widetilde{\mathbf{Y}}$ and replacing *with the corresponding value of $\bar{\mathbf{Y}}$. The noise passed into the generator is $(1 - \mathbf{M}) \odot$ because the target distribution is $P(\mathbf{Y}|\widetilde{\mathbf{Y}})$

The discriminator **D** in PGAIN-VS is considered an adversary for training G to determine the observed or predicted components. **D**: Y → $[0,1]^d$ with the component D($\widehat{\mathbf{Y}}$) at the *i-th* position following the probability that the value of $\widehat{\mathbf{Y}}$ at the *i-th* position is observed by sensor .

PGAIN-VS still uses the "hint" defined in [7] in the process of missing data imputation. $\widehat{\mathbf{Y}}$ and **H** are accepted as the inputs passed to **D**. **H** is formed as follows:

$$\mathbf{H} = \mathbf{U} \odot \mathbf{M} + 0.5 \odot (1 - \mathbf{U}) \quad (4)$$

where $\mathbf{U} \in \{0,1\}^d$ is defined by uniformly sampling k from {1, 2, ..., d} at random and then setting Eq. (5):

$$\mathbf{U}_j = \begin{cases} 1\ if\ j \neq k \\ 0\ otherwise \end{cases}$$

In PGAIN-VS, **D** is trained to increase the likelihood that it will accurately predict **M**, and **G** is trained to

decrease the likelihood that **D** will predict **M**. The general objective function and loss function are briefly described as Eq. (6) and Eq. (7) below:

$$L(D,G) = \begin{cases} min(G) \\ max(D) \end{cases} \tag{6}$$

$$L(D,G) = E_{\widehat{Y}|Pc,M,H}[M^T \log D \left(\widehat{\mathrm{Y}}|\mathbf{Pc},\mathbf{M},\mathbf{H}\right) + (1-\mathbf{M})^T \log\left(1 - D\left(\widehat{\mathrm{Y}}|\mathbf{Pc},\mathbf{H}\right)\right) \tag{7}$$

In general, the PGAIN-VS algorithm can be expressed in detail by the following pseudocode (pseudocode for the PGAIN-VS algorithm):

Algorithm. Pseudocode of the PGAIN-VS algorithm

Pearson correlation and sensor arrangement calculations

Input: Missing data of sensor Sf and data frame of other sensors Sc

Output: Numpy 2-dimensional array with Pearson arrangement

for $i = 0.n_{Sc}$ **do**

$p_i \leftarrow Pc(Sf, \mathrm{Sc}_i)$

end for

The sc list was rearranged in descending order based on Pearson's

Convert the Sc data frame to a 2-d numpy array

Begin training of PGAIN-VS

Input: D is the size of batch n_D, G is the size of batch n_G. **Output**: Imputed data.

Stochastic gradient descent (SGD) is applied.

while until the training loss converges:

(A) **Optimize D**

Generate n_D samples through $\{\widetilde{y}|pc,m\}$, noise and hint

for $i = range(n_D)$:

$y_i \leftarrow G(\widetilde{y_i}|pc,m_i,n_i)$

$\widehat{y_i} \leftarrow m_i \odot \widetilde{y_i} + \bar{y}_i \odot (1-m_i)$

$h_i \leftarrow u_i \odot m_i + 0.5 \odot (1-u_i)$

end for

Use SGD to update D.

(B) **Optimize G**

Generate n_D in the same way as above

for $i = range(n_G)$:

$h_i \leftarrow u_i \odot m_i + 0.5 \odot (1-u_i)$

end for

Use SGD to update G.

end while

a. Sensor rotational measurement description

In theory, there is always a correlation among collected data, which indicates relationships and similarities among spatial positions. Thus, we define **C** = (C_1...,C_d) as a correlation variable of sensor S_i with the remaining values in the range $[0,1]^d$ and then apply PGAIN-VS to the dataset with its features arranged by Pearson correlation C to obtain imputed values $\bar{\mathbf{Y}}$. R = (R_{S1} ...,R_{Sd}) is defined as a ranking variable calculated with the Borda voting method, and the weight of sensor S_i is determined by R_i. We define a formula to calculate the weight of a single sensor S_i as shown in Eq. (8):

$$w_i = 1 - \left(\frac{\varepsilon}{max_{i\in\{1,\ldots,d\}}}\varepsilon_i\right) \tag{8}$$

where υ_i is determined by adding the 95th percentile of the absolute errors. Then, the vote is defined as Eq. (9):

$$R_{S_i} = \sum_{i=0,i\neq j}^{d} [n_{sensors} - pos_i(j)]\, w_i \tag{9}$$

where $pos_i(j)$ represents the position of sensor j, the total number of sensors used is $n_{sensors} = d$, and w_i is the weight calculated.

The whole set of rotational measurement data can be described as follows (Eq. (10)):

$$Mat = \begin{bmatrix} y_1^{R_{S_1}} & * & y_1^{R_{S_3}} & * & \cdots & y_1^{R_{S_d}} \\ y_2^{R_{S_1}} & * & y_2^{R_{S_3}} & * & \cdots & y_2^{R_{S_d}} \\ * & y_3^{R_{S_2}} & * & y_3^{R_{S_4}} & \cdots & y_3^{R_{S_d}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ * & y_n^{R_{S_2}} & * & y_n^{R_{S_4}} & \cdots & y_n^{R_{S_d}} \end{bmatrix} \tag{10}$$

where $y_i^{R_{S_i}}$ are normal observed values with the *i-th* rank arranged in descending order from left to right and * are unobserved values during the time when sensor was not present. The number of * on each feature and the number of features having * are determined by solving the optimization mentioned in Eq. (11). Ambitiously, we want to have more * appearing in the dataset, but the reliability of our solution is still preserved.

b. Objective

SRM is considered a black-box optimization (BBO) with constraints, as PGAIN-VS is a machine learning model; thus, there is no analytical form for the objective function. PGAIN-VS is the vector-valued black-box function for the SRM BBO described as $f(x)$: X $\rightarrow$ R^d, where X is the search space of interest and has

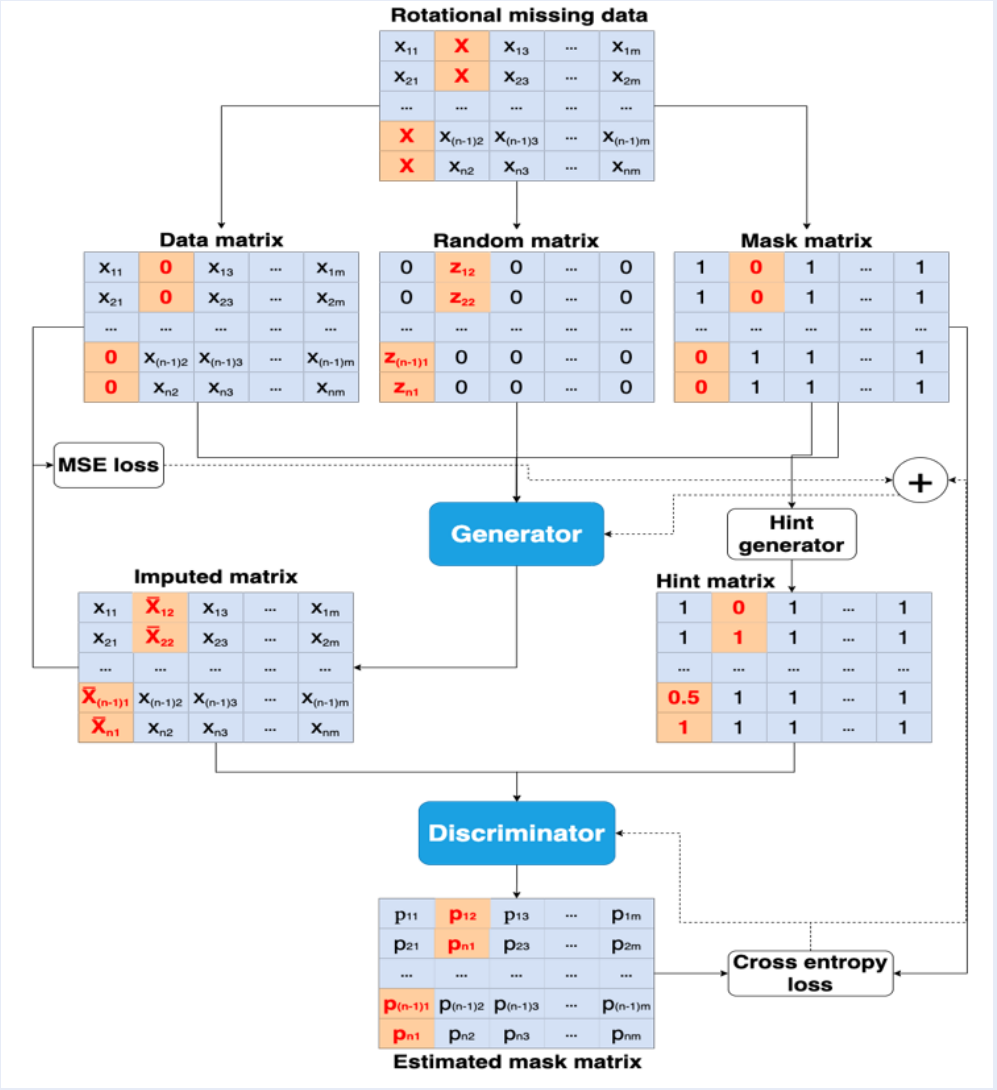


Figure 1: Sensor rotational measurement architecture

the same d-dimension as S defined earlier. Identifying the set of Pareto optimal solutions is the main goal presented in Eq. (11):

$$P^* = f(x)\, s.t. \nexists x' \in X : f(x') < f(x) \tag{11}$$

such that every advancement in one objective implies a decline in another. We calculate the finite Pareto set $\mathbf{P}$ from the observed data $\{(x_i)\}_{i=1}^{n}$ to estimate $\mathbf{P}^*$.

In this paper, the PGAIN-VS virtual sensor inspired by GAIN [7] was used as the main mediator to impute missing values. The Borda count[22] method was applied to determine the weight of the sensors before solving the SRM BBO. The SRM BBO is considered the objective optimization method because the minimum RMSE is set as the main objective function, the maximum number of reduced physical sensors is set, and the maximum measurement interval is set. Mathematically, the equation of the SRM can be written as follows:

$$O_{SRM} = \begin{cases} f_e^{min}(x)\, g_n^{max}(x)\, h_i^{max}(x) : x \in F \\ f_e(x) \le c_1 \\ 0 < g_n(x) \le c_2 \\ 0 < h_i(x) \le c_3 \end{cases} \tag{12}$$

where x is the solution, F is the feasible set, f_e is the minimum RMSE objective function, the two others are g_n, the maximum number of physical sensors possibly involving the SRM, and h_i, the maximum observation interval. c_1, c_2 , *and* c_3 are the threshold constraints of the functions, and *min/max* is the combined object operation.

c. Sensor rotational measurement algorithm

The following approach, described in pseudocode, was used to evaluate the ranking supported by Pearson correlation using PGAIN-VS on the 20% of test data that were missing to automatically choose a selection of sensors to be utilized as predictors.

Algorithm. Pseudocode for the sensor ranking

Input: Data of the whole sensors in the list *Sl*

Output: The rank of the sensors in descending order

foreach *sen* in *Sl* do i i← 0

for $k = 0.n_{Sl} - 1$ do

Discarded sensors: $S_{discarded} \leftarrow Sl[0: k]$

Remaining sensors: S_{remain}

$remain \leftarrow set(Sl) - set(Sdiscarded) - set([sen])$

$error_i \leftarrow$ PGAIN-VS(S_i, S_{remain})

Empirical cumulative distribution function at the 95% percentile:

$ecdf\_95_i \leftarrow$ ecdf(percentile(95, $error_i$))

end for

i++

end foreach

for $i = 0..npredicted\_sensors$ **do**

Weight: $w_i \leftarrow sum(ecdf\ 95, groupby(predicted\_sensor))$

Borda vote for sensors: $b_i \leftarrow Borda(n_{Sremain}, w_i)$

end for

As previously noted, the SRM is regarded as a black-box optimization problem; hence, OpenBox, an open-source and general-purpose BBO service, was utilized to resolve this problem. The authors claim that OpenBox implements a wide variety of optimization algorithms to achieve excellent performance in diverse BBO issues and that it is able to select the appropriate algorithms and settings based on the parameters of the incoming assignment. Because it can handle situations where the input space has conditions, there are more parameters in the input space, or there are more trials than hundreds of times, we selected the probabilistic random forest (PRF)[23] to solve the problem in this study.

The top influencing sensors can be selected as candidates for applying SRMs to the data. By being arranged very close to one another, the entire data distribution of datasets can be learned and preserved better. Based on this fact, the algorithm for finding the expected results can be defined by the following SRM pseudocode.

Algorithm. Pseudocode of SRM

Input: Data of the whole sensors, *Sl,* arranged by the Borda ranking

Output: RMSE, number of SRM sensors and interval

Definition of the configuration space

$d \leftarrow len(SI)$

$srm\_num \leftarrow 1..(d/2)$

$miss\_rate \leftarrow 0.01..0.2$

Add the *srm_num* and *miss_rate* variables to the OpenBox space *cs*

Definition of the RMSE objective function with configuration space cs

(*srm_num*, *miss_rate*) ← *cs.get*()

$\widetilde{y} \leftarrow create\_rotation\_data(srm\_num, miss\_rate)$

$\bar{y} \leftarrow$ PGAIN VS ($\widehat{y}$)

$rmse \leftarrow rms\_loss\ \widetilde{y}$

return *objs*[*rmse*]

Run OpenBox with the objective

result ← *optimizer.run*()

Calculation of results

Min RMSE value: min(results.get(*rmse*))

Number of locations: result.get(*srm_num*)

Interval: result.get(*miss_rate*) x $interval_2$ *adjacent_observation*

RESULTS

To evaluate PGAIN-VS, we conducted experiments on the Solar Power (21 sensors)[24,25], Traffic (207 detectors; 50 random sensors were selected)[26] and Rasipihat temperature (12 sensors; 50000 random samples were extracted)[9] datasets, which were also widely used in other works. The details of the datasets are described in **Table 1**.

The answers for the SRM problem are given in **Table 2**. We reported the RMSE and R^2values to show the accuracy and reliability of our approach when used to predict missing data with two parameters, the number of reduced sensors and the measurement interval found by OpenBox. We set 100 trials as the maximum number of runs, and the time limit per trial was 180 seconds. A probabilistic random forest was used as the surrogate model for OpenBox to solve the SRM optimization problem.

We extracted the top three feasible results for each experiment corresponding to each dataset with different values of RMSE and R^2, but they were still guaranteed to be within an ideal boundary of approximately 0.10. In addition, to highlight the effectiveness of arranging sensors based on Borda counting, we listed both the ranking and nonranking results for visual comparison. The output shows that the RMSE and R^2obtained using the Borda voting arrangement tend to be significantly better. More importantly, the number of sensors can be reduced, and the time required for a rotation turn seems to increase. This can be clearly observed in the Solar dataset when there is a large distance between the minimum value and the maximum

Table 1: Characteristics of the datasets

Parameters	Solar	Raspihat	Traffic
Samples	24000	50000	34272
Mean	9601	30.2	53.0
STD	19789	2.0	20.1
Min	0	11.9	0
Max	307891	44.4	70
25%	0	29.5	51.3
50%	1393	30.1	61.6
75%	10170	30.8	65.6

value, as well as for the other parameters described in **Table 1**. For the two remaining datasets, especially for Raspihat, we cannot see equivalence because of minor differences in the observed data among positions, but generally, the expectation was still preserved.

Instead of providing the best output for the SRM optimization problem, our approach also yields a few possible results, as shown in **Table 2**, from which a suitable choice can be selected depending on the needs of the real circumstances. For example, in the case of the Solar dataset, removing fewer than **4** sensors may be a better selection, but **6** sensors with an RMSE of **0.077** and an R^2 of **0.829** should be selected when considering economic considerations (up to an approximately **28%** reduction). Similarly, a list of possible combinations was given with up to **5** sensors removed (approximately **45%** reduction) for the Raspihat dataset and up to **6** sensors (approximately **12%** reduction) for the Traffic dataset. Therefore, depending upon the decision to implement a trade-off strategy, the greater the prediction error accepted is, the more sensors can be saved.

Figures 2, 3 and 4 show images of the imputed and actual data of the sensors applying the SRM after solving the BOO problem. We can easily realize that the distance between the predicted and observed data is not too large. The imputed data still assure the overall data distribution of the datasets, so the SRM virtual sensing solution is potentially able to produce accurate data. Obviously, the results found in OpenBox for the optimization problem are reliable. We chose the results of each dataset reported in **Table 2** with all the sensors involving the SRM to construct the performance graph. Three sensors with imputed data were selected to illustrate the accuracy of the data.

DISCUSSION

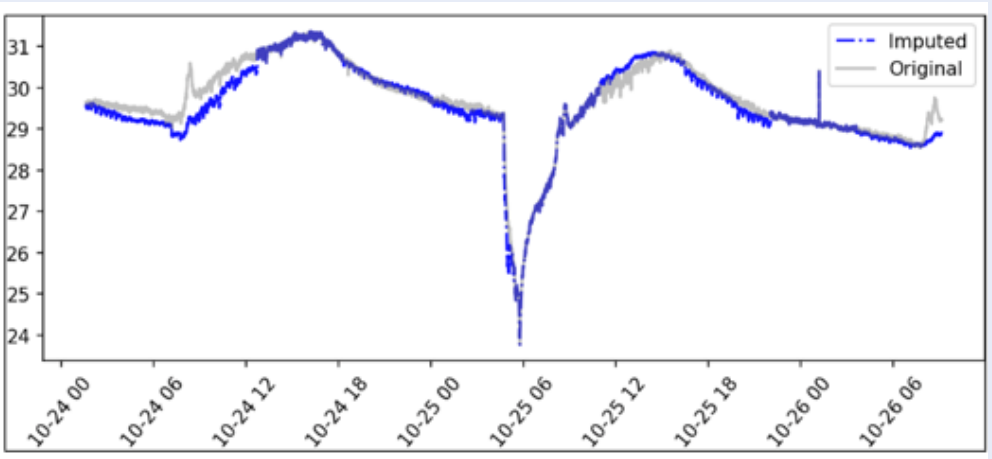
The promising results can be explained by the fact that GAN-based approaches are machine learning models in which two neural networks compete with each other to increase the accuracy of their predictions. Hence, arranging the most impactful sensors adjacent helps the data distribution of the whole dataset be learned and preserved better. In addition, as designed, OpenBox is able to recognize the distribution of datasets after several iterations, so it tends to converge at a point; then, the minima of the RMSE are found with different adjustments of the duration and the number of physical sensors to be cut off.

The experimental results affirm the potential benefit for businesses when the number of sensors used in reality can be lower, so the deployment and management costs are surely lower, and the productivity is greater. In addition, the ability of the PGAIN-VS to impute multiple missing-data dimensions was transparently demonstrated through the SRM.

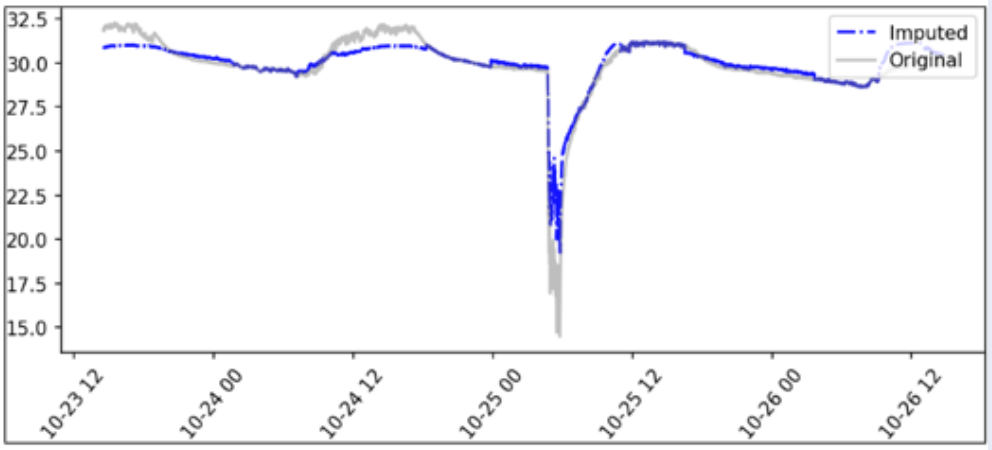
More importantly, a new introduction to rotational measurements may lead to more and more ideas for dynamic observation in the future. Nevertheless, our approach is expected to work well with continuous environmental data, so it requires measurements taken on an interval scale. In contrast, the accuracy of these methods may not be guaranteed.

CONCLUSIONS AND FUTURE WORK

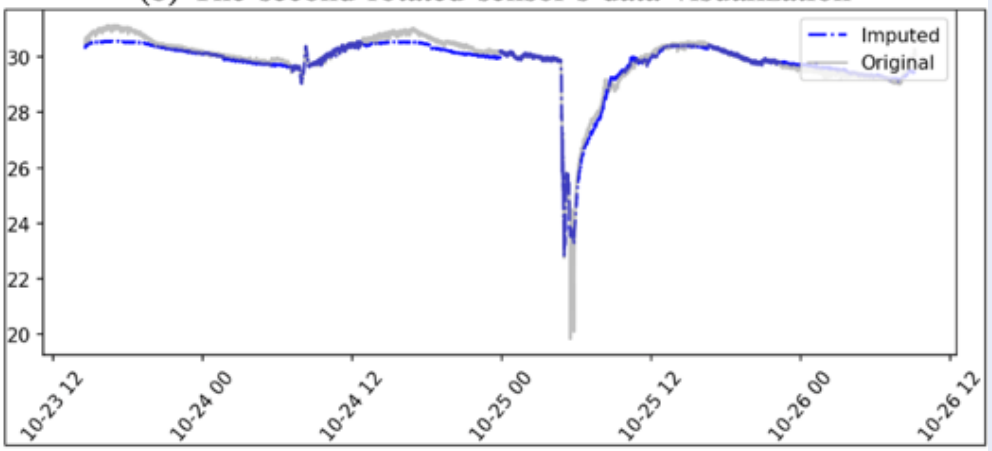
In this paper, a novel dynamic sensor rotation measurement approach, SRM, which takes advantage of PGAIN-VS to determine the best subset of physical sensors, was introduced. Various experiments with real-world datasets show that the SRM is capable of finding and replacing real sensors with virtual ones when limitations may arise from its economical or natural nature. The SRM allows dynamic data observation when physical sensors can be moved around positions to perform their jobs. More importantly, SRM provides businesses with a financially beneficial



(a) The first rotated sensor's data visualization

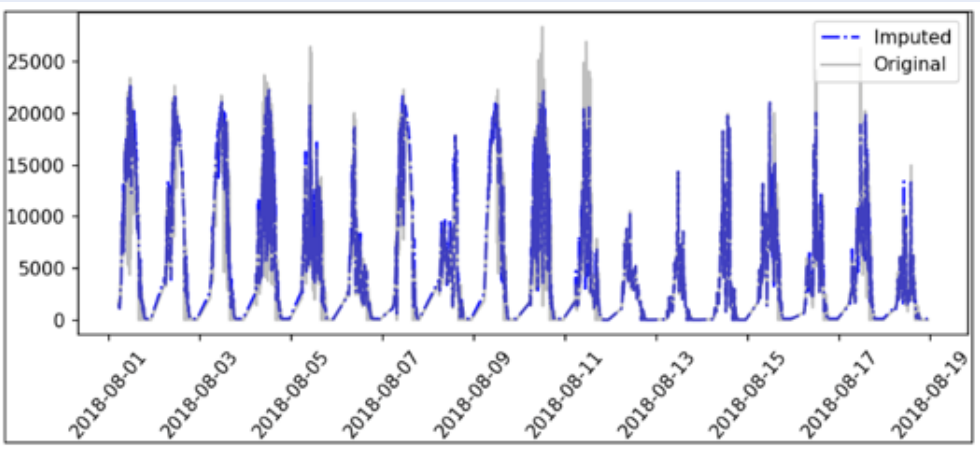


(b) The second rotated sensor's data visualization

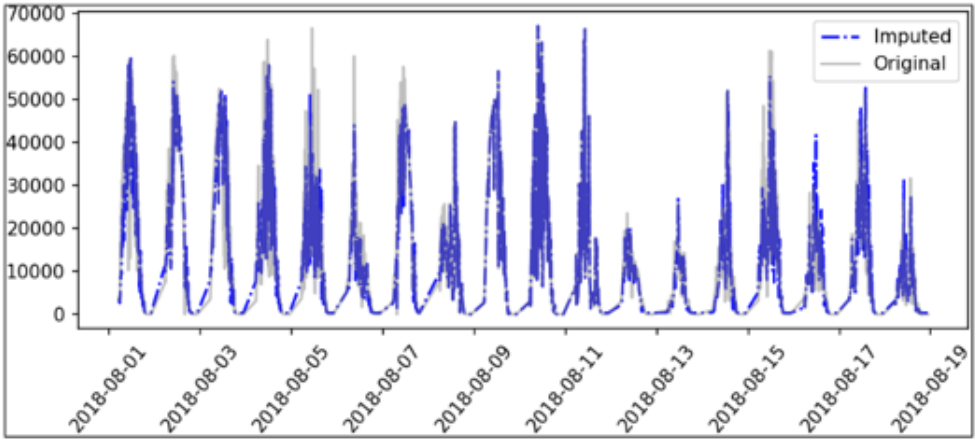


(c) The third rotated sensor's data visualization

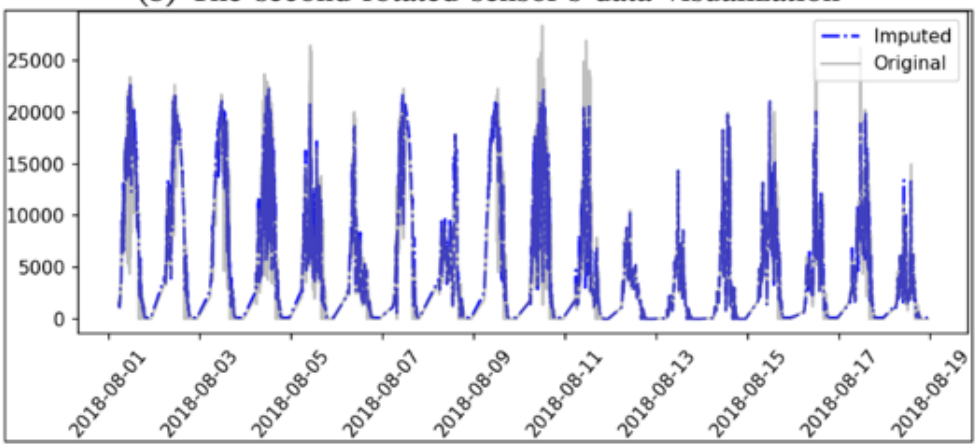
Figure 2: **Performance on the Raspihat dataset for three sensors**; RMSE = 0.032.



(a) The first rotated sensor's data visualization



(b) The second rotated sensor's data visualization



(c) The third rotated sensor's data visualization

Figure 3: **Performance on the Solar Power dataset for three sensors**; RMSE = 0.077

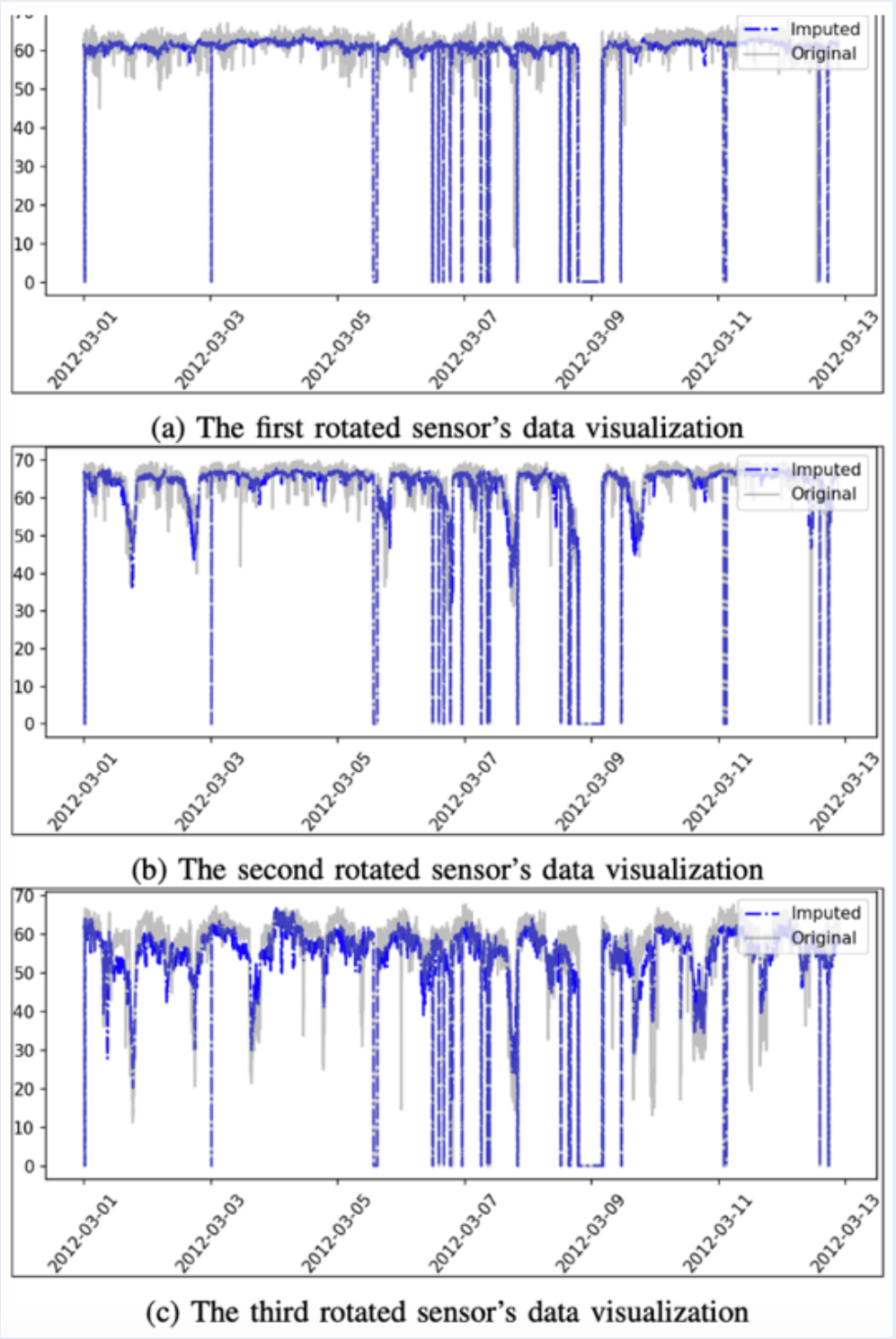


Figure 4: **Performance on the traffic dataset for three sensors**; RMSE = 0.096

trade-off strategy, as it helps reduce the cost of equipment and the deployment and management of a number of physical sensors to be installed in a spatial area. Finally, the SRM proves the ability of PGAIN-VS to impute multiple missing-data dimensions.

Future work will investigate how to add deeper time series analysis to the solution since it includes temporal information; thus, more valuable information can be extracted and used in the process of training virtual sensors. In addition, for data compression techniques for optimal sensor placement, other correlation metrics should be considered because they help reduce the need for more physical sensors and work with various types of data.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGEMENT

We acknowledge Ho Chi Minh City University of Technology (HCMUT) and VNU-HCM for supporting this study. We acknowledge the support of time and facilities from the High Performance Computing Laboratory, HCMUT. We also acknowledge the support and collaboration from the TIST Laboratory.

This research was conducted within 58/20-DTDL. The CN-DP Smart Village project was sponsored by the Ministry of Science and Technology of Vietnam.

AUTHORS CONTRIBUTIONS

All authors significantly contributed to this work. All authors read and approved the final version of the manuscript for publication.

REFERENCES

1. IHS. Markit. 2018: The top transformative technologies to watch this year;8;Available from: Com/www/pdf/IHS-Markit-2018-Top-TransformativeTechnology-Trends.pdf.
2. Liu L, Kuo SM, Zhou M. Virtual sensing techniques and their applications. In: International Conference on Networking, Sensing and Control IEEE. Vol. 2009; 2009. p. 31-6;.
3. Li Haorong, Yu D, Braun JE. A review of virtual sensing technology and application in building systems. HVAC R Res. 2011;17(5):619-45;Available from: https://doi.org/10.1080/10789669.2011.573051.
4. Quan NT, Hung NQ, Thoai N. Virtual sensor data imputation using generative adversarial imputation nets and Pearson correlation. In: Xin-She Yang, editors. Proceedings of the eighth international congress on information and communication technology. Singapore: Springer Nature Singapore. ISBN: 978-981-99-3236-8; 2024. p. 507-16;Available from: https://doi.org/10.1007/978-981-99-3236-8_40.
5. Grandi U, Loreggia A, Rossi F, Saraswat V. A Borda count for collective sentiment analysis. Ann Math Artif Intell. 2016;77(3-4):281-302;Available from: https://doi.org/10.1007/s10472-015-9488-0.
6. Goodfellow I et al. Generative adversarial nets. Adv Neural Inf Process Syst. 2014;27;.
7. Yoon J, Jordon J, Schaar M. Gain: missing data imputation using generative adversarial nets. In: International conference on machine learning. PMLR; 2018. p. 5689-98;.
8. Li Y, Shen Y, Zhang W, Chen Y, Jiang H, Liu M et al. Openbox: A generalized black-box optimization service. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining; 2021. p. 3209-19;Available from: https://doi.org/10.1145/3447548.3467061.
9. Brunello A, Urgolo A, Pittino F, Montvay A, Montanari A. Virtual sensing and sensors selection for efficient temperature monitoring in indoor environments. Sensors (Basel). 2021;21(8):2728;PMID: 33924423. Available from: https://doi.org/10.3390/s21082728.
10. Najjar N, Gupta S, Hare J, Kandil S, Walthall R. Optimal sensor selection and fusion for heat exchanger fouling diagnosis in aerospace systems. IEEE Sens J. 2016;16(12):4866-81;Available from: https://doi.org/10.1109/JSEN.2016.2549860.
11. Palmer KA, Bollas GM. Sensor selection embedded in active fault diagnosis algorithms. IEEE Trans Control Syst Technol. 2019;29(2):593-606;Available from: https://doi.org/10.1109/TCST.2019.2955042.
12. Jin H, Su L, Chen D, Nahrstedt K, Xu J. Quality of information aware incentive mechanisms for mobile crowd sensing systems. In: Proceedings of the 16th ACM international symposium on mobile Ad Hoc networking and computing; 2015. p. 167-76;Available from: https://doi.org/10.1145/2746285.2746310.
13. Jawaid ST, Smith SL. Submodularity and greedy algorithms in sensor scheduling for linear dynamical systems. Automatica. 2015;61:282-8;Available from: https://doi.org/10.1016/j.automatica.2015.08.022.
14. Vitus MP, Zhang W, Abate A, Hu J, Tomlin CJ. On efficient sensor scheduling for linear dynamical systems. Automatica. 2012;48(10):2482-93;Available from: https://doi.org/10.1016/j.automatica.2012.06.092.
15. Gupta V, Chung TH, Hassibi B, Murray RM. On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage. Automatica. 2006;42(2):251-60;Available from: https://doi.org/10.1016/j.automatica.2005.09.016.
16. Manohar K et al. Data-driven sparse sensor placement for reconstruction: demonstrating the benefits of exploiting known patterns. IEEE Control Syst. 2018;38(3):63-86;Available from: https://doi.org/10.1109/MCS.2018.2810460.
17. Wu J, Jia Q, Johansson KH, Shi L. Event-based sensor data scheduling: trade-off between communication rate and estimation quality. IEEE Trans Autom Control. 2012;58(4):1041-6;Available from: https://doi.org/10.1109/TAC.2012.2215253.
18. Mo Y, Ambrosino R, Sinopoli B. Sensor selection strategies for state estimation in energy constrained wireless sensor networks. Automatica. 2011;47(7):1330-8;Available from: https://doi.org/10.1016/j.automatica.2011.02.001.
19. Hare JZ, Gupta S, Wettergren TA. POSE: prediction-based opportunistic sensing for energy efficiency in sensor networks using distributed supervisors. IEEE Trans Cybern. 2018;48(7):2114-27;PMID: 28809721. Available from: https://doi.org/10.1109/TCYB.2017.2727981.
20. Sheltami T, Musaddiq M, Shakshuki E. Data compression techniques in wireless sensor networks. Future Gener Comput Syst. 2016;64:151-62;Available from: https://doi.org/10.1016/j.future.2016.01.015.
21. Ercan T, Papadimitriou C. Optimal sensor placement for reliable virtual sensing using modal expansion and information theory. Sensors (Basel). 2021;21(10):3400;PMID: 34068203. Available from: https://doi.org/10.3390/s21103400.
22. Ludwin WG. Strategic voting and the Borda method. Public Choice. 1978;33(1):85-90;Available from: https://doi.org/10.1007/BF00123946.
23. Hutter F, Hoos HH, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration. In: International conference on learning and intelligent optimization. Springer; 2011. p. 507-23;Available from: https://doi.org/10.1007/978-3-642-25566-3_40.

Table 2: The results of SRM solved as black-box optimization by openbox

Dataset	No. sensors reduced ⇓	RMSE		R^2	
		Ranking	Nonranking	Ranking	Nonranking
Solar	4	0.064	0.073	0.875	0.848
	5	0.070	0.075	0.851	0.858
	6	0.077	0.121	0.829	0.762
Raspihat	3	0.032	0.042	0.928	0.829
	4	0.046	0.056	0.856	0.748
	5	0.114	0.102	0.344	0.241
Traffic	5	0.088	0.129	0.868	0.778
	6	0.096	0.147	0.848	0.711
	7	0.146	0.265	0.665	0.047

24. Ilyas EB, Fischer M, Iggena T, Tonjes R. Virtual sensor creation to replace faulty sensors using automated machine learning techniques. In: Global Internet of Things Summit (GIoTS). Vol. 2020. IEEE Publications; 2020. p. 1-6;PMID: 32412188. Available from: https://doi.org/10.1109/GIOTS49054.2020.9119681.
25. ITK digital. Solar Power Panel Dataset at Open Data DK;Available from: https://www.opendata.dk/city-of-aarhus/solcelleanlaeg.
26. Li Y et al. Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In: International Conference on Learning Representations (ICLR '18); 2018;.