IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 July 2024, accepted 7 August 2024, date of publication 16 August 2024, date of current version 30 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3444046

RESEARCH ARTICLE

Sensor Monitoring Techniques in Edge Computing Using Spatio-Temporal Correlation Anomaly Detection Algorithms

RUI ZHANG, LIDE ZHOU, AOQI MEI, AND YIPENG HE
Dongguan Power Supply Bureau of Guangdong Power Grid Corporation, Dongguan 523000, China
Corresponding author: Rui Zhang (Z_z2222@163.com)

ABSTRACT Due to the rapid development of technologies such as cloud computing and the Internet of Things (IoT), wireless sensor networks are becoming increasingly popular in the field of environmental monitoring. Anomaly detection algorithm is often used as the main method of sensor data detection. The development and application of IoT technology has led to a significant increase in data traffic. However, current anomaly detection methods are difficult to effectively detect heterogeneous data sequences from multiple sources. In this study, the sensor monitoring model of an intelligent greenhouse is constructed by using the spatio-temporal correlation anomaly detection algorithm in edge computing. The data is chunked by a sliding window to reduce the error of one-sided estimation of single data on the detection results. The spatial correlation anomaly detection algorithm is formed on the basis of the temporal correlation detection algorithm, fusing the two algorithms with the edge computation to construct a multi-source multi-dimensional data anomaly detection model. The results of the time-related anomaly detection algorithm experiment showed that the F1-score of the algorithm was 91.26%. Compared with other methods, the false alarm rate of spatial correlation anomaly detection algorithm was reduced by 56.50% $\sim$ 83.45%, and F1-score was increased by 1.37% $\sim$ 22.25%. In the case of big data, the detection time of the sensor monitoring model was 0.47s, the required energy consumption was reduced by 36.75% $\sim$ 79.20%, and the delay time was the least. The anomaly detection algorithm in this study is related to time and space, which effectively improves the detection rate and detection accuracy, thus reducing the computing load of the cloud platform, and is superior to the deep learning method in processing delay.

INDEX TERMS Sensor monitoring techniques, anomaly detection, spatio-temporal correlation, sliding window, cloud computing.

I. INTRODUCTION

Wireless sensor networks are widely used in environmental monitoring due to the explosive expansion of cloud computing, the Internet of Things (IoT), and big data. Typically, sensor nodes and aggregation nodes form a distributed network system in wireless sensor networks. The aggregation node receives the environmental data from the sensor nodes, which self-organize to gather it and then sends it via multi-hop relay [1], [2], [3]. Wireless sensor networks are widely used in industries such as transportation, agriculture, and the military, making them a significant field of interest in automation. The rapid development of information technology has led to a higher demand for abnormality detection in the field of automation [4]. However, when dealing with massive amounts of data and extreme wireless network environments, most algorithms proposed for detecting anomalies in sensor data focus solely on solving the time-continuous aspect of single-source data. These approaches include distance-based, integrated learning-based, clustering-based, and deep learning neural network methods, which often overlook the temporal and spatial correlation of multi-source data [5], [6].

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

Simultaneously, the centralized cloud computing platform struggles to handle large-scale sensor data volumes, and its operational processes result in high latency, making it challenging to achieve effective real-time monitoring of abnormal data. Consequently, this study recommends implementing a sensor monitoring model (SMM) based on edge computing (EC) that utilizes the spatio-temporal correlation anomaly detection algorithm (S-TCADA) to accurately detect outliers. A temporal correlation anomaly detection algorithm (TCADA) is proposed based on temporal correlation. To address the challenge of heterogeneous data from multiple sources and the long outlier distance of neighboring temporal data, the study introduces the sliding window (SW) technique for chunking the data. Additionally, S-TCADA is designed for multi-source heterogeneous data to improve detection accuracy and shorten detection time. A hierarchical EC model is used to construct an SMM, which is then applied to smart greenhouses to enhance the accuracy of sensor anomaly detection (AD). Therefore, this study proposes the construction of an S-TCADA-EC-based SMM for the purpose of resolving the issue of precise identification of outliers in heterogeneous data from multiple sources.

The novelty of the study lies in the application of SWs to multi-source data of neighboring time-sequence for data chunking, thus effectively avoiding the resultant errors arising from one-sided estimation of single data. Furthermore, SCADA based on TCADA is proposed for error detection of heterogeneous multi-source sensor data from the perspective of spatial correlation. To address the challenge of detecting anomalous nodes in agricultural greenhouses, the study further develops an S-TCADA-EC through the aforementioned technique, which fully considers the temporal and spatial correlation of multi-source heterogeneous data. The method significantly improves the accuracy and robustness of sensor AD.

The study introduces the use of SW for the chunking of multi-source heterogeneous data, which improves the efficiency of data processing. A TCADA has been developed within the EC framework, and multi-source data has been processed in blocks using software technology, markedly reducing the detection bias caused by isolated data points. Concurrently, a SCADA system has been developed based on the TCADA algorithm. The algorithm conducts an in-depth analysis of sensor data from the perspective of spatial correlation, thereby enhancing the accuracy of AD. The study proposes a sensor anomaly monitoring system that improves the efficiency of AD for multi-source heterogeneous data by combining TCADA with SCADA. This system fully leverages the spatio-temporal correlation of multi-source heterogeneous data, thereby enhancing the detection of anomalies. The proposed system has practical applications in the monitoring of intelligent greenhouses and actively promotes the development of agricultural intelligence. The implementation of this technical method has the potential to advance the field of data AD, furnish a dependable theoretical foundation for sensor monitoring technology, and simultaneously offer technical assistance for agricultural intelligence.

The study is structured into five parts. The first part introduces the background of the research on sensor monitoring technology and briefly describes the proposed methodology. The second part summarizes the current research results and shortcomings of sensor monitoring technology. The third part studies and designs a SMM combining EC on the basis of S-TCADA. The fourth part experiments and analyzes the proposed SMM. Finally, the fifth part summarizes the experimental results and indicates future research directions.

II. RELATED WORKS

In the midst of the digital transformation wave, data AD and EC have become prominent research topics. Data AD aims to identify irregular or anomalous patterns in large datasets to ensure data quality and security. Meanwhile, EC enhances efficiency by processing data near its source, reducing latency [7]. AD algorithms combining temporal and spatial correlation play a significant role in sensor monitoring techniques within the EC environment, improving monitoring system accuracy and efficiency [8]. A selection of relevant research conducted by these individuals is presented below.

T. Wang et al. proposed a data cleaning method based on mobile edge nodes to address the issue of traditional data cleaning methods being unable to ensure data credibility. The method employed angle-based outlier detection at the edge nodes to obtain training data, which effectively improved data cleaning efficiency. However, the method only improved data reliability and not its detection effectiveness [9]. The study proposed a spatio-temporal correlation algorithm design based on the temporal correlation algorithm to improve the effectiveness of data AD. K. Sadaf et al. proposed a deep learning-based intrusion detection approach to identify potential threats in fog computing environments. This classification approach was well-suited for fog devices as they prioritize real-time performance, but were limited to binary classification for packet classification [10]. The study put forth the proposition of employing a combination of algorithms based on EC and spatio-temporal correlation. This approach took into account both multi-source and single-source heterogeneous data. S. Tanwar et al. proposed a structure for tracking hand movements of arthritis patients using EC. However, the structure's robustness was limited in processing large amounts of data [11]. To address the issue of poor processing of large amounts of data, the study introduced SW for data classification. Additionally, P. Kumar et al. proposed an anomalous intrusion detection system that is decentralized to local fog nodes with a cloud-based security architecture to tackle the problems of high false alarm rates (FARs) and low accuracy in traditional intrusion detection systems. This system reduced the FAR and improved detection accuracy, showing potential for application in modern network security. However, the proposed method in the study took into account the spatial correlation of outlier data, which this detection system does not [12].

The importance of AD in wireless sensor networks lies in ensuring data reliability and quality. L. Chen et al. proposed a method for AD that utilized spatio-temporal correlation and information entropy to address the underutilization of spatio-temporal correlation of sensor data. However, this method did not address the detection of outlier data, despite its significant improvement in reducing false positive and false negative rates [13]. For AD of outlier data, N. Berjab et al. proposed a cross-correlation-based method for acquiring spatio-temporal relationships of sensors. The method achieved better results on the validation of real-world datasets. However, the design of the method did not take into account the bandwidth consumption, and its detection of large amounts of data requires more bandwidth [14]. The framework combined spatio-temporal correlation algorithms with EC, effectively reducing the required bandwidth consumption for detection. The study proposed a dynamic framework for modeling spatio-temporal traffic data to diagnose anomalies and improve the quality of service of the current transportation system. The authors, X. Wang and L. Sun, aimed to use clear and concise language to explain their approach. A time-varying vector auto-regressive model was used to characterize the system dynamics and modeled with a low-rank tensor structure. However, while the method showed superiority in dynamically monitoring anomalous transportation networks, it lacked accuracy in monitoring anomalies in massive data [15]. H. Liu et al. proposed a collaborative intrusion detection mechanism based on distributed federated learning and secured the aggregated model using block-chain technology to manage the storage and sharing of the training model. However, the study aimed to enhance the security of the transportation network system and did not extensively investigate the effectiveness of anomaly data detection [16]. The study applied algorithms based on EC and spatio-temporal correlation to design a sensor monitoring system for greenhouse cultivation. This approach expanded the range of applications for AD algorithms. In addition, A. Haj-Hassan et al. proposed a framework for outlier detection for wireless human sensor networks. However, the method had some weaknesses for massive data [17].

Scholars have conducted extensive research on AD of network data. One area of focus is the detection of spatio-temporal data anomalies, which aims to enhance network system security and reliability [18], [19]. However, this research does not comprehensively consider factors such as network bandwidth and outlier data. Contrastingly, EC has been studied and applied in data AD for an extended period. The present study suggests a new SMM approach. The spatial correlation anomaly detection algorithm (SCADA) is further designed based on the proposed TCADA, and the SMM for smart greenhouses is constructed by integrating the two algorithms in EC. The study introduces SW innovatively as a multi-source data chunking method for data processing, which is different from other data detection models and effectively avoids the resultant error caused by one-sided estimation of single data. To demonstrate the feasibility of the proposed method, the study compares it with the AD methods reviewed in the previous section. The comparison results are presented in Table 1.

TABLE 1. Comparison of the proposed methodology of the study with the existing literature.

Research purpose	Research methods	Contributions	Reference
Inadequate utilization of spatio-temporal correlation of sensor data	Spatio-temporal correlation; information entropy; weighted coefficient of variation	False-positive and false-negative rates were significantly reduced	L. Chen et al [13]
Poor spatio-temporal correlation of heterogeneous sensors	Multivariate attributes; cross-correlation; SW; median absolute deviation	The average accuracy rate is 96.50%, the average precision rate is 88.69%, and the recall rate is 93.00%.	N. Berjab et al [14]
Enhancing the quality of service of current transportation systems	Time-varying vector auto-regressive model; low-rank tensor structure	Demonstrated superiority in dynamic monitoring of abnormal transportation networks	X. Wang et al [15]
To improve the effectiveness of data anomaly monitoring in transportation networks	Distributed federated learning; block-chain technology	Improving the safety of transportation network systems	H. Liu et al [16]
To improve the accurate detection of multi-source heterogeneous data anomalies and enhance multi-source data spatio-temporal correlation	EC; SW; spatio-temporal correlation	Aiming to improve the accuracy of anomaly detection of heterogeneous data from multiple sources while applying to sensor monitoring of smart greenhouses	This study

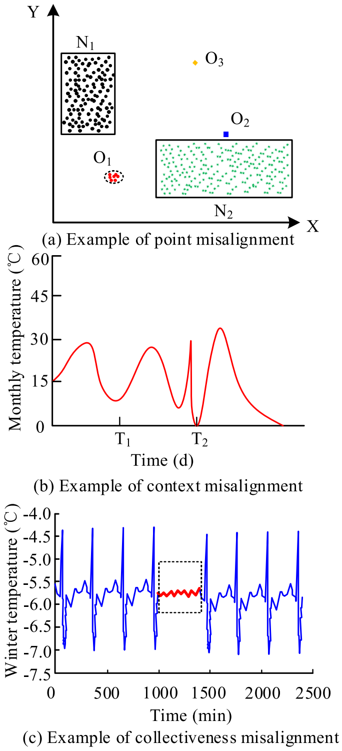
III. SENSOR MONITORING MODEL COMBINING S-TCADA AND EC

Current algorithms for anomaly data detection mostly focus on single-source data, ignoring the correlation of node space and the detection of heterogeneous data from multiple sources. Therefore, a SMM constructed using S-TCADA in

EC is studied and designed. Firstly, SW is introduced in TCADA for data chunking, and SCADA is further proposed based on this. Secondly, S-TCADA is fused to construct a sensor model in conjunction with EC, and it is applied to data malfunction monitoring in smart greenhouses with cloud platform technology and so on.

A. DESIGN BASED ON TCADA

Traditional AD methods ignore the relevant features of multi-source data when processing streaming data, which leads to the reduction of accuracy and efficiency. For this reason, the study proposes a TCADA for multi-source data, which determines the anomaly state by calculating its anomaly score for multi-source data at adjacent time and chunks the data using SW. Considering the heterogeneity of multi-source data, the study categorizes the scene malfunctions into point, context, and collective malfunctions, as shown in Figure 1.



(a) Example of point misalignment

(b) Example of context misalignment

(c) Example of collectiveness misalignment

FIGURE 1. Schematic diagram of three types of abnormal scenarios.

Point anomaly means that a single data is far away from other data of the same type, so it is judged as a point anomaly or belonging to an isolated point. In Figure 1(a), the points O$_1$, O$_2$, and O$_3$ are far away from the planes N$_1$ and N$_2$, which can be considered as belonging to point derangement. Contextual derangement is a situation where an individual instance is normal in other scenarios but appears deranged in a specific context. As shown in Figure 1(b), individual instances at moment t2 exhibit contextual malfunctions. In contrast, collective dissonance refers to a situation in which individual instances within a group collection appear to be discordant. This may occur even when the individual instances themselves are normal. However, when multiple instances within the collection exhibit dissonance, it is referred to as collective dissonance.

Based on the concept of scenario malfunction, the study uses three time-sequence to measure the data, namely, single-source data, multi-source heterogeneous data, and SW, and calculates the relative distances between data objects using the Euclidean distance equation [20], [21], [22]. Among them, single-source data are collected by a single sensor, which can be used to reflect the changes or development trends produced by things over time, and the specific expression equation is shown in equation (1).

$$A_i = \{a_1, a_2, \ldots, a_i, \ldots, a_n\} \quad (1 \leq i \leq n) \tag{1}$$

In equation (1), A_i denotes the set of single-source time-sequence, a_1 denotes the data collected at the 1st observation moment. a_i denotes the data collected at the i-th observation moment, a_n denotes the data collected at the nth observation moment. i denotes the ordinal number of the moment of observation. Moreover, n denotes the number of observations. Equation (2) is a simplified representation of multi-source heterogeneous data, which are gathered from many sensors. The data take the form of time-sequence data.

$$\begin{cases} TA_m = \{A_1, A_2, \ldots, A_i, \ldots, A_m\} \quad (1 \leq i \leq m) \\ A_i = \left\{a_1, a_2, \ldots, a_j, \ldots, a_m\right\} \end{cases} \tag{2}$$

In equation (2), TA_m denotes the time-sequence data set of multi-source sensing data, a_j denotes the data value collected at a specific moment. m denotes the number of data in the set, j denotes the ordinal number of the observation moment, but $j \neq i$. Meanwhile, the study innovatively introduces the time SW to chunk the data, and the expression of a SW data in a wireless sensor is shown in equation (3).

$$\begin{cases} W = \{a_{e+1}, a_{e+2}, \ldots, a_{e+i}, \ldots, a_{e+w}\} \quad (1 \leq i \leq w) \\ E = \{e_1, e_2, \ldots, e_i, \ldots, e_w\} \\ F = \{f_1, f_2, \ldots, f_i, \ldots, f_w\} \end{cases} \tag{3}$$

In equation (3), W denotes data from a SW in the sensor, and w denotes the length of the SW. E and F denote two sets of neighboring multidimensional data, respectively, and e_i and f_i denote a certain data value of the multidimensional data collected at a certain moment. Equation (3) is able to be used for statistical and monitoring purposes in practical situations, where time-sequence are localized and computed based on the length of the units formulated. Moreover, to enhance the calculation of data information, the study invokes the time SW for categorization. This SW dynamically adjusts the window size based on the characteristics of the data stream. In the event of rapid alterations in the data pattern, the window can be reduced in order to capture more recent data. Conversely, if the data is relatively stable, the window can be expanded in order to include more historical data. Concurrently, the study postulates that each time the data of the entire time-sequence is accessed, its sub-sequence will be calculated in accordance with the specifications delineated in equation (4).

$$A_{sub}^{t} = \{a_{t-w+1}, a_{t-w+2}, \ldots, a_t\} \tag{4}$$

In equation (4), A_{sub}^{t} denotes the sub-sequence of each visit. Therefore, the proposed TCADA process for heterogeneous data from multiple sources is shown in Figure 2.

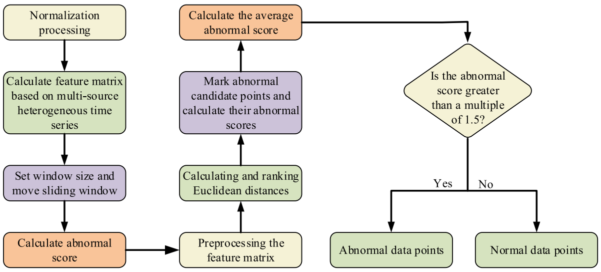

Normalization processing
Calculate feature matrix based on multi-source heterogeneous time series
Set window size and move sliding window
Calculate abnormal score
Preprocessing the feature matrix
Calculating and ranking Euclidean distances
Mark abnormal candidate points and calculate their abnormal scores
Calculate the average abnormal score
Is the abnormal score greater than a multiple of 1.5?
Yes
No
Abnormal data points
Normal data points

FIGURE 2. Algorithm flow for detecting time related anomalies.

In Figure 2, TCADA initially normalizes the data in order to mitigate the adverse impact of disparate scales on the data. Secondly, the time-sequence with multiple sources of heterogeneity are input, and the calculation of the feature matrix is performed when SW is 1. The specific calculation is shown in equation (5).

$$\begin{cases} D_{t(e,f)} = \dfrac{\sum_{i=0}^{w-1} a_{t-i}^{e} a_{t-i}^{f}}{w} \\ a^e = \left\{a_{t-w+1}^{e}, a_{t-w+2}^{e}, \ldots, a_t^{e}\right\} \\ a^f = \left\{a_{t-w+1}^{f}, a_{t-w+2}^{f}, \ldots, a_t^{f}\right\} \end{cases} \tag{5}$$

In equation (5), $D_{t(e,f)}$ denotes the characterization matrix of the e-th row and f-th column. a^e denotes the expression in row e and a^f denotes the expression in column f. In accordance with the aforementioned rationale, the window size setting is implemented, and SW shifting is conducted with the objective of obtaining the subsequent sequence at the next moment. The anomaly score is calculated by determining the degree of outlier distance shift of neighboring feature evidences. Concurrently, the feature matrix is subjected to pre-processing, and the distance calculation of neighboring nodes is performed. The nearest neighbor distance difference algorithm is a commonly used method for calculating data sets of various dimensions. The Euclidean distance is employed to calculate the distance formula between two neighboring nodes, as illustrated in Equation (6).

$$L(i,j) = \left\{\left|A_i - A_j\right|, 1 \le i \le n, 1 \le j \le n\right\} \tag{6}$$

In equation (6), $L(i,j)$ denotes the distance between two neighboring nodes. This nearest neighbor distance difference calculation equation is applicable to data sets of various dimensions. In this case, the expression equation for the dataset is shown in equation (7).

$$\begin{cases} A = \left\{a_i \left| a_i \in R^N, 1 \le i \le M\right.\right\} \\ a_i = \{a_{i1}, a_{i2}, \cdots, a_{iN}\} \end{cases} \tag{7}$$

In equation (7), M denotes the total data objects in the dataset, and N denotes the data dimensions. Furthermore, the study analyzed and ranked the degree of influence of the malfunction at this moment under each dimension on the basis of the malfunction judgment rule. The algorithm then ranked the malfunction scores in order from the largest to the smallest, thus facilitating the rapid localization of the sensors that may trigger the malfunction through the detected anomalies.

B. DESIGN BASED ON SCADA

According to [23], there is frequently a strong spatial correlation among the data gathered by nodes located in close proximity within the monitoring area. Therefore, with the goal to increase the accuracy and efficiency of AD, the study further proposes SCADA on the basis of TCADA, which analyzes the multi-sensor data from the same type of sensors according to the sensor locations. Spatial correlation refers to the existence of a quantitative functional relationship between the data values of nodes within a specific spatial range. This concept is particularly relevant in the context of sensor networks. Among them, it is essential to understand the degree of similarity in the data collected by sensor nodes that are located in close proximity to each other within the monitoring area. Based on the spatial correlation, the study classifies the spatial sequence data collected by the environmental monitoring of smart greenhouses into three categories, including edge nodes aggregated into multiple different types of unit sensor nodes, continuous time data flows formed by multiple edge nodes, and covariance and Pearson's coefficient [24], [25]. The study initially delineates the formation of an edge node, which is constituted by the aggregation of multiple unitary sensor nodes of disparate types. The multidimensional matrix on the basis of the time-sequence represented by its collected data is shown in equation (8).

$$B_i^j(t) = \begin{bmatrix} A_1^1 & A_1^2 & \cdots & A_1^j \\ A_2^1 & A_2^2 & \cdots & A_2^j \\ \vdots & \vdots & \cdots & \vdots \\ A_i^1 & A_i^2 & \cdots & A_i^j \end{bmatrix} (i, j \in [1, N]) \tag{8}$$

In equation (8), $B_i^j(t)$ denotes the data collected by the j-th sensor of the i-th edge node at the t-th moment. A_i^j denotes the time-sequence of the jth sensor of the i-th edge node at the t-th moment. This is followed by a continuous temporal data flow formed by multiple edge nodes, as shown in equation (9).

$$B_i^T = \begin{bmatrix} B_i^1(1) & B_i^1(2) & \cdots & B_i^1(T) \\ B_i^2(1) & B_i^2(2) & \cdots & B_i^2(T) \\ \cdots & \cdots & \cdots & \cdots \\ B_i^N(1) & B_i^N(2) & \cdots & B_i^N(T) \end{bmatrix} \tag{9}$$

In equation (9), B_i^T denotes a matrix of data flow over a period of time constituted by all the sensors of the i-th base station. According to the data flow equation, the other edge nodes of the shed are numbered in order to perform data collection, and the numbering equation is shown in equation (10).

$$\begin{cases} A = \{A1, A2, \cdots, An\} \\ A' = \left\{A'1, A'2, \cdots, A'n\right\} \\ A'' = \left\{A''1, A''2, \cdots, A''n\right\} \end{cases} \tag{10}$$

In equation (10), A denotes the data set of the i-th edge node. $A1$, $A2$, and An denote the 1-st, 2-nd, and n-th data of the i-th edge node, respectively. A' is the data set of the $i+1$-th edge node. $A'1, A'2$, and $A'n$ denote the 1-st, 2-nd, and n-th data of the $i+1$-th edge node, respectively. A'' is the data set of the $i+2$-th edge node. $A''1$, $A''2$, and $A''n$ denote the 1-st, 2-nd, and n-th data of the $i+1$-th edge node, respectively. Finally, the mean, standard deviation and other correlation coefficients of the data samples are calculated based on the covariance and Pearson's correlation coefficient, where the equation for mean, standard deviation and variance is shown in equation (11).

$$\begin{cases} \bar{A} = \sum_{i=1}^{n} a_i \\ S = \sqrt{\dfrac{\sum_{i=1}^{n}(a_i - \bar{A})^2}{n-1}} \\ S^2 = \dfrac{\sum_{i=1}^{n}(a_i - \bar{A})^2}{n-1} \end{cases} \tag{11}$$

In equation (11), $\bar{A}$ denotes the mean, S denotes the standard deviation of the sample data, and S^2 denotes the variance of the sample data. The equation for covariance between random variables is then shown in equation (12).

$$\begin{cases} Cov(X, Y) = \mathrm{E}\left\{[X - \mathrm{E}(X)]\,[Y - \mathrm{E}(Y)]\right\} \\ \rho_{x,y} = \dfrac{\mathrm{Cov}(X, Y)}{\sigma x \sigma y} = \dfrac{\mathrm{E}(x-\mu_x)(y-\mu_y)}{\sigma x \sigma y} \end{cases}. \tag{12}$$

In equation (12), $Cov(X, Y)$ denotes the Pearson correlation coefficient of random variables X and Y. $\rho_{x,y}$ denotes the correlation coefficient of random variables X and Y. $\mathrm{E}\left\{[X - \mathrm{E}(X)]\,[Y - \mathrm{E}(Y)]\right\}$ denotes the covariance of random variables X and Y. σx and σy denote the standard deviation of two random variables, respectively. For the nodes in the edge layer of the smart greenhouse, it is investigated to compute and analyze the correlation of the matrices of different edge nodes in the same time situation, which in turn achieves the result of alleviating the computational task. Therefore, the computational steps of the proposed TCADA-based SCADA are shown in Figure 3.

In Figure 3, the data is first normalized using min-max normalization and sorted according to the number of edge nodes. Secondly, abnormal node judgment is performed by comparing the correlation of two edge nodes. The vector similarity calculation of multi-source data is performed using the correlation coefficient. Where the normalization expression is shown in equation (13).

$$\bar{B}_i^j(t) = \frac{B_i^j(t) - \min(B_i^j)}{\max(B_i^j) - \min(B_i^j)} \tag{13}$$

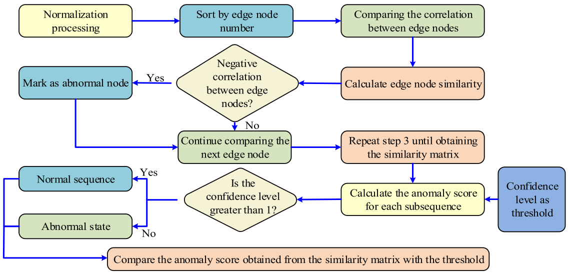

Normalization processing
Sort by edge node number
Comparing the correlation between edge nodes
Mark as abnormal node
Yes
Negative correlation between edge nodes?
Calculate edge node similarity
No
Continue comparing the next edge node
Repeat step 3 until obtaining the similarity matrix
Normal sequence
Yes
Is the confidence level greater than 1?
Calculate the anomaly score for each subsequence
Confidence level as threshold
Abnormal state
No
Compare the anomaly score obtained from the similarity matrix with the threshold

FIGURE 3. Steps of spatial correlation-based anomaly detection algorithm.

In equation (13), $\max(B_i^j)$ and $\min(B_i^j)$ denote the maximum and minimum values of the monitored data, respectively. SCADA proceeds to the next edge node after marking the edge node as an anomaly. On the basis of this, a similarity matrix $C_{i,j}$ is computed from the two N-dimensional matrices, which in turn solves for the anomaly score of a certain sub-sequence. The specific calculation formula is presented in equation (14).

$$\begin{cases} H_i = \dfrac{\frac{\sum_{j=1}^{N} C_{i,j}}{N}}{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}}{N^2}} = N \cdot \dfrac{\sum_{j=1}^{N} C_{i,j}}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}} \\ C_{i,j} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1j} \\ C_{21} & C_{22} & \cdots & C_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ C_{i1} & C_{i2} & \cdots & C_{ij} \end{bmatrix} \end{cases} \tag{14}$$

In equation (14), $C_{i,j}$ denotes the similarity matrix of the j-th sensor of the i-th edge node. H_i denotes the anomaly score of the i-th sub-sequence. $\sum_{j=1}^{N} C_{i,j}$ denotes the average value of each row element in the similarity matrix and $\frac{\sum_{j=1}^{N} C_{i,j}}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}}$ denotes the average value of all elements in the similarity matrix. Considering the correlation between the derangement score and the detection result, the study introduced a threshold value as the detection range of derangement and utilized the confidence level (CL) as the measurement criterion for derangement detection [26], [27], [28].The specific expression equation of CL is shown in equation (15).

$$CL = \frac{H_i}{\frac{\sum_{i=1}^{N} H_i}{N}} \tag{15}$$

In equation (15), $\frac{\sum_{i=1}^{N} H_i}{N}$ represents the average of the anomaly scores of the whole sub-sequence. According to equation (15), the anomaly score of the anomalous sub-sequence is greater than that of the normal sub-sequence. The anomaly score of the overall time-sequence is equal to 1. Therefore, the confidence CL of the normal sub-sequence should be close to 1. Combining the above, the SCADA obtained on the basis of the TCADA consists of two parts: the algorithm for generating similarity matrices and the AD. In this case, the metric equation for evaluating the AD is shown in equation (16) [29], [30].

$$\begin{cases} Sen = \dfrac{TD}{TP} \times 100\% \\ Spe = \dfrac{FP}{G - TD} \times 100\% \end{cases} \tag{16}$$

In equation (16), *Sen* denotes the detection rate, whose value is equal to 1 means that all abnormal patterns have been successfully detected. *Spe* denotes the FAR, whose smaller value indicates that the accuracy of the detection method is higher. The numbers *TD* and *TP* represent the number of anomalies found experimentally and appropriately, respectively. The numbers *FP* and *G* represent the number of regular sub-sequences and the number of sub-sequences that are found to be anomalous.

C. SENSOR MONITORING MODEL CONSTRUCTION COMBINING S-TECADA AND EC

In order to improve the detection accuracy of massive heterogeneous monitoring data, the study integrates TCADA and SCADA, forming S-TCADA. Moreover, the specific steps are shown in Figure 4.

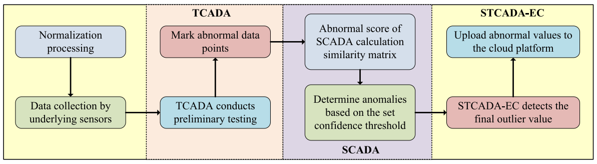

TCADA
STCADA-EC
Normalization processing
Data collection by underlying sensors
Mark abnormal data points
TCADA conducts preliminary testing
Abnormal score of SCADA calculation similarity matrix
Determine anomalies based on the set confidence threshold
SCADA
Upload abnormal values to the cloud platform
STCADA-EC detects the final outlier value

FIGURE 4. Algorithm for anomaly detection-based on spatio-temporal correlation.

As illustrated in Figure 4, the S-TCADA algorithm initially gathers multi-source data through the nodes of the underlying sensors and performs normalization and pre-processing. Secondly, the underlying sensor nodes transmit the collected time-sequence multi-source heterogeneous data. TCADA conducts preliminary monitoring of the multi-source heterogeneous data, calculates the outlier distance of each node within the SW, and filters out the nodes with anomaly scores more than 1.5 times of the mean as the candidate anomalous nodes, and labels them. SCADA derives the anomaly scores by calculating the similarity of the multi-source heterogeneous data of different edge nodes. Subsequently, a threshold comparison is conducted. Then, S-TCADA identifies any anomalies and transmits them to the cloud platform. At the same time, the study combines the hierarchical EC to construct a multi-source and multi-dimensional data AD model for intelligent greenhouse spatio-temporal correlation anomaly detection algorithm-edge computing (S-TCADA-EC), which is capable of realizing the functions of collecting, detecting and transmitting the design of environmental monitoring data for greenhouses. This AD model consists of three parts: remote cloud platform, edge layer nodes and sensing layer nodes, and the specific architecture is shown in Figure 5.

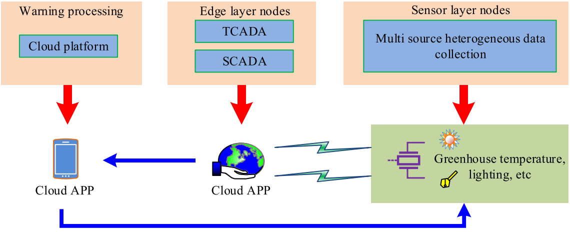


FIGURE 5. Overall architecture of multi-source and multi-dimensional data malfunction detection modeling.

In Figure 5, the model is mainly oriented to intelligent greenhouses. Among them, the sensor nodes are mainly responsible for the acquisition of multi-source heterogeneous data such as temperature, humidity, light, and multi-dimensional heterogeneous data. The edge layer nodes act as the core of EC, and the study embeds the proposed S-TCADA into the edge nodes for spatial correlation AD of multi-source data, while the edge nodes utilize time-sequence continuity and node spatial location correlation for the detection of out-of-whack data appearing in real-time sensing data. The edge layer is connected to the sensing layer plus points downward and interfaces with the cloud platform upward. The sensing layer, which is at the base of the model, uses little power to gather data from the field devices. It is fitted with sensors that measure things like light, temperature, humidity, carbon dioxide, and other variables. It is the cloud platform's responsibility to gather edge layer feedback. The edge layer supplies data support for cloud services, and the cloud platform is in charge of gathering the AD results from there. Figure 6 shows the intelligent greenhouse monitoring system. Its system hardware uses an STM32 chip, and the LCD display and WiFi connection are on this chip.

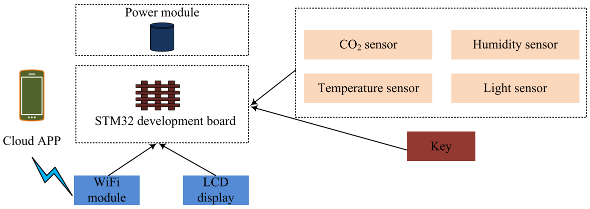


FIGURE 6. Specific hardware structure framework.

Figure 6 depicts a schematic diagram of the hardware structural framework of an edge layer node in a smart greenhouse. The edge layer node may be regarded as a small server, comprising multiple distributed potential network administrators or servers. The edge layer node represents the primary site for AD and serves as the central element in the model that integrates the S-TCADA with EC. The AD sequence delineates the stages of the AD task, which are depicted in Figure 7.

As illustrated in Figure 7, the AD process encompasses the acquisition of data, storage of data, analysis of anomalies, and provision of feedback regarding anomalies. Upon transmission of the multi-source heterogeneous data collected by the sensor layer node to the edge layer node, the latter will immediately analyze the multi-source data for anomalies and

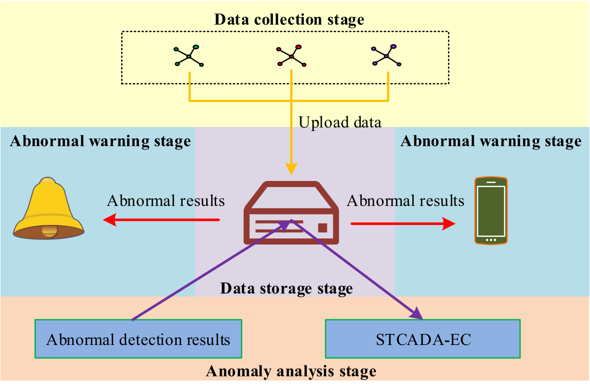


FIGURE 7. Abnormal warning structure.

provide the user with the detection results via the cloud platform. In the malfunction feedback stage, users can view the data collected by the sensors in real time, as well as the warning information fed back by the edge layer node after analysis and processing of the data. Furthermore, the study employs precision rate, recall rate, and F1-score as evaluation metrics for the algorithm. The expression formula for precision rate is provided in Equation (17).

$$precision = \frac{TP}{TP + FP} \tag{17}$$

The expression for the recall rate is shown in equation (18).

$$recall = \frac{TP}{TP + FN} \tag{18}$$

In equation (18), FN denotes the number of data that are actually normal but are detected as out of order. The formula for calculating the F1-score is shown in equation (19).

$$F1 = \frac{2\, precision \cdot recall}{precision + recall} \tag{19}$$

IV. EXPERIMENTAL ANALYSIS OF SENSOR MONITORING COMBINING S-TCADA AND EC

The study is conducted to validate and analyze the method with multi-source and multi-dimension data collected from greenhouses. Firstly, the value of SW is determined with test metrics such as precision and recall, and the performance of TCADA is evaluated. Secondly, the SW size of SCADA is further determined, and the algorithms are compared in terms of performance metrics. Finally, based on the proposed application process of STCADA-EC in sensor monitoring, other methods are introduced for comparison to verify the performance of STCADA-EC.

A. EXPERIMENTAL SETUP

The experimental data are derived from the actual environmental data of a smart agriculture experimental garden. The data attributes encompass temperature, humidity, light, greenhouse wind speed, soil pH value, and other indicators. The study set the greenhouse sensors to collect the data of the above indicators every 10 min, 10 edge nodes for data collection, each edge node to monitor a greenhouse, and the collection time is 24 h. Based on the information from the collected data, the study validates AD for three variables: temperature, humidity, and light in the greenhouse. Some of the characteristics of the dataset are shown in Table 2.

TABLE 2. Data characteristics of selected datasets.

Time	Temperature	Humidity	Sunlight light
10:00	29.3	38.5	43.2
10:10	29.5	38.6	43.6
10:20	29.6	39.0	43.9
10:30	30.1	39.9	42.3
10:40	34.5	40.5	43.6
10:50	30.6	40.2	44.1
11:00	31.2	40.3	44.4
…	…	…	…

The data characteristics of some real datasets are shown in Table 2. In the whole simulation experiment, the amount of validated data is 5000, the sensors is 150, the first 3000 data records are training data, and the last 2000 data records are test data. Based on this, the study introduces the K-nearest neighbor (KNN) algorithm [31], fuzzy-theoretic algorithm (FTA) [32], random forest algorithm [33], XGBoost [34], LightGBM [35], and deep learning algorithm auto-encoder [36]. Since the experimental validation data in the proposed process of the comparison methods are different, the study set the above comparison methods are compared using the greenhouse data collected by the study. Meanwhile, K-fold cross-validation is used to confirm the hyper-parameters of KNN algorithm. Cosine similarity is used to calculate the similarity of FTA. Grid search cross-validation is used for the parameter optimization of random forest algorithm. Moreover, the hyper-parameters of XGBoost and LightGBM are decision tree and control tree structure respectively. The hyper-parameters of the above machine learning comparison methods are determined through the steps of segmenting the data, training and evaluating the models, calculating the performance metrics, analyzing the results, selecting the best model, final testing, and reporting the results. The input format files for all machine learning models are comma-separated values. The input features are of continuous numerical type and are normally distributed. The model training loss function is cross Entropy loss. In addition, for the KNN and random forest algorithms do not have the function of unbalanced dataset processing, the study uses the evaluation index F1-score for data unbalance processing. To ensure the robustness of the proposed method in practical scenarios, stress tests including high-load testing, diversity testing, anomaly pattern change testing, and adversarial attack testing are conducted before the experimental validation.

Among them, the hyper-parameters of KNN algorithm are set as follows: n_neighbors=5, weights=10, and P=1. The parameters related to Random Forest algorithm are set as follows: max_depth=5 and n_estimators=300. The parameters related to XGBoost are set as follows: max_depth=5, min_child_ weight=3, gamma=0, subsample=0.85, colsample_bytree=0.9, reg-alpha=0.1, learning_rate=0.1, and n_estimators=0.1. Weight=3, gamma=0, subsample=0.85, colsample_bytree=0.9, reg-alpha=0.1, learning_rate=0.1, n_estimators=301, and objective='binary: logistic'. LightGBM related parameters are set as follows: max_depth=7, num_leaves=80, min_child_samples=20, min_child_weight=0.001, colsample_bytree=0.8, reg-alpha=0.1, learning_rate=0.1, n_estimators=301, and objective='binary: logistic'. Bytree=0.8, subsample=0.6, learning_rate=0.1, n_estimators=230, boosting_type="gbdt", and objective='binary: logistic'.

B. EXPERIMENTAL VALIDATION BASED ON TCADA

According to the proposed TCADA algorithm, the study experimentally verified it with precision rate, recall rate and F1 score. Since the value of the SW will directly affect the running time of the algorithm as well as the evaluation results, the study compares and analyses the precision rate, recall rate and F1 score of the algorithm by setting different SWs. The specific experimental results are shown in Figure 8.

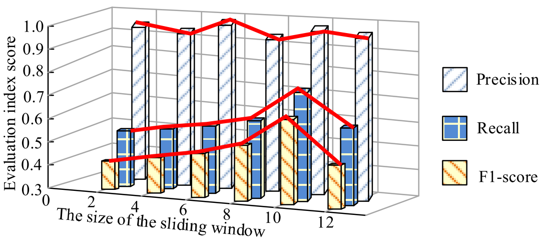
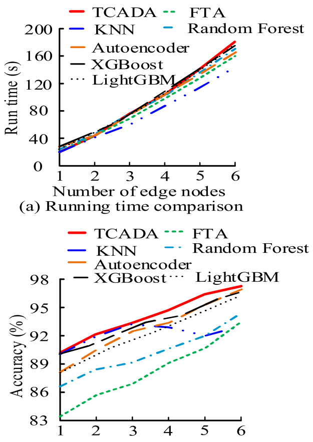


FIGURE 8. The impact of different sliding window sizes on algorithm performance evaluation metrics.

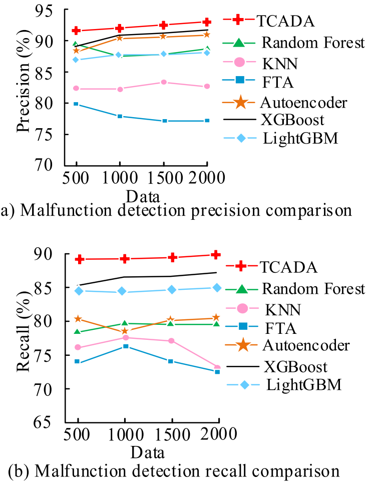
In Figure 8, the precision rate varies less at different SW sizes, while the recall rate and F1-score fluctuate more overall. When the SW is 10, both the recall and F1-scores are at their highest values with the best results. Therefore, the study set the size of SW as 10. Firstly, the original data set is initialized, and the feature matrix of the input multi-source heterogeneous time-sequence is extracted. The window size is then set to 10. Secondly, a window shift is performed to obtain a sub-sequence of the subsequent moment. Anomalous scoring points are then determined based on the leptokurtic distances of the neighboring matrices. Meanwhile, in order to further validate the performance of the TCADA algorithm proposed in the study, the performance of other methods is compared with the proposed method. The specific comparison results are shown in Figure 9.

In Figure 9(a), when the number of edge nodes is small, the difference in the running time of several algorithms is minimal. However, as the amount of data increases, the running time of several algorithms increases to varying degrees, with TCADA's running time being higher than that of the other four methods at six edge nodes. This may be due to the fact that the double computation of data by the introduced SW causes it to take too much time to run. However, TCADA exerts a significant advantage in terms of market detection accuracy. The accuracy gap with KNN is smaller when the amount of data is smaller, but it is still better than the other methods. Moreover, as the amount of data increases, the detection ability of TCADA is even more superior. When the edge node is 6, TCADA increases the accuracy by 0.27%-4.89% over the other four methods. This suggests that the runtime due to the introduction of the SW is worthwhile in the case of improved detection accuracy. Therefore, it can be assumed that when the detection accuracy is improved, the increase in running time does not affect the detection performance of the algorithm. The study ran multiple trials of the five approaches on a bigger volume of data in order to more thoroughly examine the detection performance of TCADA. The final results of the comparison of precision and recall are shown in Figure 10.



(a) Running time comparison

FIGURE 9. Comparison of runtime and accuracy of different algorithms.



(a) Malfunction detection precision comparison

(b) Malfunction detection recall comparison

FIGURE 10. Comparison of precision and recall of different algorithms.

As the data volume increases, the accuracy rates of several algorithms at different data volumes show different trends, but TCADA is overall higher than the other six algorithms. When the data volume is 2000, the precision rate of TCADA is as high as 92.58%, which is 18.74% more than FTA. Comparison of recall rates for different data volumes shows

TABLE 3. Comparison results of abnormal detection using different algorithms.

Performance index (%)	Algorithm model						
	TCADA	KNN	FTA	Random Forest	Autoencoder	XGBoost	LightGBM
Precision	92.58	82.98	77.97	88.95	91.12	90.67	89.89
Recall	89.97	73.74	73.51	79.86	80.78	89.40	88.72
F1-score	91.26	78.09	75.67	84.16	85.64	89.73	89.30

that FTA is significantly lower than the other six algorithms, while the proposed TCADA recall rate of the study is still superior to the other methods. Table 3 compares the precision, recall, and F1-scores of the seven methods at a data volume of 2000 in order to allow for a more natural comparison of the detection outcomes of the algorithms.

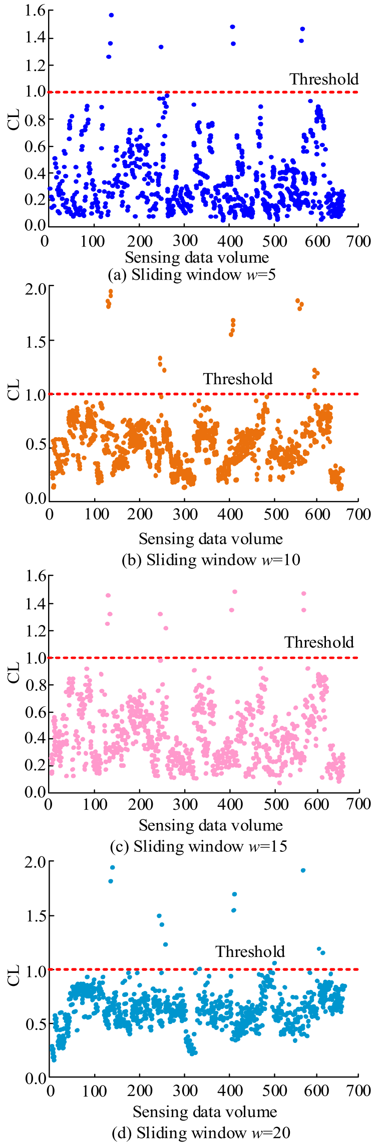
In the comparison of precision, recall and F1-score of the seven algorithms, it can be noted that TCADA has the best AD effect, and the F1-score is increased by 6.56%-20.60% compared to the other six algorithms. The KNN algorithm is ranked worse among the seven algorithms, probably because it calculates the malfunctioning score by the distance between the current data to the K-th nearest neighbor, and therefore the detection effect is not ideal. Moreover, FTA is based on fuzzy theory, and its effect is the worst among all algorithms. The unsupervised model's AD effect is noticeably worse than the deep model's, and when it comes to handling the anomaly problem of heterogeneous data from numerous sources, TCADA performs somewhat better than the other six algorithms.

C. EXPERIMENTAL VALIDATION BASED ON TCADA

In order to make further judgment on the proposed TCADA, the study first validates the performance of SCADA based on TCADA. Since the size of SW affects the detection results of the algorithm and it is known that TCADA performs best at a SW of 10, the study further compares the effectiveness of SCADA for AD with different SW sizes and different data volumes, as shown in Figure 11.

In Figure 11, combined with the set threshold CL=1.0 it can be observed that there are data points exceeding the threshold line at SW sizes of 5, 10, 15 and 20, with the most points exceeding the CL at w=10. This indicates that SCADA has the best detection at a SW size of 10. Figure 12 displays the specific detection results for different SW. The results include the number of false detections (NFD), number of false alarms (NFA), False Detection Rate (FDR), and FAR. It is evident that the FDR and FAR are the lowest when the SW is 10, while the AD accuracy is as high as 80%. Thus, the validation experiment analysis of SCADA still employs a SW size of 10.

Meanwhile, the study further compares the variation of SCADA performance at different CL levels when SW is 10. The details are shown in Table 4.



(a) Sliding window w=5

(b) Sliding window w=10

(c) Sliding window w=15

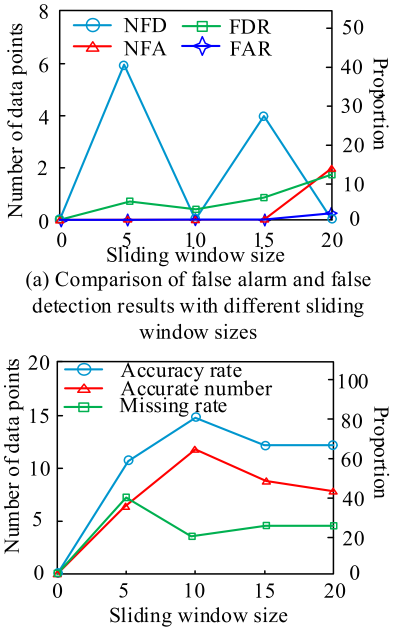
(d) Sliding window w=20

FIGURE 11. The effect of different sliding window sizes on abnormal detection performance.

TABLE 4. Effect of different CL levels on SCADA performance at SW=10.

Detection indicators	CL level					
	0.0	0.2	0.4	0.6	0.8	1.0
FDR (%)	22.12	20.01	17.61	15.50	12.87	6.87
FAR (%)	16.78	14.32	10.09	8.87	4.21	0.00

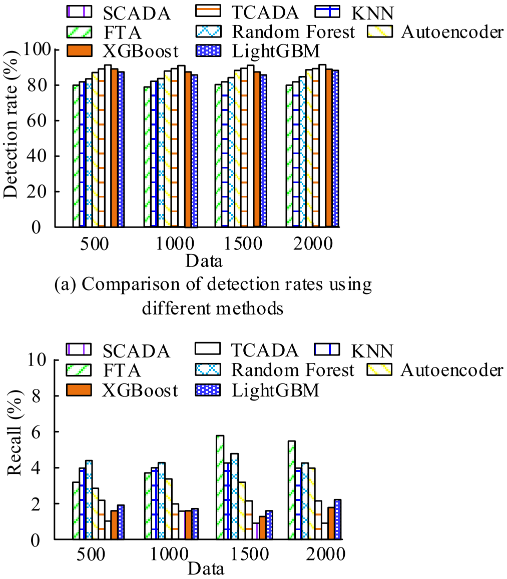
In Table 4, FDR and FAR of SCADA differed at different CL levels. As the CL level increases, both FDR and FAR tend to decrease. When CL takes the value of 1.0, the FDR of the algorithm is 6.87% and FAR is 0%. This indicates that SCADA is most effective for AD when CL takes the value of 1.0. Therefore, the study conducts subsequent validation experiments with SW set to 10 and CL set to 1. Different parameters have a certain impact on the AD results of the algorithm, and appropriate parameter settings are conducive to giving full play to the detection effect of the algorithm.



(a) Comparison of false alarm and false detection results with different sliding window sizes

(b) Comparison of accuracy and miss detection results with different sliding window sizes

FIGURE 12. Detection results of different sliding window sizes.



(a) Comparison of detection rates using different methods

(b) Comparison of false alarm rates among different methods

FIGURE 13. Comparison of detection rate and false alarm rate between different methods.

As a result, the study evaluates the detection capabilities of SCADA in comparison to other methods. Figure 13 displays the findings of the comparison of the detection rate and FAR under various data volumes.

As a further improved algorithm based on TCADA, SCADA demonstrates a gradual leveling off of its detection rate as the data volume increases. At both data volumes of 1500 and 2000, the detection rate reaches 94.58%, which is

TABLE 5. Comparison of detection performance between different methods.

Performance index (%)	Algorithm model							
	SCADA	TCADA	KNN	FTA	Random Forest	Autoencoder	XGBoost	LightGBM
AD	94.58	92.26	83.96	82.15	86.37	89.80	92.44	91.75
FAR	0.97	2.23	3.94	5.86	4.25	3.96	1.74	2.42
F1-score	92.51	91.26	78.09	75.67	84.16	85.64	90.21	90.03

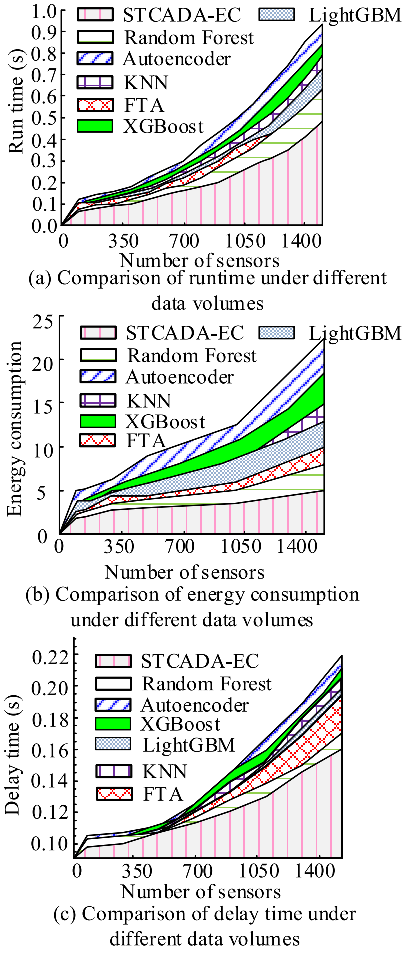
significantly better than several other methods. When the data volume is 2000, the final detection rate of SCADA increases by 2.51% over TCADA and 5.32%-15.13% over several other algorithms. FAR comparison results show that SCADA's FAR shows a large increase when the data volume is 1000, with a final FAR of 0.97% and TCADA's FAR of 2.23%. In addition, as the amount of data increases, the FTA FAR is the highest among all the methods, followed by Random Forest. This indicates that the proposed SCADA further reduces the FAR for the detection of out-of-order data on the basis of temporal correlation, and it also shows that the proposed SCADA has better detection capability. The results of the comparison of detection rate, FAR, and F1-score of the six methods are shown in Table 5.

The FAR of TCADA is reduced by 44.25%-83.45% compared to several other methods, while the F1-score is increased by 1.37%-22.25%, which indicates that TCADA is more effective in the detection of anomalous data. It can be noted that algorithmic improvement based on TCADA using spatial correlation is beneficial to the optimization of AD results, and the detection rate and F1-score of SCADA are increased to different degrees compared with TCADA, while the FAR is significantly reduced by 56.50%. Therefore, the proposed SCADA based on temporal correlation effectively improves the detection accuracy of the anomaly data and significantly reduces the FAR, which is superior to the detection performance of other algorithms.

D. EXPERIMENTAL VALIDATION OF A SENSOR MONITORING MODEL COMBINING S-TCADA AND EC

Finally, in order to realize STCADA for real-time anomaly warning during sensor acquisition of multidimensional time-sequence data, the study further compares the detection performance of S-TCADA-EC with KNN, FTA, random forest, XGBoost, LightGBM, and auto-encoder for different number of sensors.

The detection runtime of all five techniques in Figure 14(a) is less than 0.15s when the number of sensors is minimal, and the seven methods' needed runtime increases linearly with the number of sensors, which causes a huge increase in data. Among them, the auto-encoder algorithm takes the most time with the huge amount of data, while the study proposed S-TCADA-EC has the lowest runtime among all the methods. The configuration of the hardware affects the energy



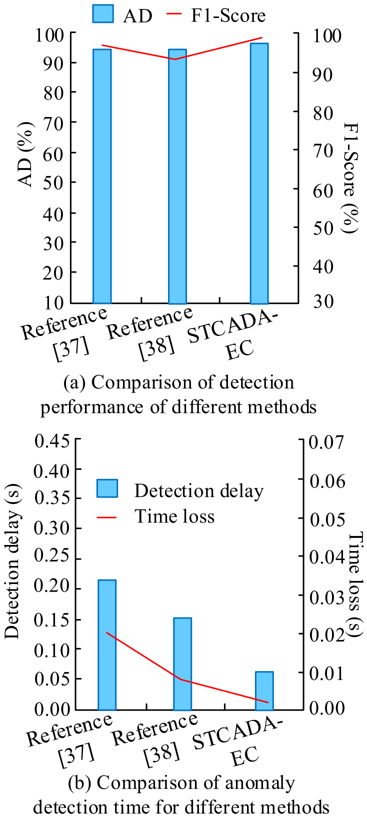
(a) Comparison of runtime under different data volumes

(b) Comparison of energy consumption under different data volumes

(c) Comparison of delay time under different data volumes

FIGURE 14. Performance comparison of different methods.

consumption of the algorithms during the runtime. From Figure 14(b), S-TCADA-EC requires the lowest energy consumption among all the methods, whereas auto-encoder and other auto-encoder algorithms utilizing neural network-based algorithms consume the most energy. This indicates that the combination of EC and S-TCADA is beneficial in reducing the energy consumption, especially for AD with large amounts of data. In addition, the amounts of data has a greater impact on the delay time of the algorithm running process, while the traditional detection methods have some advantages in the environment of smaller amount of data. However, in Figure 14(c), as the amount of data increases, that algorithms such as S-TCADA-EC and random forest have less delay time. The proposed S-TCADA-EC shows a significant advantage in delay time comparison, which may be due to the fact that the computational tasks are distributed to multiple edge nodes with distributed features, which improves the operation efficiency of the cloud platform to a certain extent, and thus the time delay is better. The prolongation time of methods such as auto-encoder, KNN, and so on, is too long, which may be due to the increase in the amount of data, which leads to an increase in their running time, and accordingly the delay time grows. Meanwhile, the latest SMM proposed by current scholars has been introduced to compare the performance with STCADA-EC. The specific comparison results are shown in Figure 15 [37], [38].



(a) Comparison of detection performance of different methods

(b) Comparison of anomaly detection time for different methods

FIGURE 15. Comparison of STCADA-EC performance with other methods.

In Figure 15(a), STCADA-EC has a significant advantage in AD compared to other anomaly monitoring models. Compared with literature [37] and literature [38], the F1-score of STCADA-EC increases by 4.58% on average. In Figure 15(b), a comparison of the detection latency and time loss under the three models reveals that STCADA-EC is the most advantageous. The literature [37] method exhibits a greater degree of delay and time loss than the other two methods. This is likely due to the fact that the method employs a probabilistic model for the detection of anomalous nodes, followed by the use of a discrete-time Markov chain for further analysis. This detection process takes more computational time and overhead. Overall, the research proposed STCADA-EC has significant advantages in greenhouse AD data. Compared with other methods, the proposed method has superior performance, which further indirectly verifies the application performance of this method in the benchmark dataset.

V. CONCLUSION

To achieve real-time detection of anomalous data in large-scale sensor data using a centralized cloud platform, the study proposes TCADA for detecting time-sequence of heterogeneous data from multiple sources. For the spatio-temporal correlation of sensor data, SCADA is designed. Based on this, the study constructs the AD model

STCADA-EC by introducing the hierarchical EC model and S-TCADA. The validation showed that the TCADA algorithm achieves a precision rate of 92.58% and a recall rate of 89.97% across various data volumes. The FAR of the SCADA algorithm was reduced by 56.50% compared to TCADA. Additionally, the STCADA-EC runtime was reduced by 14.54%-51.55% compared to other methods. The study's results indicated that the AD method proposed in the study can effectively improve the accuracy of AD for large-scale data and has more significant advantages in AD for multi-source heterogeneous data.

However, there are still some shortcomings in the study and some outliers in the real collected data still exist. Prior to analyzing the data, the study performed thorough data cleaning, including removing or correcting obvious input errors, missing values, and outliers. The use of semi-supervised learning methods can utilize unlabeled data to improve the performance of AD. Therefore, the efficiency of automatic selection of model SW size will be further updated in the future to build more adaptive mode SW. Neural network algorithms and others are considered to be invoked for data training, including techniques such as convolutional neural networks, gated recurrent units, and graph neural networks. It is possible to integrate neural networks into the data preprocessing process, thereby enabling automatic learning of features. Alternatively, neural networks can be trained end-to-end with other components, such as feature extractors.

ABBREVIATIONS

Abbreviations	Full name
IoT	Internet of Things.
EC	Edge Computing.
S-TCADA	Spatio-Temporal Correlation Anomaly Detection Algorithm.
TCADA	Temporal Correlation Anomaly Detection Algorithm.
SW	Sliding Window.
AD	Anomaly Detection.
SCADA	Spatial Correlation Anomaly Detection Algorithm.
CL	Confidence Level.
S-TCADA-EC	Spatio-Temporal Correlation Anomaly Detection Algorithm-Edge Computing.
KNN	K-Nearest Neighbor.
FTA	Fuzzy-Theoretic Algorithm.
NFD	Number of False Detections.
NFA	Number of False Alarms.
FDR	False Detection Rate.
FAR	False Alarm Rate.

REFERENCES

[1] V. Chauhan and S. Soni, "Energy aware unequal clustering algorithm with multi-hop routing via low degree relay nodes for wireless sensor networks," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 2469–2482, Feb. 2021, doi: 10.1007/s12652-020-02385-1.

[2] S. N. G. Aryavalli and G. H. Kumar, "Futuristic vigilance: Empowering chipko movement with cyber-savvy IoT to safeguard forests," *Arch. Adv. Eng. Sci.*, vol. 1, no. 8, pp. 1–16, Sep. 2023, doi: 10.47852/bonviewaaes32021480.

[3] K. Jaiswal and V. Anand, "A grey-wolf based optimized clustering approach to improve QoS in wireless sensor networks for IoT applications," *Peer-to-Peer Netw. Appl.*, vol. 14, no. 4, pp. 1943–1962, Apr. 2021, doi: 10.1007/s12083-021-01099-1.

[4] T. Ge and O. Darcy, "Study on the design of interactive distance multimedia teaching system based on VR technology," *Int. J. Continuing Eng. Educ. Life-Long Learn.*, vol. 32, no. 1, p. 65, Mar. 2022, doi: 10.1504/ijceell.2022.121221.

[5] S. Wu, X. Li, W. Dong, S. Wang, X. Zhang, and Z. Xu, "Multi-source and heterogeneous marine hydrometeorology spatio-temporal data analysis with machine learning: A survey," *World Wide Web*, vol. 26, no. 3, pp. 1115–1156, May 2023, doi: 10.1007/s11280-022-01069-4.

[6] S. Evangelatos and A. L. Moustakas, "Detection of transmission state of multiple wireless sources: A statistical mechanics approach," *Telecom*, vol. 4, no. 3, pp. 649–677, Sep. 2023, doi: 10.3390/telecom4030029.

[7] M. Taneja, N. Jalodia, J. Byabazaire, A. Davy, and C. Olariu, "SmartHerd management: A microservices-based fog computing-assisted IoT platform towards data-driven smart dairy farming," *Softw. Pract. Exper.*, vol. 49, no. 7, pp. 1055–1078, Jul. 2019, doi: 10.1002/spe.2704.

[8] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, "PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3310–3322, Mar. 2021, doi: 10.1109/JIOT.2020.3023126.

[9] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. Liu, "Big data cleaning based on mobile edge computing in industrial sensor-cloud," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1321–1329, Feb. 2020, doi: 10.1109/TII.2019.2938861.

[10] K. Sadaf and J. Sultana, "Intrusion detection based on autoencoder and isolation forest in fog computing," *IEEE Access*, vol. 8, pp. 167059–167068, 2020, doi: 10.1109/ACCESS.2020.3022855.

[11] S. Tanwar, J. Vora, S. Kaneriya, S. Tyagi, N. Kumar, V. Sharma, and I. You, "Human arthritis analysis in fog computing environment using Bayesian network classifier and thread protocol," *IEEE Consum. Electron. Mag.*, vol. 9, no. 1, pp. 88–94, Jan. 2020, doi: 10.1109/MCE.2019.2941456.

[12] P. Kumar, G. P. Gupta, and R. Tripathi, "Design of anomaly-based intrusion detection system using fog computing for IoT network," *Autom. Control Comput. Sci.*, vol. 55, no. 2, pp. 137–147, May 2021, doi: 10.3103/s0146411621020085.

[13] L. Chen, L. Xu, and G. Li, "Anomaly detection using spatio-temporal correlation and information entropy in wireless sensor networks," in *Proc. Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData) IEEE Congr. Cybermatics (Cybermatics)*, Nov. 2020, pp. 121–128, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00037.

[14] N. Berjab, H. H. Le, and H. Yokota, "A spatiotemporal and multivariate attribute correlation extraction scheme for detecting abnormal nodes in WSNs," *IEEE Access*, vol. 9, pp. 135266–135284, 2021, doi: 10.1109/ACCESS.2021.3115819.

[15] X. Wang and L. Sun, "Diagnosing spatiotemporal traffic anomalies with low-rank tensor autoregression," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7904–7913, Dec. 2021, doi: 10.1109/TITS.2020.3044466.

[16] H. Liu, S. Zhang, P. Zhang, X. Zhou, X. Shao, G. Pu, and Y. Zhang, "Blockchain and federated learning for collaborative intrusion detection in vehicular edge computing," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 6073–6084, Jun. 2021, doi: 10.1109/TVT.2021.3076780.

[17] A. Haj-Hassan, C. Habib, and J. Nassar, "Real-time spatio-temporal based outlier detection framework for wireless body sensor networks," in *Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2020, pp. 1–6, doi: 10.1109/ANTS50601.2020.9342827.

[18] N. Berjab, H. H. Le, C.-M. Yu, S.-Y. Kuo, and H. Yokota, "Abnormal-node detection based on spatio-temporal and multivariate-attribute correlation in wireless sensor networks," in *Proc. IEEE 16th Int. Conf. Dependable, Autonomic Secure Comput., 16th Int. Conf. Pervasive Intell. Comput., 4th Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2018, pp. 568–575, doi: 10.1109/DASC/PiCom/DataCom/CYBERSCITEC.2018.00106.

[19] M. Zhao, H. Takizawa, and T. Soma, ''Spatiotemporal anomaly detection for large-scale sensor data,'' in *Proc. 12th Int. Symp. Parallel Architectures, Algorithms Program. (PAAP)*, Dec. 2021, pp. 162–168, doi: 10.1109/PAAP54281.2021.9720310.

[20] J. Leukel, J. González, and M. Riekert, ''Machine learning-based failure prediction in industrial maintenance: Improving performance by sliding window selection,'' *Int. J. Quality Rel. Manage.*, vol. 40, no. 6, pp. 1449–1462, May 2023, doi: 10.1108/ijqrm-12-2021-0439.

[21] S. Wijuniamurti, S. Nugroho, and R. Rachmawati, ''Agglomerative nesting (AGNES) method and divisive analysis (DIANA) method for hierarchical clustering on some distance measurement concepts,'' *J. Statist. Data Sci.*, vol. 1, no. 1, pp. 7–11, Mar. 2022, doi: 10.33369/jsds.v1i1.21009.

[22] A. C. Bailey, M. Vincenzi, D. Scolnic, J.-C. Cuillandre, J. Rhodes, I. Hook, E. R. Peterson, and B. Popovic, ''Type Ia supernova observations combining data from the *Euclid* mission and the Vera C. Rubin Observatory,'' *Monthly Notices Roy. Astronomical Soc.*, vol. 524, no. 4, pp. 5432–5441, Jul. 2023, doi: 10.1093/mnras/stad2179.

[23] Z. Ma, G. Mei, E. Prezioso, Z. Zhang, and N. Xu, ''A deep learning approach using graph convolutional networks for slope deformation prediction based on time-series displacement data,'' *Neural Comput. Appl.*, vol. 33, no. 21, pp. 14441–14457, May 2021, doi: 10.1007/s00521-021-06084-6.

[24] V. Maiwald, V. Vrakking, P. Zabel, D. Schubert, R. Waclavicek, M. Dorn, L. Fiore, B. Imhof, T. Rousek, V. Rossetti, and C. Zeidler, ''From ice to space: A greenhouse design for moon or Mars based on a prototype deployed in Antarctica,'' *CEAS Space J.*, vol. 13, no. 1, pp. 17–37, Jan. 2021, doi: 10.1007/s12567-020-00318-4.

[25] R. Mitra and A. L. MacLean, ''RVAgene: Generative modeling of gene expression time series data,'' *Bioinformatics*, vol. 37, no. 19, pp. 3252–3262, May 2021, doi: 10.1093/bioinformatics/btab260.

[26] A. Mahmoudi and S. A. Javed, ''Probabilistic approach to multi-stage supplier evaluation: Confidence level measurement in ordinal priority approach,'' *Group Decis. Negotiation*, vol. 31, no. 5, pp. 1051–1096, Aug. 2022, doi: 10.1007/s10726-022-09790-1.

[27] I. Karatas and T. Oktem, ''Investigation of the relationships between self-confidence levels and job finding anxiety of faculty of sports sciences students,'' *Educ. Quart. Rev.*, vol. 5, no. 1, pp. 291–302, Mar. 2022, doi: 10.31014/aior.1993.05.01.440.

[28] A. Nanda, D. B. B. Mohapatra, A. P. K. Mahapatra, A. P. K. Mahapatra, and A. P. K. Mahapatra, ''Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error,'' *Int. J. Statist. Appl. Math.*, vol. 6, no. 1, pp. 59–65, Jan. 2021, doi: 10.22271/maths.2021.v6.i1a.636.

[29] J. Mao, H. Wang, and B. F. Spencer, ''Toward data anomaly detection for automated structural health monitoring: Exploiting generative adversarial nets and autoencoders,'' *Struct. Health Monitor.*, vol. 20, no. 4, pp. 1609–1626, Jul. 2021, doi: 10.1177/1475921720924601.

[30] Y. Wang, N. Masoud, and A. Khojandi, ''Real-time sensor anomaly detection and recovery in connected automated vehicle sensors,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1411–1421, Mar. 2021, doi: 10.1109/TITS.2020.2970295.

[31] S. Ying, B. Wang, L. Wang, Q. Li, Y. Zhao, J. Shang, H. Huang, G. Cheng, Z. Yang, and J. Geng, ''An improved KNN-based efficient log anomaly detection method with automatically labeled samples,'' *ACM Trans. Knowl. Discovery Data*, vol. 15, no. 3, pp. 1–22, Apr. 2021, doi: 10.1145/3441448.

[32] S. Huang, Y. Guo, N. Yang, S. Zha, D. Liu, and W. Fang, ''A weighted fuzzy C-means clustering method with density peak for anomaly detection in IoT-enabled manufacturing process,'' *J. Intell. Manuf.*, vol. 32, no. 7, pp. 1845–1861, Oct. 2021, doi: 10.1007/s10845-020-01690-y.

[33] P. Biswas and T. Samanta, ''Anomaly detection using ensemble random forest in wireless sensor network,'' *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 2043–2052, Oct. 2021, doi: 10.1007/s41870-021-00717-8.

[34] S. T. Ikram, A. K. Cherukuri, B. Poorva, P. S. Ushasree, Y. Zhang, X. Liu, and G. Li, ''Anomaly detection using XGBoost ensemble of deep neural network models,'' *Cybern. Inf. Technol.*, vol. 21, no. 3, pp. 175–188, Sep. 2021, doi: 10.2478/cait-2021-0037.

[35] H. Lu, M. Du, K. Qian, X. He, and K. Wang, ''GAN-based data augmentation strategy for sensor anomaly detection in industrial robots,'' *IEEE Sensors J.*, vol. 22, no. 18, pp. 17464–17474, Sep. 2022, doi: 10.1109/JSEN.2021.3069452.

[36] Z. Cheng, S. Wang, P. Zhang, S. Wang, X. Liu, and E. Zhu, ''Improved autoencoder for unsupervised anomaly detection,'' *Int. J. Intell. Syst.*, vol. 36, no. 12, pp. 7103–7125, Jul. 2021, doi: 10.1002/int.22582.

[37] H. Gao, L. Zhou, J. Y. Kim, Y. Li, and W. Huang, ''Applying probabilistic model checking to the behavior guidance and abnormality detection for A-MCI patients under wireless sensor network,'' *ACM Trans. Sensor Netw.*, vol. 19, no. 3, pp. 1–24, Mar. 2023, doi: 10.1145/3499426.

[38] V. Shanmuganathan and A. Suresh, ''Markov enhanced I-LSTM approach for effective anomaly detection for time series sensor data,'' *Int. J. Intell. Netw.*, vol. 5, pp. 154–160, May 2024, doi: 10.1016/j.ijin.2024.02.007.

RUI ZHANG was born in Xianyang, Shaanxi, in October 1988. He received the bachelor's degree in electrical engineering and automation from Northeast Electric Power University, in 2014, with a focus on distribution network dispatching automation and digital power grid. Since 2018, he has been a Senior Operator with the Distribution Network Dispatch Automation Team, System Operation Department (Power Dispatch Control Center), Dongguan Power Supply Bureau of Guangdong Power Grid Corporation. He holds seven invention patents and seven articles.

LIDE ZHOU was born in Zhanjiang, Guangdong, in December 1987. He received the bachelor's degree in electrical automation technology from the Tianhe College, Guangdong Normal University, in 2010, and the master's degree in software engineering from Central South University, in 2018. From 2010 to 2013, he was a Relay Protection Team Member of the Substation Management Institute, Dongguan Power Supply Bureau. From 2013 to 2017, he was a Distribution Automation Team Leader with Dongguan Power Supply Bureau, Songshanhu Branch. Since 2017, he has been the Distribution Automation Team Leader with the Power Dispatch Control Center, Dongguan Power Supply Bureau of Guangdong Power Grid Corporation.

AOQI MEI was born in Huanggang, Hubei, in August 1995. He received the bachelor's degree in electrical engineering and automation from Southeast University, in 2018. Since 2018, he has been a member of the Main Grid Dispatch Automation Team, Dongguan Power Supply Bureau of Guangdong Power Grid Corporation. In 2022, he has published articles in EI ''Hierarchical Optimal Scheduling of Regional Integrated Energy Power System Based on Multi-Objective Particle Swarm Optimization Algorithm.'' He holds five invention patents. His research interests include power dispatch automation and digital power grids.

YIPENG HE was born in Dongguan, Guangdong, in November 1987. He received the bachelor's degree in electrical information engineering and the master's degree in electrical engineering and automation from South China University of Technology, in 2010 and 2018, respectively. Since 2020, he has been a Technician with the Distribution Network Dispatch Automation Team, System Operation Department (Power Dispatch Control Center), Dongguan Power Supply Bureau of Guangdong Power Grid Corporation. He holds two invention patents. His research interests include distribution network dispatch automation and digital power grids.