frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 26 October 2023
DOI 10.3389/fenvs.2023.1295526

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
University of Waterloo, Canada

REVIEWED BY
Olena Pavliuk,
Silesian University of Technology, Poland
Parvaiz Ahmad Naik,
Youjiang Medical University for Nationalities, China

*CORRESPONDENCE
Ivan Izonin,
ivan.v.izonin@lpnu.ua

RECEIVED 16 September 2023
ACCEPTED 06 October 2023
PUBLISHED 26 October 2023

CITATION
Izonin I, Tkachenko R, Krak I, Berezsky O, Shevchuk I and Shandilya SK (2023), A cascade ensemble-learning model for the deployment at the edge: case on missing IoT data recovery in environmental monitoring systems. *Front. Environ. Sci.* 11:1295526. doi: 10.3389/fenvs.2023.1295526

A cascade ensemble-learning model for the deployment at the edge: case on missing IoT data recovery in environmental monitoring systems

Ivan Izonin[1]*, Roman Tkachenko[2], Iurii Krak[3,4], Oleh Berezsky[5], Ihor Shevchuk[1] and Shishir Kumar Shandilya[6]

[1]Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine, [2]Department of Publishing Information, Lviv Polytechnic National University, Lviv, Ukraine, [3]Department of the Theoretical Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, [4]Intelligence Communicative Information Laboratory, Glushkov Cybernetics Institute, Kyiv, Ukraine, [5]Department of Computer Engineering, West Ukrainian National University, Lviviska, Ukraine, [6]School of Data Science and Forecasting, Devi Ahilya University, Indore, Madhya Pradesh, India

In recent years, more and more applied industries have relied on data collection by IoT devices. Various IoT devices generate vast volumes of data that require efficient processing. Usually, the intellectual analysis of such data takes place in data centers in cloud environments. However, the problems of transferring large volumes of data and the long wait for a response from the data center for further corrective actions in the system led to the search for new processing methods. One possible option is Edge computing. Intelligent data analysis in the places of their collection eliminates the disadvantages mentioned above, revealing many advantages of using such an approach in practice. However, the Edge computing approach is challenging to implement when different IoT devices collect the independent attributes required for classification/regression. In order to overcome this limitation, the authors developed a new cascade ensemble-learning model for the deployment at the Edge. It is based on the principles of cascading machine learning methods, where each IoT device that collects data performs its analysis based on the attributes it contains. The results of its work are transmitted to the next IoT device, which analyzes the attributes it collects, taking into account the output of the previous device. All independent at-tributes are taken into account in this way. Because of this, the proposed approach provides: 1) The possibility of effective implementation of Edge computing for intelligent data analysis, that is, even before their transmission to the data center; 2) increasing, and in some cases maintaining, classification/regression accuracy at the same level that can be achieved in the data center; 3) significantly reducing the duration of training procedures due to the processing of a smaller number of attributes by each of the IoT devices. The simulation of the proposed approach was performed on a real-world set of IoT data. The missing data recovery task in the atmospheric air state data was solved. The authors selected the optimal parameters of the proposed approach. It was established that the developed model provides a slight increase in prediction accuracy while significantly reducing the duration of the

training procedure. However, in this case, the main advantage is that all this happens within the bounds of Edge computing, which opens up several benefits of using the developed model in practice.

KEYWORDS

environmental monitoring, cascading, missing data recovery, edge computing, machine learning, a developed network of IoT devices

1 Introduction

Modern systems for monitoring the state of the environment, in particular air, are based on data collection by IoT devices (Li et al., 2019). Considering the type and scale of the surveillance object, such a process can take place using a single IoT device or a developed network of IoT devices. The data collected by both options is generally sent to a centralized data center for analysis (Yassine et al., 2019). Most data centers are located in cloud services for storing and processing information (Shakerkhan and Abilmazhinov, 2019; Saxena et al., 2021). This approach provides the possibility of using knowledge-intensive and, at the same time, resource-consuming methods of intelligent data analysis and returning instructions to information-gathering devices based on high-precision prediction or classification. It is ensured by the enormous computing power of data centers that can be used to obtain the highest possible accuracy in analyzing large volumes of data (Kumar et al., 2019). On the other hand, this approach requires significant time and energy costs for transferring vast amounts of data for analysis. In particular, the time of receiving information plays a key role here since reaction delays while waiting for decisions from the data center can lead to significant losses (Bisikalo et al., 2020a; Bisikalo et al., 2020b).

The problem of transmitting a large amount of data in real time with minimal costs from remote IoT sensors can be mitigated by the use of intelligent devices at the Edge. (Li et al., 2019). These devices can perform pre-processing of data immediately at the places of their collection and transfer not high-dimensional volumes of data to the data centers but their knowledge in a format of smaller dimensions. It will reduce the necessary energy and time costs for data transmission, which, in general, will increase the efficiency of environmental monitoring systems (Alakbarov and Alakbarov, 2018).

For this purpose, in recent years, the concept of edge computing has become widespread, where part of the data can be processed in local networks, in which each device, having specific computing capabilities, can become a node of edge computing (Raj et al., 2022). Accordingly, a set of similar nodes can provide preliminary processing of IoT data.

In (Hassan et al., 2018), existing technologies of intelligent data analysis at the Edge are analysed. The authors overviewed many literary sources and summarized the features of existing Edge Computing paradigms. Probably the greatest contribution of this theoretical work is the description of the main requirements for the implementation of Edge Computing in practice according to various scenarios.

The application of this concept for data pre-processing seems appropriate when the data collected by IoT devices contain gaps or anomalies. Gaps in such systems arise for a variety of reasons, which have been investigated in (Arroyo et al., 2018). The main one is the failure of one or more sensors of a specific node (IoT device). If such a sensor collects data on harmful chemical impurities in the air that threaten human health or even life, then such data are critical for analysis. Therefore, there is a problem of filling gaps in such data to function the entire ecological monitoring system based on IoT fully. Efficiency here is determined by the accuracy and time of solving this problem to preserve the integrity of the data even before it is transferred to the data center (until the sensor is repaired or replaced). A rational solution to this problem will ensure the reliability and uninterrupted operation of environmental monitoring systems, in particular subsystems, for assessing the state of the air environment.

Simple gap-filling methods, such as averaging and others, do not provide sufficient accuracy in the work (Mishchuk et al., 2020). This will be reflected in the accuracy of decisions made by the data center, and as a result, in the effectiveness of the entire system of monitoring the state of environmental monitoring. That is why there is a need to use machine learning tools to increase the accuracy of solving the given problem. Analysis of several independent attributes using machine learning methods demonstrates a significantly higher prediction/classification accuracy than other methods (Mishchuk et al., 2020).

In the case when a single IoT device collects data, that is, all independent changes are in one place, this problem can be solved by simply adding more computing power to the device for the possibility of implementing training procedures in the device itself (Kryvonos et al., 2016; Kotsovsky et al., 2020). In particular, (Savaglio and Fortino, 2021), solved the missing data recovery task using new, EdgeMiningSim methodology, which is focused on Edge computing. The authors developed an application that was used to evaluate the monitoring system in various scenarios.

However, when a developed network of IoT devices collects data (Ageyev et al., 2022), the solution to the given task becomes significantly more complicated. Since all the independent attributes are collected by different IoT devices, transferring them all to one device to implement the training procedure within Edge computing is not optimal (Geche et al., 2022). In addition, such an option requires enormous power for such a super device, and there is a high risk of its breakdown or failure.

The detailed discussion of this problem are presented in (D'Agostino et al., 2019). The authors considered one of the bioinformatics tasks, which is based on a huge volume of data collected by IoT devices. The paper proposes an approach to combining Edge and Cloud computing to increase the efficiency of the entire process of medical data analysis. This approach looks effective both in terms of computing resources and speed of work and can be used in other areas of medicine as well (Tabassum et al., 2021a).

A similar problem arises in the manufacturing area (Chen et al., 2018). However, for its solution in this case, the authors limited themselves to the use of Edge computing only. In particular, they developed a new architecture of edge computing for IoT-based manufacturing. It provides a significant increase in the production process (till 96%), which has been studied in detail from various aspects. This is one of the few applied works devoted to the problem of effective deployment of such systems. However, in (Hung, 2021) an improved scheme for the implementation of Edge computing in manufacturing is proposed. It is based on the use of ensemble learning (boosting). The proposed approach ensured high accuracy of solving the predicting yield failure task in a semiconductor manufacturing process. However, the boosting strategy has a number of disadvantages that significantly affect its practical use specifically for the implementation of Edge computing. The first is that noise in the input data can lead to unstable predictions. In addition, Boosting ensemble can be extremely time-consuming for training, especially if the model contains a large number of underlying models (e.g., decision trees).

In (Eddine et al., 2022) an improved boosting scheme is proposed. The basis of the intelligent component for the implementation of Edge computing here is the use of Random Forest algorithm. It provides a higher prediction accuracy compared to the previous method, but it is also characterized by a number of disadvantages. Random Forest can be slow on large datasets, especially when the number of trees in the forest is large. Usually, for big data, other algorithms are used, which can be faster and more efficient (Tabassum et al., 2021b). In addition, if the data set is represented by a large number of features, which is typical for IoT data, and they have different importance, it may happen that Random Forest will use less useful features and this may lead to less accurate predictions. In this regard, the implementation of Edge computing, seems appropriate using a cascade of machine learning methods (Vermesan et al., 2018; Mochurad, 2021) based on the entire network of IoT devices that:

1) Collect independent attributes important for analysis;
2) Have the computing power to implement the training procedure.

In this case, the number of devices that collect the attributes necessary for intelligent analysis will actually determine the number of cascade levels. This approach involves using a machine learning method inside each device, which will process only the attributes it collects. At the same time, the results of its work will be transmitted as an additional attribute to the next level of the cascade (the next IoT device) as a generalization of those features during the analysis that was collected at the previous level.

It should be noted that the composition of such a cascade should be sufficiently simple and accurate for the effective implementation of Edge computing (Izonin et al., 2019). In addition, the fundamental basis of machine learning methods that can be used for similar purposes should be high speed and accurate while analyzing large sets of collected IoT data. It will ensure the high accuracy of the analysis with a minimum duration of training procedures, reducing the total data pre-processing time. In addition, such methods should be as simple as possible to implement. It will allow data to be processed inside each IoT device of the cascade, characterized by rather limited computing capabilities. Because of this, it seems appropriate to use linear, i.e., high-speed methods of analysis of large datasets for implementing Edge computing in systems of ecological monitoring of the state of the air environment. To increase the accuracy of their work, we can use various variants of non-linear expansion of the inputs (Mishchuk et al., 2020).

Therefore, this paper aims to develop a cascade of machine learning methods for implementing Edge computing when solving the task of filling gaps in data from an extended network of Internet of Things (IoT) devices.

The research object is the preliminary processing of IoT data for ecological monitoring of the state of the air environment.

The research subject is the ensemble machine learning methods for deployment at the Edge to fill data gaps from a developed network of IoT devices.

The main contribution of this article is as follows:

1. We developed a new ensemble machine learning method for data pre-processing from a developed network of Internet of Things devices, which is based on the principles of cascading machine learning methods and the linearization of the response surface. The method is based on the idea of processing part of the data inside each smart IoT device from the developed network that it collects and transfers the predicted value as an additional feature for the next smart IoT device (linearization principle). Machine learning is performed using the fast linear SVR, and the increase in accuracy is achieved using the non-linear expansion of the inputs at each level of the method by the Kolmogorov-Gabor polynomial;
2. We investigated the order of inclusion of various devices from the developed IoT network in the proposed cascade on the method's accuracy. The optimal combination of it was determined experimentally;
3. We investigated the influence of different orders of the Kolmogorov-Gabor polynomial as a tool for the non-linear expansion of the inputs of each cascade node on the accuracy of the method. The optimal value of this indicator was established experimentally.

The practical value of the proposed solution is the following:

- The possibility of effective implementation of Edge computing to fill gaps in data at the places of their collection, i.e., even before they are transferred to the data center;
- Preserving, and in some cases, increasing the accuracy (based on various performance indicators) of filling data gaps at the same level that can be achieved in the data center;
- A significant reduction in the duration of training procedures due to the processing of a smaller number of attributes by each of the IoT devices in comparison with the use of a similar method of IoT data analysis in a data center located in a cloud environment;
- Increasing the reliability and uninterrupted operation of environmental monitoring systems, in particular, subsystems for assessing the state of the air environment based on ensuring the integrity of data transmitted to the data center.

This paper is organized as follows: Section 2 presents the background and proposed cascade method, its training, and

application algorithms. Section 3 gave the modeling and results of the optimal parameters selection procedures. In Section 4 the authors compared similar methods that can be applied only in the cloud and discussed the results. Conclusion are in Section 5.

2 Materials and methods

This paper proposes a new ensemble model of machine learning algorithms that can be implementing at the Edge. The basis of the proposed model is using one of the most accurate classes of ensemble methods—cascading. The authors proposed combining a linear version of the regressor based on the Support Vector Machine and the Ito decomposition for non-linear input expansion. Such a combination in the proposed cascade provides both high-accuracy and high-speed solving of prediction tasks and the possibility of their implementation at the Edge, i.e., near the immediate locations of IoT data collection.

2.1 Linear SVR

The support vector machine (SVM) is one of the well-known methods for solving classification and regression problems (Piletskiy et al., 2020; Babenko et al., 2021). The SVM is based on the idea of finding an optimal hypersurface that would minimize losses in multidimensional space (Mamat et al., 2023). Essentially, SVM searches for regression coefficients by minimizing quadratic losses. If the current value falls within the region of the constructed optimal hypersurface, then the losses are zero. Otherwise, the loss is equal to the difference between the actual and predicted values of the desired quantity. A detailed mathematical description of this well-known machine-learning method is given in (Shang et al., 2018).

SVR provides the flexibility of analysis due to the possibility of using different methods' kernels. In the case of non-linear kernels, this method allows efficient analysis of non-linear response surfaces (Pasieka et al., 2020).

In this paper, we will use the linear SVR kernel since this approach provides high processing speed, and high accuracy will be kept due to the use of the Ito decomposition for the non-linear transformation of the input data.

2.2 Ito decomposition

Nowadays, various methods of non-linear expansion of inputs are very often used to increase the accuracy of linear machine learning methods. One of the possible options for implementing this is the use of a discrete Volterra series—the Ito decomposition. Second-order Ito decomposition for a given set of input parameters $x_1, ..., x_n$ can be represented in the following form (α_i are coefficients of Ito decomposition) (FERREIRA, 2002):

$$Y(x_1, ..., x_n) = \alpha_i + \sum_{i=1}^{n} \alpha_i x_i + \sum_{i=1}^{n} \sum_{j=i}^{n} \alpha_{i,j} x_i x_j \quad (1)$$

Such a technique has high approximation properties and significantly reduces errors, particularly in linear machine learning methods (Mochurad et al., 2020). In addition, in the proposed scheme, the Ito decomposition will also model the connection between the sensor data and the output of the previous cascade level, which will also positively affect the method's accuracy.

2.3 Designed cascade-based ensemble-learning model

Modern monitoring systems for various purposes are based on data collection by IoT devices. The scale of such systems requires many devices to collect data and transfer them to the data center for processing. This approach involves a lot of energy and resources to transmit data to the data center. Therefore there is a need to find alternative ways of data analysis, particularly in the devices that collect them. In the case of data collection by a developed network of IoT devices, the application of machine learning methodology is significantly limited.

In this paper, a new cascade machine learning model is developed for deployment in the Edge to collect all the attributes necessary for the analysis from the developed network of IoT devices. It eliminates the shortcomings mentioned above while ensuring a significant increase in the speed of training procedures while maintaining the level of accuracy that can be obtained in a data center.

The approach is based on using all IoT devices of a developed network that collect the attributes necessary for analysis. In essence, data analysis occurs in each of these devices based on the attributes it collects. Their fusion is ensured using the principles of cascading (Bodyanskiy and Tyshchenko, 2019), where the predicted value of the searched attribute from the previous IoT device is passed to the next one as an additional feature. In addition, the authors used the Ito decomposition at each new level of the cascade. This decomposition made it possible to model the dependencies between predicted output from the previous IoT device and the attributes of the current IoT device. In turn, such interconnections provided additional information for the selected machine learning method at each new level of the proposed cascade scheme (Babichev et al., 2018). Combined with very high approximation properties of the Ito decomposition, its use made it possible to increase the prediction accuracy of the desired attribute at each level of the cascade, as well as the overall accuracy of the proposed scheme.

The flow chart of the proposed approach is shown in Figure 1.

The algorithmic implementation of the training procedure of the machine learning model implemented in this paper requires the following steps to be followed sequentially:

1. To determine the number of IoT devices that collect data necessary for analysis. The number of devices, in this case, will determine the number of levels of the proposed model;
2. To set the order of inclusion of devices in the cascade scheme;
3. To perform the Ito decomposition on the attributes collected by the first IoT device of the cascade and conduct training of the linear variant of the SVR
4. To transfer the output of the SVR obtained in the previous step as an additional attribute to the next IoT device; to perform the Ito decomposition over N+1 attributes of the current IoT device and train the next linear SVR

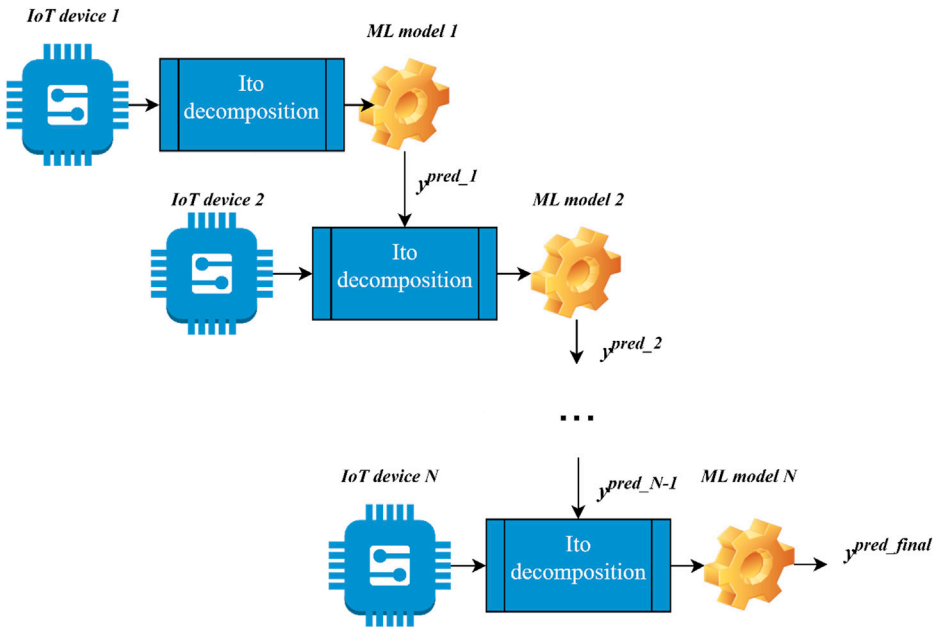

FIGURE 1
Flow-chart of the proposed cascade ensemble-learning model for the deployment at the Edge in case of data collection by the developed network of IoT devices.

5. To repeat step 4 until the last IoT device on the cascade is reached.

Then, we have a pre-trained model with a certain number of levels in the application mode. We sequentially apply the test data set or the current vector to each previously trained model level according to steps 3–4 of the training algorithm. The last level of the model, i.e., obtained by the prediction on the last of the IoT devices of the cascade model, will determine the predicted value of the desired attribute.

3 Modeling and results

When various IoT devices collect a large number of features about a specific observation, the use of artificial intelligence tools for implementing Edge computing is significantly limited (since all features should be collected in one place) (Al Shahrani et al., 2022). The developed method in this paper is based on an approach that allows a smart IoT device to process only that part of the data it collects and transfer the predicted value as an additional attribute to the next one smart IoT device from the developed network. This approach is based on the ensemble of machine learning methods (Li et al., 2019) and is justified by the principles of linearization of the response surface (Izonin et al., 2023). It is the basis for the effective implementation of Edge computing.

The modern element base and computing capabilities of smart IoT devices allow the implementation of machine learning methods at its core (Rocha Neto et al., 2018). This, together with the cascade ensemble developed by the authors of this paper, provides the possibility of implementing basic operations of data pre-processing using artificial intelligence tools in the places of data collection, that is, inside of smart IoT devices.

The basis of the simulation carried out in this section is the need to determine and analyze the effectiveness of applying the proposed cascade method for solving the stated task. This simulation aims to determine whether the proposed approach, which can be implemented in the conditions of Edge computing, will provide the same accuracy and speed of work that can be obtained by transmitting data from all IoT devices of the developed network to the cloud, that is, to the data center.

In this paper, we operated with real-world data collected by IoT devices and studied the parameters of operation and the overall efficiency of the proposed cascade method for solving the stated task in the situation of the need to process data at the places of their collection.

The proposed approach was simulated on the laptop with the following characteristics: Intel Core i5-600U processor (2.40 GHz), *p* 8.00 GB RAM, and a 64-bit operating system.

3.1 Dataset descriptions

Modeling of the proposed approach to filling gaps in data is performed on a real environmental monitoring task (Izonin et al., 2019; Mishchuk et al., 2020). The implementation of the method took place on a data set that contains 9,000 observations of hourly averaged responses from metal oxide chemical sensors for determining air quality.

We used the data from devices that were located on the field in a heavily polluted area at the road level. The dataset consists of hourly averaged concentrations of carbon monoxide (CO), tungsten oxides (WO), benzene (C_6H_6), titanium (Ti), non-methane hydrocarbons (SnO_2), total nitrogen oxides (NO), nitrogen dioxide (NO_2),

TABLE 1 Performance indicators for the proposed model in test mode.

Case N/performance indicators	Case 1	Case 2	Case 3	Case 4
Second-degree Ito decomposition				
MAE	0.271	0.271	0.274	0.269
MSE	0.229	0.229	0.232	0.229
MAPE	0.195	0.196	0.196	0.195
RMSE	0.479	0.479	0.481	0.479
MaxE	8.483	8.460	8.512	8.491
MedAE	0.167	0.169	0.170	0.163
Training time	2.81	2.7	2.98	2.92
Third-degree Ito decomposition				
MAE	0.264	0.264	0.266	0.262
MSE	0.213	0.214	0.216	0.213
MAPE	0.191	0.192	0.191	0.191
RMSE	0.462	0.462	0.465	0.461
MaxE	8.541	8.544	8.536	8.503
MedAE	0.165	0.165	0.164	0.161
Training time	6.67	6.51	8.35	8.7

tungsten dioxide (WO_2), air temperature (T), indium oxide (InO) relative (RH) and absolute (AH) air humidity (De Vito, 2016).

Since the carbon monoxide (CO) column contains the most significant number of omissions (De Vito, 2016), the simulation was carried out precisely to restore the missing data for this indicator of air composition.

3.2 Optimal parameters selection for the proposed cascade scheme

The developed cascade ensemble-learning model requires the selection of several parameters to obtain optimal predicted values of the sought variables. The following should be noted among them:

- The order for processing smart IoT devices in the proposed model;
- Degree of Ito decomposition at each level of the cascade;
- Optimal SVR operating parameters.

The paper considers missing data recovery tasks collected by IoT sensors. The first device collects 8 attributes (CO, WO, SnO_2, NO_2, C_6H_6, NO, Ti, WO_2, and InO). The second IoT device collects one attribute (T), and the third IoT device collects 2 features (RH and AH). In order to determine the optimal order of submission of IoT devices for processing in the model proposed in this work, we will consider all possible variants. It should be noted that since the second device collects only one attribute, it cannot stand in the first place of the developed model. Therefore, there were 4 different options for three devices:

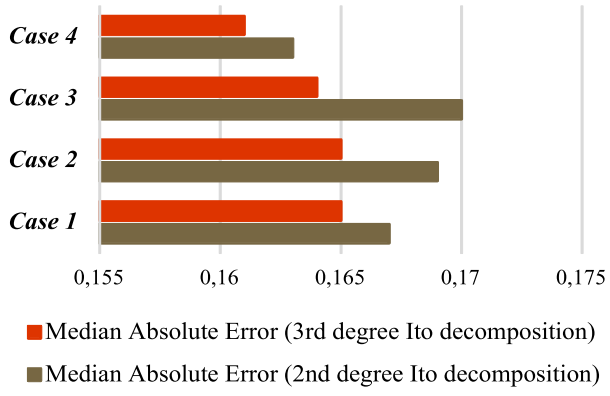

FIGURE 2
MedAE for implementation of the 4 cases of the proposed model in test mode with 2-th and 3-rd Ito decomposition degree.

1. Case 1: firsts device—second device—third device;
2. Case 2: firsts device—third device—second device;
3. Case 3: third device—second device—firsts device;
4. Case 4: third device—firsts device—second device.

The results of this experiment for the second and third orders of the Ito decomposition are summarized in Table 1.

The higher orders of this decomposition were not taken into account because they significantly increase the dimensionality of the input data space and, therefore, greatly complicate the training procedure and can even provoke overfitting.

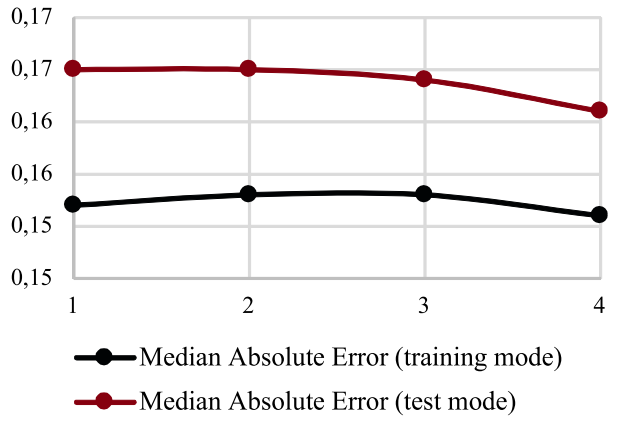

FIGURE 3
MedAE for the implementation of the 4 cases of the proposed model in training and test modes.

To visualize the results of this experiment, Figure 2 summarizes MedAE errors for four cases when using second and third-degree polynomials.

As can be seen from Figure 2, the most accurate results are obtained when using case 4 and the third order of the Ito decomposition. In addition, Figure 3 summarizes MedAE errors for training and application modes for all 4 cases.

As can be seen from Figure 3, all 4 cases provide a reasonably high degree of generalization. However, the best results were obtained for case 4.

All performance indicators for the proposed model using its optimal parameters are presented in Table 2.

4 Discussion

In order to check the effectiveness of the application of the developed model at the Edge, this work compares its work with basic methods that can be implemented in a data center. In particular, the combined use of the Ito decomposition of the second and third orders and SVR with linear kernel were chosen.

Also, we have compared our cascade with the existing SVR (with RBF kernel)-based cascade scheme (Izonin et al., 2023). The last method was developed using similar principles but without an additional module of non-linear input extension because RBF functions did the same. Also SVR (RBF)-based cascade scheme can't be implemented at the Edge as on each level of this cascade, SVR with RBF kernel processing part of the whole dataset (with all attributes). Other ensemble methods from the boosting, bagging, and stacking classes cannot be used to solve the stated task as they should use the whole dataset for analysis and, therefore, can be implemented at the Edge.

The performance results of all studied methods when selecting their optimal parameters based on RMSE, Maximum Error (Max error), and their training time are summarized in Figure 4.

As can be seen in Figure 4, SVR with linear kernel provides the highest speed of the training procedure but the lowest accuracy of operation. The use of the Ito decomposition, even of the second degree, ensures a significant increase in accuracy but also a significant increase in the duration of the training procedure. As expected, the third order of the Ito decomposition increases the prediction accuracy. However, the duration of the training procedure increases by more than 15 times compared to linear SVR. It is explained by a significant increase in the number of independent attributes due to the application of Ito decomposition. The existing SVR (with RBF kernel)-based cascade scheme (Izonin et al., 2023) provides a good performance indicator (Figure 4). However, due to the principles of its training and test algorithms, this method cannot be implemented at the Edge computing.

The proposed model demonstrated the highest prediction accuracy among all studied methods. Moreover, it showed an increase in accuracy even compared to using SVR with 3-rd degree Ito decomposition, which can be applied in a data center. In addition to this advantage, the proposed model provided more than three times the acceleration of the training procedure. This happened due to processing a significantly smaller number of features compared to the method mentioned above.

However, the effectiveness of using the developed cascade scheme largely depends on the machine learning method that will be its basis. It should be noted that efficiency in this case is determined by accuracy and speed. In the case of big data analysis, the classic SVR will not provide either sufficient speed or satisfactory prediction accuracy. In order to avoid both of these drawbacks, further research will be conducted in the direction of using non-iterative artificial neural networks at each step of the cascade (Medykovskvi et al., 2018). They will

TABLE 2 Performance indicators for the optimized model in each level of the ensemble in the test mode.

Case N/performance indicators	IoT device 3	IoT device 1	IoT device 2	Final result
MAE	1.094	0.264	0.261	0.261
MSE	2.302	0.213	0.212	0.212
MAPE	0.756	0.191	0.190	0.19
RMSE	1.517	0.462	0.460	0.460
MaxE	9.879	8.516	8.510	8.51
MedAE	0.844	0.166	0.159	0.16
Training time	0.49	7.52	0.68	8.7 s

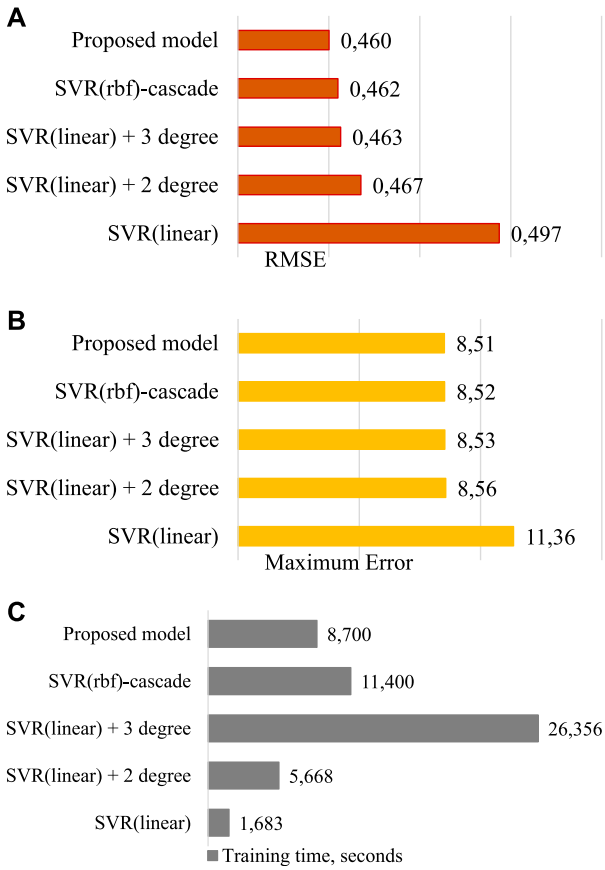

FIGURE 4
Comparison with existing methods: **(A)** RMSE; **(B)** Maximum Error; **(C)** training time, seconds.

ensure both high accuracy of work and high speed due to the non-iterative nature of their training procedure (Mishchuk et al., 2020).

To sum up, a cascade ensemble-learning model developed in this work ensured both an increase in accuracy and a significant reduction in the duration of the training process compared to a similar method that can be used in a data center. However, in this case, the main advantage is the possibility of deploying the developed model to implement Edge computing in the case of data collection by a developed network of IoT devices. It opens up several advantages for applying the developed model in practice.

5 Conclusion

This paper is devoted to monitoring the state of the air environment based on data collected by IoT devices. Because modern monitoring systems are quite large, they use many IoT devices to collect data. In the vast majority, data is collected by a developed network of IoT devices in the data center. In this case, using traditional machine learning methods and paradigms fully justifies itself. However, the process of transferring vast volumes of data to data centers is accompanied by considerable consumption of energy and other resources, which necessitates the search for more effective ways of analyzing such data. One of them is the use of the concept of Edge computing. However, when a developed network of IoT devices collects data, the use of classical machine learning approaches at the border is significantly complicated.

In this paper, we developed a new cascade ensemble of machine learning methods for pre-processing data from a developed network of IoT devices based on a linear regressor with non-linear input mapping. It provides the ability to implement Edge computing effectively; significantly reducing the duration of training procedures; preserving, and in some cases, increasing the prediction accuracy.

The modeling was carried out on a real-world set of IoT data. The task of filling gaps in atmospheric air data was solved. The optimal values of the parameters of the developed model were determined experimentally. We show the optimal order of including IoT sensors in the proposed three-level cascade for the investigated task is the following: IoT devise 3, IoT devise 2, and IoT devise 3. At the same time, the optimal value of input expansion for this task is the use Kolmogorov-Gabor polynomial of the third order. Based on this, we established a reduction in the duration of the training procedure by more than three times (from 26.4 to 8.7 s) and even a slight increase in prediction accuracy compared to a similar method that can be applied in the data center (in the cloud).

Among the prospects for further research is the possibility of using: 1) feature selection techniques and PCA to reduce the dimensionality of the input data space, 2) neural networks to increase the prediction accuracy, and 3) non-iterative machine learning algorithms to reduce the duration of learning the cascade model will be considered. Also, the proposed approach can be used not only for filling data gaps, but also for other data preprocessing tasks (for example, anomaly detection) that are collected by a developed network of Internet of Things devices.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

II: Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft. RT: Conceptualization, Formal Analysis, Resources, Supervision, Writing–review and editing. IK: Conceptualization, Formal Analysis, Supervision, Visualization, Writing–review and editing. OB: Conceptualization, Formal Analysis, Resources, Supervision, Writing–review and editing. IS: Data curation, Software, Validation, Visualization, Writing–review and editing. SS: Conceptualization, Formal Analysis, Funding acquisition, Supervision, Writing–review and editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. The

National Research Foundation of Ukraine funded this research under the project № 03/0103.

Acknowledgments

The authors would like to thank: 1) the Armed Forces of Ukraine because this paper has become possible only because of the resilience and courage of the Ukrainian Army; 2) the reviewers for the correct and concise recommendations that helped present the materials better; 3) Mr. Myroslav Havrylyuk and Mr. Kyrylo Yemets for their excellent technical support. This research is supported by the British Academy's Researchers at Risk Fellowships Programme.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ageyev, D., Radivilova, T., Mulesa, O., Bondarenko, O., and Mohammed, O. (2022). "Traffic monitoring and abnormality detection methods for decentralized distributed networks," in *Information security technologies in the decentralized distributed networks*. Editors R. Oliynykov, O. Kuznetsov, O. Lemeshko, and T. Radivilova (Cham: Springer International Publishing). doi:10.1007/978-3-030-95161-0_13

Al Shahrani, A. M., Alomar, M. A., Alqahtani, K. N., Basingab, M. S., Sharma, B., and Rizwan, A. (2022). Machine learning-enabled smart industrial automation systems using Internet of Things. *Sensors* 23 (1), 324. doi:10.3390/s23010324

Alakbarov, G., and Alakbarov, R. (2018). Effective use method of cloudlet resources by mobile users. *IJCNIS* 10 (2), 46–52. doi:10.5815/ijcnis.2018.02.06

Arroyo, Á., Herrero, Á., Tricio, V., Corchado, E., and Woźniak, M. (2018). Neural models for imputation of missing ozone data in air-quality datasets. *Complexity* 2018, 1–14. doi:10.1155/2018/7238015

Babenko, V., Panchyshyn, A., Zomchak, L., Nehrey, M., Artym-Drohomyretska, Z., and Lahotskyi, T. (2021). Classical machine learning methods in economics research: macro and micro level examples. *WSEAS Trans. Bus. Econ.* 18, 209–217. doi:10.37394/23207.2021.18.22

Babichev, S., Lytvynenko, V., Škvor, J., Korobchynskyi, M., and Voronenko, M. (2018). "Information Technology of gene expression profiles processing for purpose of gene regulatory networks reconstruction," in 2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP), Lviv, Ukraine, 21-25 August 2018 (IEEE), 336–341.

Bisikalo, O. V., Kovtun, V. V., and Kovtun, O. V. (2020a). "Modeling of the estimation of the time to failure of the information system for critical use," in 2020 10th International Conference on Advanced Computer Information Technologies (ACIT), Deggendorf, Germany, 16-18 September 2020 (IEEE). doi:10.1109/ACIT49673.2020.9208883

Bisikalo, O. V., Kovtun, V. V., Kovtun, O. V., and Romanenko, V. B. (2020b). "Research of safety and survivability models of the information system for critical use," in 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT), Kyiv, Ukraine, 14-18 May 2020 (IEEE). doi:10.1109/DESSERT50317.2020.9125061

Bodyanskiy, Y. V., and Tyshchenko, O. K. (2019). A hybrid cascade neuro–fuzzy network with pools of extended neo–fuzzy neurons and its deep learning. *Int. J. Appl. Math. Comput. Sci.* 29 (3), 477–488. doi:10.2478/amcs-2019-0035

Chen, B., Wan, J., Celesti, A., Li, D., Abbas, H., and Zhang, Q. (2018). Edge computing in IoT-based manufacturing. *IEEE Commun. Mag.* 56 (9), 103–109. doi:10.1109/mcom.2018.1701231

D'Agostino, D., Morganti, L., Corni, E., Cesini, D., and Merelli, I. (2019). Combining Edge and Cloud computing for low-power, cost-effective metagenomics analysis. *Future Gener. Comput. Syst.* 90, 79–85. doi:10.1016/j.future.2018.07.036

De Vito, S. (2016). UCI machine learning repository: air quality data set. Available from: http://archive.ics.uci.edu/ml/datasets/air+quality (Accessed March 17, 2019).

Eddine, M. M., Benkirane, S., Guezzaz, A., and Azrour, M. (2022). Random forest-based IDS for IIoT edge computing security using ensemble learning for dimensionality reduction. *IJES* 15 (6), 467. doi:10.1504/ijes.2022.129803

FERREIRA (2002). Evolution of Kolmogorov-Gabor polynomials. Available from: https://www.gene-expression-programming.com/GepBook/Chapter4/Section4/SS1.htm (Accessed December 5, 2022).

Geche, F., Mitsa, O., Mulesa, O., and Horvat, P. (2022). "Synthesis of a two cascade neural network for time series forecasting," in 2022 IEEE 3rd International Conference on System Analysis and Intelligent Computing (SAIC), Kyiv, Ukraine, 04-07 October 2022 (IEEE). doi:10.1109/SAIC57818.2022.9922991

Hassan, N., Gillani, S., Ahmed, E., Yaqoob, I., and Imran, M. (2018). The role of edge computing in Internet of Things. *IEEE Commun. Mag.* 56 (11), 110–115. doi:10.1109/mcom.2018.1700906

Hung, Y. H. (2021). Improved ensemble-learning algorithm for predictive maintenance in the manufacturing process. *Appl. Sci.* 11 (15), 6832. doi:10.3390/app11156832

Izonin, I., Greguš, ml. M., Tkachenko, R., Logoyda, M., Mishchuk, O., and Kynash, Y. (2019). "SGD-based wiener polynomial approximation for missing data recovery in air pollution monitoring dataset," in *Advances in computational intelligence*. Editors I. Rojas, G. Joya, and A. Catala (Cham: Springer International Publishing), 781–793.

Izonin, I., Tkachenko, R., Holoven, R., Shavarskyi, M., Bukin, S., and Shevchuk, I. (2023). "Multistage SVR-RBF-based model for heart rate prediction of individuals," in *Advances in artificial systems for medicine and education VI*. Editors Z. Hu, Z. Ye, and M. He (Cham: Springer Nature Switzerland). doi:10.1007/978-3-031-24468-1_19

Kotsovsky, V., Batyuk, A., and Yurchenko, M. (2020). "New approaches in the learning of complex-valued neural networks," in 2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP), Lviv, Ukraine, 21-25 August 2020 (IEEE), 50–54.

Kryvonos, I. G., Krak, I. V., Barmak, O. V., and Bagriy, R. O. (2016). New tools of alternative communication for persons with verbal communication disorders. *Cybern. Syst. Anal.* 52 (5), 665–673. doi:10.1007/s10559-016-9869-3

Kumar, V., Laghari, A. A., Karim, S., Shakir, M., and Anwar Brohi, A. (2019). Comparison of fog computing and cloud computing. *IJMSC* 5 (1), 31–41. doi:10.5815/ijmsc.2019.01.03

Li, B., Zhou, S., Cheng, L., Zhu, R., Hu, T., Anjum, A., et al. (2019b). A cascade learning approach for automated detection of locomotive speed sensor using imbalanced data in ITS. *IEEE Access* 7, 90851–90862. doi:10.1109/access.2019.2928224

Li, X., Zhao, N., Jin, R., Liu, S., Sun, X., Wen, X., et al. (2019a). Internet of Things to network smart devices for ecosystem monitoring. *Sci. Bull.* 64 (17), 1234–1245. doi:10.1016/j.scib.2019.07.004

Mamat, N., Mohd Razali, S. F., and Hamzah, F. B. (2023). Enhancement of water quality index prediction using support vector machine with sensitivity analysis. *Front. Environ. Sci.* 10, 1061835. doi:10.3389/fenvs.2022.1061835

Medykovskvi, M., Pavliuk, O., and Sydorenko, R. (2018). "Use of machine learning technologys for the electric consumption forecast," in 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 11-14 September 2018 (IEEE). doi:10.1109/STC-CSIT.2018.8526617

Mishchuk, O., Tkachenko, R., and Izonin, I. (2020). "Missing data imputation through SGTM neural-like structure for environmental monitoring tasks," in *Advances in computer science for engineering and education II*. Editors Z. Hu, S. Petoukhov, I. Dychka, and M. He (Cham: Springer International Publishing), 142–151.

Mochurad, L. (2021). Optimization of regression analysis by conducting parallel calculations. *CEUR-WS.Org.* 2870, 982–996.

Mochurad, L., Shakhovska, K., and Montenegro, S. (2020). "Parallel solving of fredholm integral equations of the first kind by tikhonov regularization method using OpenMP Technology," in *Advances in intelligent systems and computing IV*. Editors N. Shakhovska and M. O. Medykovskyy (Cham: Springer International Publishing). doi:10.1007/978-3-030-33695-0_3

Pasieka, N., Khimchuk, L., Sheketa, V., Pasieka, M., Romanyshyn, Y., and Lutsan, N. (2020). "Research of dynamic mathematical models of adaptation of members of teams

of developers of infocommunication systems," in 2020 IEEE International Conference on Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, Ukraine, 06-09 October 2020 (IEEE). doi:10.1109/PICST51311.2020.9468086

Piletskiy, P., Chumachenko, D., and Meniailov, I. (2020). "Development and analysis of intelligent recommendation system using machine learning approach," in *Integrated computer technologies in mechanical engineering*. Editors M. Nechyporuk, V. Pavlikov, and D. Kritskiy (Cham: Springer International Publishing). doi:10.1007/978-3-030-37618-5_17

Raj, P., Saini, K., and Surianarayanan, C. (2022). *Edge/fog computing paradigm: the concept, platforms and applications*. First edition. Cambridge, Oxford London: Elsevier, 537.

Rocha Neto, A., Soares, B., Barbalho, F., Santos, L., Batista, T., Delicato, F. C., et al. (2018). "Classifying smart IoT devices for running machine learning algorithms," in *Anais do Seminário Integrado de Software e Hardware (SEMISH)* (João Pessoa: JBCS). Available from: https://sol.sbc.org.br/index.php/semish/article/view/3429.

Savaglio, C., and Fortino, G. (2021). A simulation-driven methodology for IoT data mining based on edge computing. *ACM Trans. Internet Technol.* 21 (2), 1–22. doi:10.1145/3402444

Saxena, S., Zubair Khan, M., and Singh, R. (2021). Green computing: an era of energy saving computing of cloud resources. *IJMSC* 7 (2), 42–48. doi:10.5815/ijmsc.2021.02.05

Shakerkhan, K. O., and Abilmazhinov, E. T. (2019). Development of a method for choosing cloud computing on the platform of paas for servicing the state agencies. *IJMECS* 11 (9), 14–25. doi:10.5815/ijmecs.2019.09.02

Shang, Q., Yang, Z., Gao, S., and Tan, D. (2018). An imputation method for missing traffic data based on FCM optimized by PSO-SVR. *J. Adv. Transp.* 2018, 1–21. doi:10.1155/2018/2935248

Tabassum, M. F., Akram, S., Mahmood-ul-Hassan, S., Karim, R., Naik, P. A., Farman, M., et al. (2021b). Differential gradient evolution plus algorithm for constraint optimization problems: a hybrid approach. *Int. J. Optim. Control, Theor. Appl. (IJOCTA).* 11 (2), 158–177. doi:10.11121/ijocta.01.2021.001077

Tabassum, M. F., Farman, M., Naik, P. A., Ahmad, A., Ahmad, A. S., and Hassan, S. M. U. (2021a). Modeling and simulation of glucose insulin glucagon algorithm for artificial pancreas to control the diabetes mellitus. *Netw. Model. Anal. Health Inf. Bioinforma.* 10 (1), 42. doi:10.1007/s13721-021-00316-4

Vermesan, O., Eisenhauer, M., and Serrano, M. (2018). "3 the next generation Internet of Things – hyperconnectivity and embedded intelligence at the edge," in *Next generation Internet of Things distributed intelligence at the edge and human machine-to-machine cooperation* (River Publishers). doi:10.1201/9781003338963-3

Yassine, A., Singh, S., Hossain, M. S., and Muhammad, G. (2019). IoT big data analytics for smart homes with fog and cloud computing. *Future Gener. Comput. Syst.* 91, 563–573. doi:10.1016/j.future.2018.08.040