IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 August 2022, accepted 4 September 2022, date of publication 6 September 2022, date of current version 15 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3204760

RESEARCH ARTICLE

Unsupervised Outlier Detection Mechanism for Tea Traceability Data

HONGGANG YANG[1], SHAOWEN LI[2], LIJING TU[1], RONGRONG MA[1], AND YIN CHEN[1]
[1]School of Information and Computer Science, Anhui Agricultural University, Hefei, Anhui 230036, China
[2]Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment, Hefei, Anhui 230036, China

Corresponding author: Shaowen Li (shwli123@163.com)

This work was supported in part by the Anhui Province Agricultural Competitiveness Enhancement Science and Technology Action Agricultural Blockchain Pilot Project and in part by the Anhui Provincial Department of Agriculture and Rural Affairs Project under Grant 717 (Wan Cai Nong [2019]) and Grant 700 (Wan Nong Ji Cai Han [2020]).

ABSTRACT The presence of outliers in tea traceability data can mislead customers and have a significant impact on the reputation and profits of tea companies. To solve this problem, an unsupervised outlier detection mechanism for tea traceability data is proposed. Firstly, tea traceability data is uploaded to the MySQL database, and then the data is preprocessed to aggregate features based on relevance, which makes it easier to identify abnormal features. Secondly, the LOKI algorithm based on Local Outlier Factor (LOF), Isolation Forest (IForest), and K-Nearest Neighbors (KNN) algorithms is used to achieve unsupervised outlier detection of tea traceability data. In addition, a Density-Based Spatial Clustering of Applications with Noise (DBSCAN-based) tuning method for unsupervised outlier detection algorithms is also provided. Finally, the types of anomalies among the identified outliers are identified to investigate the causes of the anomalies in order to develop remedial procedures to eliminate the anomalies, and the analysis results are fed back to the tea companies. Experiments on real datasets show that the DBSCAN-based tuning method can effectively help the unsupervised outlier detection algorithm optimize the parameters, and that the LOF-KNN-IForest (LOKI) algorithm can effectively identify the outliers in tea traceability data. This proves that the unsupervised outlier detection mechanism for tea traceability data can effectively guarantee the quality of tea traceability data.

INDEX TERMS Feature combination, LOKI algorithm, machine learning, outlier detection mechanism, parameter tuning method, tea traceability.

I. INTRODUCTION

Tea originated in China and has a lengthy history. Tea drinks are one of the world's three most popular beverages. In China, there are six tea families: green tea, yellow tea, oolong tea, black tea, dark tea, and white tea. China was the world's first country to discover and use tea as well as the first to trade tea commodities. Chinese tea has also played an essential role in economic growth, enhancing China's international trade efficiency. Pesticide residues and heavy metals have harmed the quality and safety of tea in recent years and have had an influence on the tea industry's development. As a result of globalization, more regulatory authorities have focused on the traceability of tea safety and reliability, and customer expectations for tea quality are increasing. The majority of existing tea quality monitoring tools offer customers traceability information, but there are few tools that can be used by businesses to examine and manage this information. Tea traceability data analysis can assist tea businesses in identifying issues in the production management process and can be used to control tea quality at the source.

Traceability data show how things have evolved and may be used to investigate the root and source of things. The gathering of traceability data may be classified into three categories based on the input method used: manual, semi-automatic, and sensor input. With the rapid growth of the

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

internet and IoT technologies, more and more traceability data application scenarios, such as agricultural product traceability [1], [2], medication traceability [3], and food traceability [4], [5], are becoming available. Tea traceability data are information about a tea's traceability from manufacture to sale.

Tea traceability data are the tracking information for all parts of tea production and sales and may offer customers information about all aspects of tea, from planting to selling [6]. Consumers are particularly worried about the quality and safety of tea. It is difficult for businesses to acquire trustworthy tea traceability data, since the tea-producing environment is significantly influenced by uncontrollable external factors such as the soil and climate [7], and the data obtained become increasingly convoluted. As a result, anomalies in the tea traceability data gathering process are common, resulting in a low traceability data quality, customers being deceived by incorrect information, and the enterprise's credibility being harmed. High-quality tea traceability information may add value to the product and raise the selling price of tea. It is easier to ensure the quality and safety of tea that can be traced back to its source. With the growth of the economy and the rising affluence, customers are prepared to spend more money on traceable tea for the benefit of their health. High-quality traceability data may also be used by tea enterprises to enhance production and operational issues. As a result, tea traceability data outlier detection techniques for tea enterprises are required.

Outlier detection methods aim to find unusual data that differ considerably from other data and are created by various mechanisms. Depending on whether there are labels, outlier detection methods may be classified as unsupervised [8], [9], [10], [11], semi-supervised [12], [13], and supervised [14], [15]. The original data set is generally partitioned into a disjoint training set and a test set for the supervised outlier detection approach, and the training data have accurate category labels. The training set is used to improve the model's fit to the data so that the supervised algorithm can perform better on the experimental data. However, in the real-world case of outlier identification, the data are frequently unlabeled, and the disparities between outliers are considerable; thus, the supervised algorithm is ineffective. The normal data in the dataset have labels for the semi-supervised outlier detection method, but the outliers do not. The outlier detection algorithm splits normal data with labels into training and test sets, which are used for model training and performance verification, respectively, and then labels the unlabeled data using the trained model. There are no labels on the unsupervised outlier detection method's training data. The anomaly score for each data point is calculated using the general features of the data, and the anomaly scores are correlated to the data's anomaly degree. Finally, some of the data with the greatest anomaly scores are printed. Statistical-based methods, density-based methods, distance-based methods, clustering-based methods, tree-based and subspace-based methods, angle-based methods, deep-learning-based methods [16], [17], [18], and linear-model-based methods are the most common unsupervised outlier detection methods.

The credibility of tea enterprises would suffer greatly if they gathered incorrect tea traceability information throughout the manufacturing process, presented it to customers, and consumers were misled by the incorrect tea traceability information. This will then harm the profits of tea enterprises. However, enhancing the quality of the traceability data can contribute to the product's value growth. High-quality tea traceability data may also be utilized to help tea enterprises resolve production and administrative problems.

In order to solve the problems caused by the poor quality of tea traceability data and to obtain the benefits from high-quality tea traceability data. The main contributions of this paper are as follows.

(1) An unsupervised outlier detection mechanism is proposed, with the goal of identifying outliers in the data, analyzing the results, and then returning the analysis results to the tea enterprises.

(2) The LOKI algorithm is proposed with the aim of combining different types of outlier detection algorithms to improve the accuracy of outlier detection.

(3) A DBSCAN-based [19] tuning method for unsupervised anomaly detection algorithms is proposed to help the unsupervised outlier detection algorithm determine the parameters.

The remainder of this work is arranged in the following manner. The study on the use of outlier detection in many domains is reviewed in Section 2. The unsupervised outlier detection mechanism for tea traceability data is described in Section 3. The experimental data and analyses are presented in Section 4. Section 5 concludes the articles, examines the limits, and proposes future research areas.

II. RELATED WORK

The use of unsupervised outlier detection is also very popular in tea traceability data as well as in other areas. There has been a significant amount of research conducted on how to identify abnormalities in complicated systems using unlabeled data. Liu *et al.* [20]. suggested the use of an incremental unsupervised anomaly detection method to rapidly analyze large-scale, real-time data from industrial control systems. This technique generates a random binary tree set from the data stream's sampled data, combines fresh data information into the current model on a continuous basis, and provides a weighting mechanism to ensure that the set's findings are reasonably stable, even if some trees are eliminated. Mikhailova [21]. employed deep learning approaches to address civil infrastructure engineering challenges and created an unsupervised system that can automatically identify the 'train event' point. Yanjun *et al.* [22]. established an anomaly detection framework and gathered more detailed data on the time series' shape and morphological characteristics through data representation for anomaly detection in order to better detect outliers in time series data. Time series data outlier identification is also commonly employed

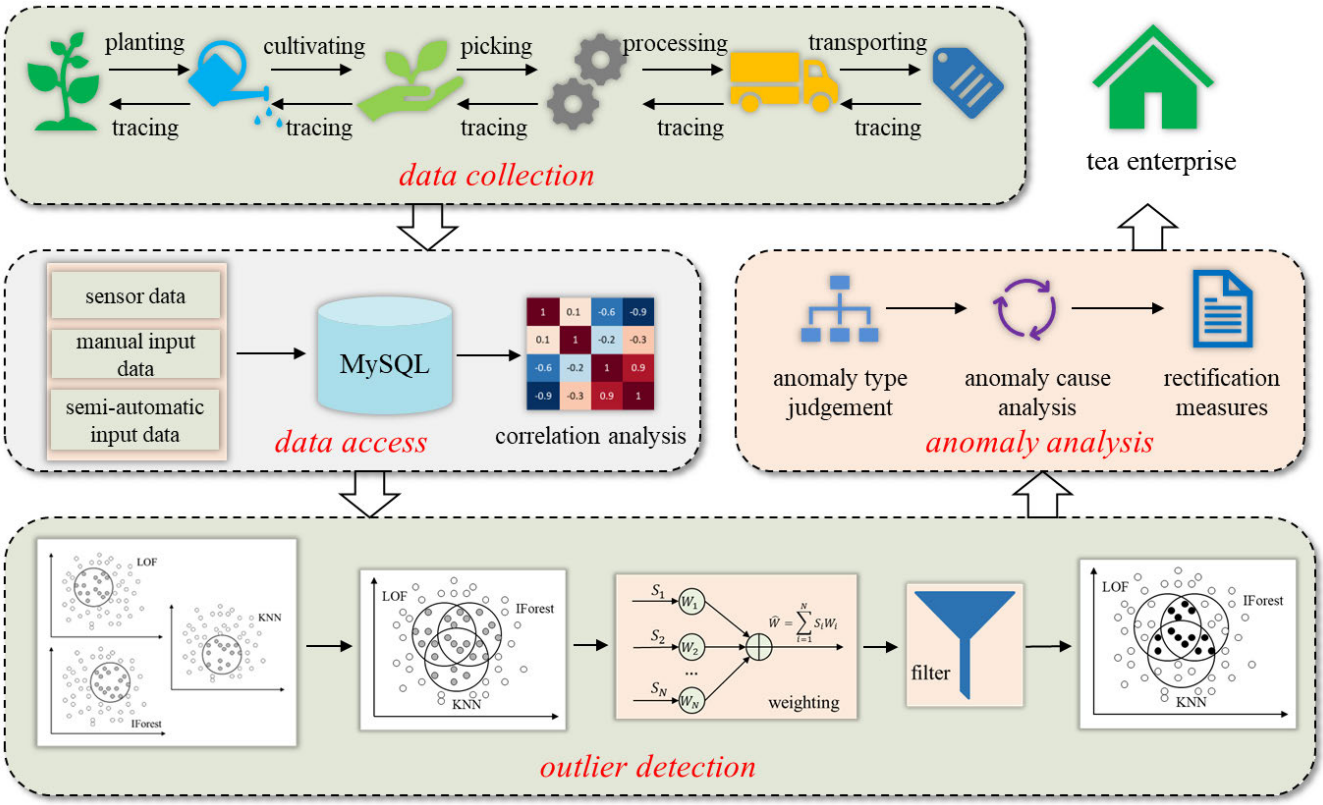

FIGURE 1. Unsupervised outlier detection mechanism used for tea traceability data.

in other domains. Álvaro *et al.* [23], for example, looked at the use of unsupervised anomaly detection technology in wood moisture content data and technology to automatically monitor abnormalities in time series data recorded from wood structures. To discover areas of vulnerability in water distribution networks and decrease false positive rates, Ane *et al.* [24]. suggested the use of a leak detection system based on self-supervised categorization of flow time series. Peng *et al.* [25]. proposed an improved Bidirectional Generative Adversarial Networks anomaly detection system to detect faults by tracking anomaly scores in order to lower the operating costs of autonomous systems operating in complex and dynamic marine environments and to achieve large-scale parallel deployment. In order to ensure successful and steady training of the generative confrontation model, the system is led by periodic supplemental prompts. Park *et al.* [26] proposed a machine anomaly detection system that combines unsupervised and non-parametric learning to detect abnormalities during machine operations using vibration data collected by the sensor.

A literature search identified very few cases of outlier detection in the world of tea traceability data. Unlike previous work, the unsupervised outlier detection mechanism proposed in this research for tea traceability data may be able to reliably discover several abnormal characteristics. To begin, the data are merged based on feature correlation to establish the types of abnormal feature combinations, and the reasons for the existence of abnormal features in each group are analyzed, followed by the implementation of appropriate improvement methods. Simultaneously, the LOKI algorithm, which combines the LOF [27], IForest [28], and KNN [29] algorithms, is proposed to increase the outlier detection accuracy by merging multiple types of outlier detection algorithms. In addition, the parameter adjustment method of an unsupervised outlier detection algorithm is suggested to aid in the optimization of parameters in an unlabeled data environment. The results of the experiments suggest that the proposed mechanism is capable of detecting outliers in tea traceability data.

III. METHOD

As illustrated in Figure 1, the tea traceability data outlier detection mechanism consists of four parts: data collection, data access, outlier detection, and anomaly analysis. Manual input, sensor input, and semi-automatic input are all examples of data collection methods. The data are uploaded to a MySQL database, which is accessible using JDBC, and the various characteristics are then integrated via correlation analysis [30]. The outlier detection part first detects outliers using the LOF, IForest, and KNN algorithms, assigns weights to the data in the detection results of the three algorithms, and finally, filters the optimal common subset of the three result sets using the weights to achieve more effective outlier detection. The anomaly analysis identifies abnormal types

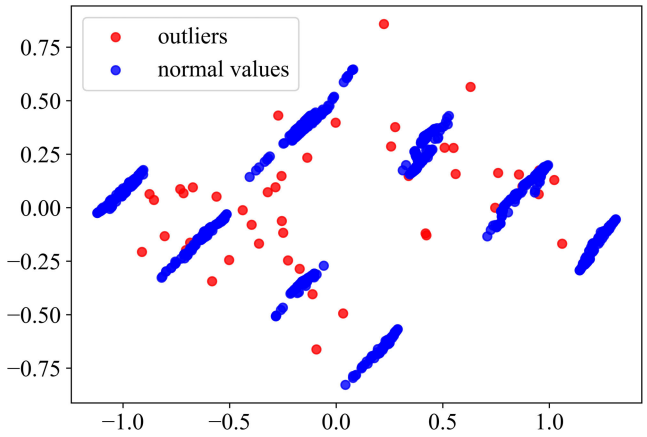

FIGURE 2. Distribution of normal values and outliers.

based on the feature combination, investigates the reasons for the occurrence of abnormal types, and lastly, provides corrective measures and feedback to the tea enterprises. A plantation information table, planting information table, inputs information table, tea information table, picking information table, processing information table, operation record table, product information table, and packaging information table are all present in MySQL. The plantation information table, for example, has fields for the plantation name, plantation number, longitude, and latitude. The longitude and latitude fields in the plantation information table are extracted to detect outliers in the traceability data.

This mechanism has three important functions: (1) It finds outliers at a finer level and identifies specific traits that appear to be outliers. In this work, the goal of the correlation analysis is to integrate characteristics with comparable causes of outlier occurrence to make the anomaly analysis easier. (2) A new algorithm combining different outlier detection algorithms is proposed to achieve more accurate outlier detection. (3) A list of the different sorts of anomalies found in the tea traceability data is compiled and the reasons for each anomaly are identified so that appropriate steps may be taken to eradicate them at their sources.

A. DATA DESCRIPTION

Tea enterprises acquire tea traceability data by sensor input, manual input, and semi-automatic input during the production process. This experiment used data (1000 records) from a tea-producing enterprise in Anhui Province. The visualization of the data [31], [32] is shown in Figure 2.

Table 1 shows the feature fields for each data set. There are 17 features and 1D labels. There are 950 normal data points and 50 outliers with outliers accounting for 5% of the total data. In the process of data collection, the longitudes and latitudes may be anomalies due to sensor failure. The tea grade, tea shape, tea color may cause anomalies in the data input due to the use of improper operation methods by employees; The weeding area, digging terraces area, planting quantity, fertilizing quantity, pruning area, picking quantity, weeding dates, digging terraces dates, planting dates, fertilizing dates, pruning dates, and picking dates may contain anomalies due to employee errors, such as repeated data entry, data omissions, and data input errors.

TABLE 1. Data feature field.

Feature Name
Longitudes
Latitudes
Weeding area
Digging terraces area
Planting quantity
Fertilizing quantity
Pruning area
Picking quantity
Tea grade
Tea shape
Tea color
Weeding dates
Digging terraces dates
Planting dates
Fertilizing dates
Pruning dates
Picking dates

B. DATA PREPROCESSING

1) NORMALIZATION

Normalization [33] involves compressing data between 0 and 1 to eliminate the order of magnitude difference between samples, ensure each data point is of the same order of magnitude, and to make the data points comparable. The normalized data follow a normal distribution, and the formula is as follows:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x_{max} represents the maximum value in the data, and x_{min} represents the minimum value in the data.

2) CORRELATION ANALYSIS

Correlation analysis is a method for analyzing the inherent links between data features. It may be used to visually illustrate the direction and degree of an intrinsic association. The linear relationship [34] between two features can be examined using the Pearson correlation coefficient. The value ranges from -1 to 1, and the closer it gets to -1, the higher the negative linear correlation between the two characteristics is. The linear correlation between two features becomes higher the closer it is to 1; the linear correlation between the two characteristics becomes smaller the closer it is to 0. The formula used to determine the Pearson correlation coefficient is

$$Cor\,(M, N) = \frac{Cov\,(M, N)}{\sqrt{Var\,(M)\,Var\,(N)}} \tag{2}$$

where $Cov(M, N)$ represents the covariance of M and N, $Var(M)$ represents the variance of M, and $Var(N)$ represents the variance of N.

C. UNSUPERVISED OUTLIER DETECTION

1) LOF

The LOF algorithm is an unsupervised outlier detection algorithm based on density, which is mainly suitable for outlier detection in low-dimensional local area space. The idea of the algorithm is to calculate the discreteness of each sample and then calculate the discreteness ratio of each sample to the sample in the field. If the obtained value is greater than a given threshold, the sample is identified as an outlier. The description of the algorithm depends on the following definitions:

Definition 1: Let $d_k(m)$ be the k distance of sample point m:

In data set D, the distance between the two sample points m, n is denoted by $d(m, n)$, if In set D, there are at least k points $n' \in S\{x \neq m\}$ that do not include m, satisfying $d(m, n') \leq d(m, n)$.

In set D, there are, at most, $k-1$ points $n' \in S\{x \neq m\}$ that do not include m, satisfying $d(m, n') < d(m, n)$.

Then, $d_k(m) = \mathrm{d}(m, n)$.

Definition 2: Let $distance_k(m, n)$ be the reachable distance from sample point n to m:

$$distance_k(m, n) = \max\{distance_k(m), d(m, n)\} \quad (3)$$

$$st.d(m, n) = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2} \quad (4)$$

The reachable distance from sample point n to m is, at least, the kth distance of sample point m. Then, the reachable distance from k nearest to sample point m is $d_k(m)$. $st.$ means subject to certain conditions.

Definition 3: Let $N_k(m)$ be the k distance neighborhood of sample point m:

$$N_k(m) = \{q \in D \setminus \{m\} | d(m, q \leq distance_k(m))\} \quad (5)$$

Definition 4. Let $lrd_k(m)$ be the local reachable density of sample point m:

$$lrd_k(m) = \frac{|N_k(m)|}{\Sigma_{n \in N_k(m)} distance_k(m, n)} \quad (6)$$

The local reachable density of sample point m represents the average reachable distance from all sample points to m in the k-neighborhood of m. If the distribution of the sample points around sample point m is relatively sparse, the k-distance neighborhood range of m is large. For sample point n of the k-distance neighborhood of sample point m, the probability of m in the k-distance field of n is small, and the probability of $distance_k(m, n) = d(m, n)$ is large, and the local reachability density of m is small. On the contrary, if the sample points around sample point m are densely distributed, the local reachability density of m is large. In short, the local reachable density explains the density of the local region of the sample points.

Definition 5: Let $LOF_k(m)$ be the local outlier factor:

$$LOF_k(m) = \frac{\Sigma_{n \in N_k(m)} \frac{lrd_k(n)}{lrd_k(m)}}{|N_k(m)|} \quad (7)$$

According to the local outlier factor algorithm, if the ratio of the local reachable density of the k nearest neighbor sample of sample point m to the local reachable density of m is close to 1, point m is more similar to its neighborhood point. If the ratio of the local reachable density of the k nearest neighbor sample of sample point m to the local reachable density of m is less than 1, the density of m is greater than that of its neighborhood point; and if the ratio of the local reachable density of the k nearest neighbor sample of sample point m to the local reachable density of sample point m is greater than 1, the density of m is less than that of its neighborhood point and it can be regarded as an isolated point, so the possibility that sample point m is an outlier is greater.

2) IFOREST

The IForest algorithm is an unsupervised fast outlier detection method based on the ensemble method, which is mainly suitable for the outlier detection of large data sets with continuous eigenvalues. The basic principle of the algorithm is to locate outliers by randomly cutting data sets. The algorithm is described as follows:

Assuming that there is a data set D, the size of the data set is n, the number of the base classifier iTrees is m, and the limit height is h.

The iTree is built and the root node of x data is randomly selected for inclusion in the iTree from the training dataset as the sample dataset for this iTree. Then, a feature p of the sample data is randomly selected to calculate the maximum and minimum values of all data in the sample data set in this feature dimension, and a data partition threshold q is randomly selected within this range. The data whose eigenvalues are less than or equal to q are put into the left subtree, and the data whose eigenvalues are greater than q are put into the right subtree. Then, the previous step is repeated in the left and right child nodes to continuously randomly divide the data until one data point in the child node reaches the limit height, so cutting is stopped and an iTree is constructed. Finally, after repeating the above method to construct m iTrees, they are merged into an IForest. Because of the big difference between normal values and outliers, outliers are more likely to be isolated faster and are more likely to appear at the root of an iTree.

When the IForest construction is completed, abnormal data points in the test data can be identified. First, the path height of the test data on each iTree is calculated as follows: The initial height of the test data is set as 0, the test data are sent to the iTree, and then look down based on the branch conditions of each node. As each node passes by, 1 path height unit is added, and the path height data are returned after finding the test data. Secondly, the average path height of the measured data in the whole IForest is calculated. Then, the anomaly score is calculated using the average path height. Finally, the

running state of the data to be measured is determined. The coefficient where the abnormal score is greater than or equal to the abnormal threshold is judged as an outlier, and the coefficient that is less than the abnormal threshold is judged as normal data.

3) KNN

The KNN algorithm is an unsupervised outlier detection algorithm based on distance, which is mainly suitable for outlier detection of low-dimensional data. The basic principle of the algorithm is as follows: for a data set, there is a new input sample, and k samples closest to the sample are found in the training data set. The class that the k samples most commonly belong to is the class of the new input sample. The algorithm first calculates and sorts the distance between the new input samples and the samples in the known category dataset. Then, k samples with the smallest distance from the new input sample are selected. Then, k samples that belong to the most categories are identified. Finally, the new input samples are determined as the k samples that belong to the most categories.

4) LOKI

In this work, the LOKI outlier detection algorithm is proposed. It was developed using the LOF, KNN, and IForest algorithms with the goal of combining multiple algorithms to increase the accuracy of outlier detection. Three high-performance algorithms were selected to complement each other by discarding the individual parts of the results of two algorithms and selecting values that are judged to be outliers by at least two algorithms. The data were first identified using the LOF, KNN, and IForest algorithms, after which the detection results from the three techniques were combined and weighted. Finally, the weights were used to determine whether the data were anomalous. The system can successfully detect outliers in tea traceability data, according to the results of the experiments.

The LOF, IForest, and KNN algorithms are the most commonly used outlier detection algorithms. The LOF performs consistently, is unaffected by the data structure, and has a good overall outlier prediction accuracy. The benefits of the IForest include its outstanding performance on low-dimensional data and its parameter insensitivity. The KNN has the benefit of having an outstanding and consistent performance with low-dimensional data. Based on the aforementioned algorithm characteristics, this work takes full advantage of the advantages of all three algorithms by merging them and then uses a screening mechanism to identify outliers. Figure 3 depicts a schematic representation of the algorithm.

The idea of the proposed LOKI algorithm is to assign weights to data in three result sets, L, I and K, from three well-performing algorithms with different types of detection results and to filter the optimal common subset P using weights to improve the detection accuracy while also improving the robustness of the algorithm. The pseudo code used in the LOKI algorithm is shown in Table 2. The algorithm inputs data set X and outputs the outliers R. First, the LOF algorithm, IForest algorithm, and KNN algorithm are used to detect the data, and the labels L_label, I_label, *and* K_label are obtained. The data points labelled 0 represent normal data, and the data points labelled 1 represent suspicious data. The suspicious data are extracted, and L, I, *and* K are obtained and merged into set M. By traversing each data point in M, the weight is calculated by the number of occurrences of each suspicious data point. The initial weight of each data point is 0, and the weights are added to 1 in set M. Finally, the suspicious data with weights greater than 1 are added to the result set R.

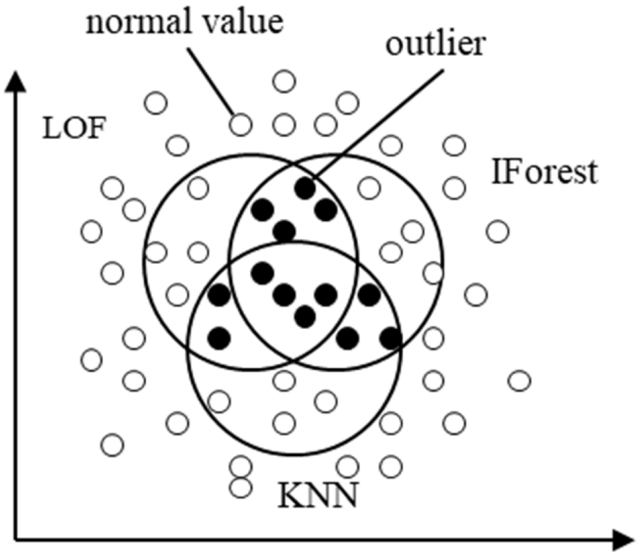

FIGURE 3. Schematic diagram of the LOKI algorithm.

TABLE 2. Algorithm pseudo code.

LOKI Algorithm
1.input: X
2.output: R
3. $L\_label \leftarrow LOF(X), I\_label \leftarrow IForest(X), K\_lable \leftarrow KNN(X)$;
4. $L \leftarrow \text{filter}(L\_label = 1), I \leftarrow \text{filter}(I\_label = 1), K \leftarrow \text{filter}(K\_label = 1)$;
5.$M \leftarrow \text{unite}(L, I, K)$;
6.for i in M
7.init$(W(i)), W(i) \leftarrow W(i) + 1$
8.if $W(i) == 2$ or $W(i) == 3$
9.$R \leftarrow \text{append}(i)$
10.endif
11.endfor
12.Return R

D. PARAMETER SELECTION

In this paper, a tuning method for an unsupervised outlier detection algorithm based on DBSCAN is proposed, unsupervised outlier detection algorithm. The algorithm's tuning

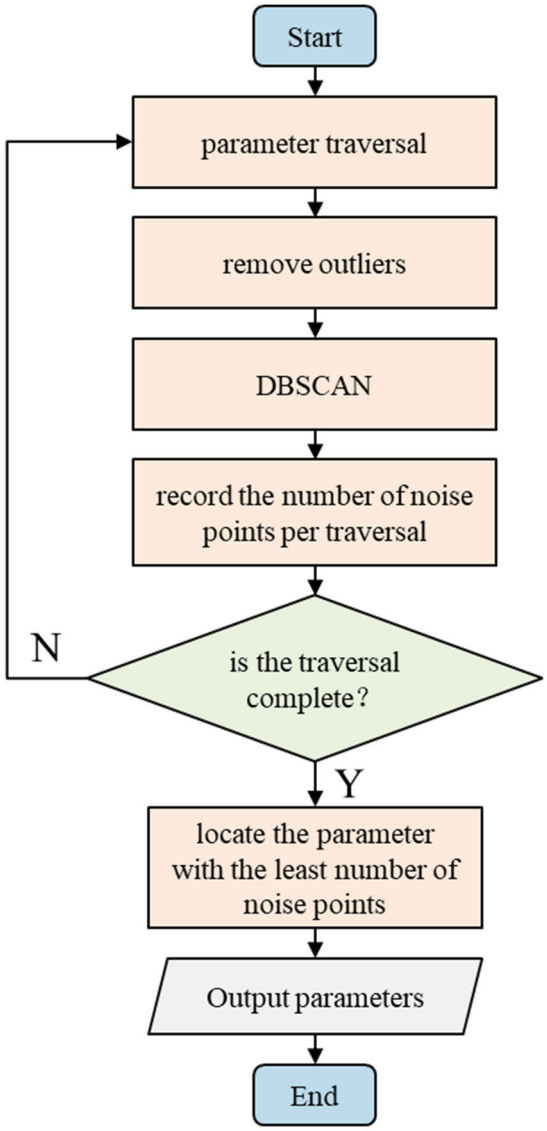

FIGURE 4. Algorithm tuning flowchart.

concept is depicted in Figure 4. First, the learning curve and grid search [35] are employed to traverse the hyperparameters using the outlier detection algorithm. Each traversal result is subtracted from the initial data. The DBSCAN technique is then used to determine the amount of noise left in the data. Finally, the quantity of noise is utilized to determine the parameters of the outlier detection algorithm, with the parameters chosen when the amount of noise is the lowest. The DBSCAN algorithm is capable of detecting noise in data sets. Because noise represents a random error or deviation in the data set that is similar to an outlier, the quantity of noise discovered may be used to assess the effectiveness of outlier detection algorithms.

The DBSCAN algorithm's key parameters during parameter adjustment are *eps* and *min_samples*. The DBSCAN's basic principle is to choose a point in a circle with a certain radius *eps* and the minimum number of nearest neighbors *min_samples*. If the point satisfies the domain circle of its radius *eps* with at least *min_samples* nearest neighbors, the center of the circle is shifted to the next sample point; if the same point does not satisfy the above conditions, the sample point is reselected and iteratively clustered according to the set radius *eps* and *min_samples*. The k-dist diagram [36] is utilized in this study to find the *eps* and *min_samples* that produce optimal clustering, providing an uniform evaluation standard for the performance of multiple outlier detection techniques with various parameters.

E. ANOMALY ANALYSIS

There are three main types of outliers in tea traceability data: outliers of sensor input data, outliers of semi-automatic input data, and outliers of manual input data.

Equipment damage and aging are the most common causes of outliers in sensor input data. To eliminate these anomalies, the following steps should be taken: (1) equipment maintenance and repair should be improved, and the equipment's key performance should be evaluated on a regular basis; and (2) Managers should be familiar with the typical state of the equipment and should debug it often in order to keep it in the best condition.

An incorrect operation method is the most common cause of outliers in semi-automatic input data. The following procedures should be taken to eliminate this type of anomaly: (1) The enterprise should develop a reasonable operating technique process based on the product's manufacturing processes; and (2) strict labor discipline should be implemented with frequent checks and supervision to ensure that staff are carrying out the manufacturing process in strict conformity with the company's operating procedures.

The major causes of outliers in manual input data include having employees who are sloppy in their production operations, do not precisely follow the enterprise's production process, and simply repeat the same activity, resulting in employee paralysis. To prevent this, (1) the staff's product quality awareness education should be strengthened and their feeling of responsibility should be increased; (2) job technical training by should strengthened by requiring each employee to learn and closely adhere to the enterprise's production workflow; (3) production and inspection employees should improve their manufacturing process control and conduct thorough process inspections; and (4) enterprises should establish an environment that allows employees to work in peace and comfort.

IV. EXPERIMENTS

A. DATASET

Before detecting outlies in tea data, the features need to be combined [37] in order to determine the type of anomaly present. The correlation heat map obtained from the correlation analysis is shown in Figure 5. The degree of linear correlation between features can be visualized. The weeding dates, digging terraces dates, planting dates, fertilizing

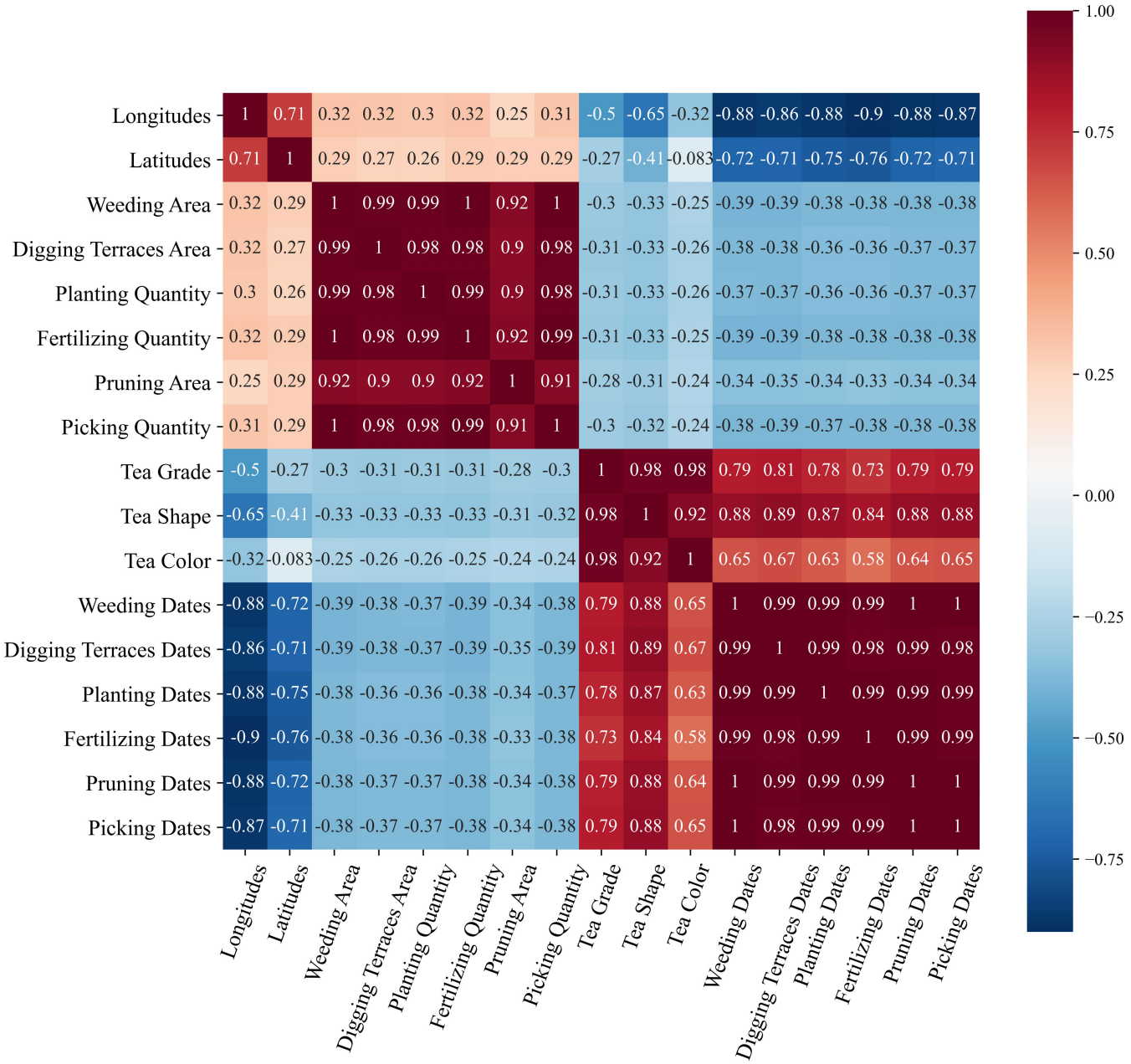

FIGURE 5. **Feature correlation heat map.**

dates, pruning dates, and picking dates all have significant connection coefficients. As a result, the aforementioned features are combined. The tea grade, tea shape, and tea color are all combined. The weeding area, digging terraces area, planting quantity, fertilizing quantity, pruning area and picking quantity are combined. The longitudes and latitudes are combined.

In this study, the outliers were oversampled using the SMOTE [38] algorithm based on the original 50 outliers and expanded it to a total data percentage of 50% with a difference of 5% to test the robustness and efficacy of the LOKI technique. Table 3 shows the proportions and volumes of data added.

B. EVALUATION INDICATORS

The Accuracy (ACC), True Negative Rate (TNR), and True Positive Rate (TPR) are used to evaluate the outlier detection performance. The specific formula is.

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

$$TNR = \frac{TN}{TN + FP} \tag{9}$$

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

where ACC stands for the accuracy rate, which is defined as the proportion of data successfully predicted by the algorithm

TABLE 3. The amounts and proportions of outliers.

Percentage of abnormal data	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Number of abnormal data	50	106	168	238	317	407	512	633	777	950
Total number of data	1000	1056	1118	1188	1267	1357	1462	1583	1727	1900

among the total data. TNR stands for the True Negative Rate, which is the ratio of properly predicted outliers to the total number of outliers predicted by the algorithm. The True Positive Rate is the proportion of normal data properly predicted by the algorithm of all normal data. The normal data points projected to be normal data points are TPs. The predicted outliers that are actually outliers are TNs; the predicted normal data points that are actually outliers are FPs; and the predicted outliers that are actually normal data points are FNs. The larger the values of the above three evaluation indicators are, the better the detection effect of the algorithm is.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) EXPERIMENT TO PROVE THE VALIDITY OF THE TUNING METHOD

In order to verify the effectiveness of the proposed tuning method, the outlier detection results of different algorithms with and without tuning are compared, namely, the density-based LOF algorithm, ensemble-based IForest algorithm, distance-based KNN algorithm, linear model-based One-Class SVM (OCSVM) algorithm [39], Cluster-Based Local Outlier Factor (CBLOF) algorithm [40], linear model-based Principal Component Analysis (PCA) algorithm [41], and Angle-Based Outlier Detector (ABOD) algorithm [42].

Each algorithm's experimental results are the average of each feature combination with the same outlier ratio. Figure 6 and Figure 8 illustrate the outcomes of the comparison. Without changing the parameters, Figure 6 shows the ACC, TNR, and TPR of seven typical methods with different outlier ratios. The ACC of each algorithm decreases as the proportion of outliers increases, as shown in Figure 6-a, and the algorithms perform erratically. The TNR of practically every method decreases as the proportion of outliers grows, as shown in Figure 6-b, and the TNR of the ABOD algorithm is always 0%, indicating that the ABOD algorithm is unable to identify outliers efficiently. Figure 6-c show the TPR trend of the algorithms with the change in the proportion of outliers, where the LOF algorithm show a decreasing trend with an increase in outliers, the OCSVM and PCA algorithms remain unchanged after increasing to a certain level, and the ABOD, KNN, IForest, and CBLOF algorithms are at a high level, indicating that they can detect normal data better but have poor detection of outliers.

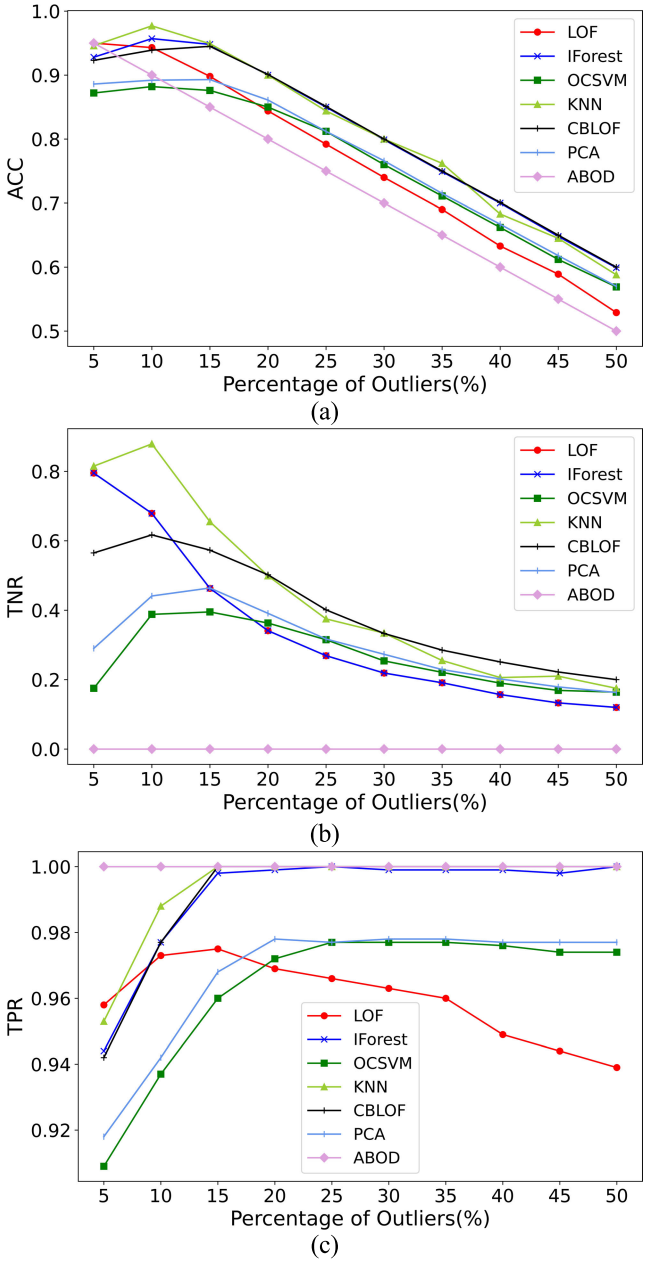

FIGURE 6. Comparison chart without tuning parameters. (a) ACC. (b) TNR. (c) TPR.

Figure 7 shows the parameter adjustment process used by the LOF algorithm for the combination of longitude and latitude features with an outlier proportion of 50%.

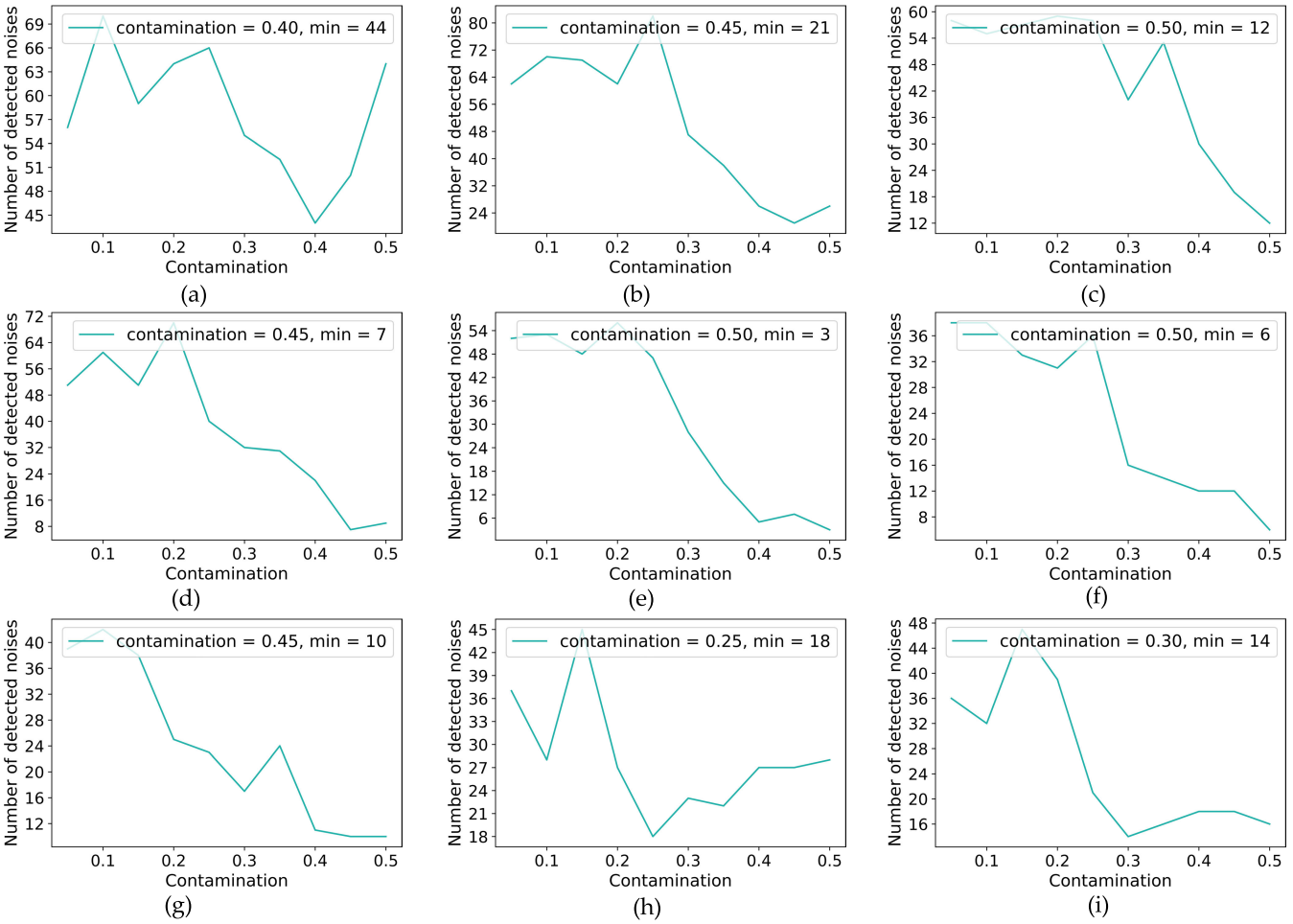

FIGURE 7. Comparison chart without tuning parameters. (a) ACC. (b)TNR. (c)TPR. (a) *n_neighbors* = 100. (b) *n_neighbors* = 200. (c) *n_neighbors* = 300. (d) *n_neighbors* = 400. (e) *n_neighbors* = 500. (f) *n_neighbors* = 600. (g) *n_neighbors* = 700. (h) *n_neighbors* = 800. (i) *n_neighbors* = 900.

The parameter adjustment process uses the grid search. The important parameters of the LOF algorithm are *n_neighbors* and *contamination*. The *n_neighbors* parameter denotes the number of neighbors used for k-nearest neighbor queries, and the *contamination* parameter denotes the proportion of outliers in the dataset. Firstly, *n_neighbors* takes 10 points 100, 200, 300, 400, 500, 600, 700, 800, 900, and then under the values of these *n_neighbors*, a learning curve of *contamination* is drawn. The value range of *contamination* is 0.05 to 0.5, and 10 points are included in the interval. Figure 7-f shows that the minimum number of noise points is detected when *n_neighbors* is taken as 500 and *contamination* is taken as 0.5, so the optimal parameters finally obtained are *n_neighbors* = 500 and *contamination* = 0.5.

Figure 8 depicts the performance of the seven most commonly used methods after using the tuning parameters. The algorithm's performance is much better after adjusting the parameters, and the algorithm's performance is more stable when compared with the situation where the settings are not tweaked. The LOF, IForest, KNN, and CBLOF algorithms perform well. Outliers can be easily spotted because the ACC, TNR, and TPR are all at high levels. As a result, this tuning method is viable.

2) COMPARISON EXPERIMENTS

The experimental results of each algorithm were averaged for each combination of characteristics under the same outlier ratio, and the LOKI algorithm was compared with the seven typical algorithms described above. The experimental results show that the LOKI algorithm is extremely reliable and better than the others in every respect.

The detection ACC of the eight techniques with varying outlier ratios is compared in Figure 9. The identification results of the PCA and OCSVM algorithms are much worse. The ACC of the IForest, ABOD, and CBLOF algorithms is slightly lower than that of the LOKI algorithm when the proportion of outliers is less, but as the proportion of outliers increases, the detection effect of the LOKI algorithm remains excellent, while the detection effects of the IForest, ABOD, and CBLOF algorithms deteriorate. The ACC of the LOKI algorithm is higher than that of the LOF and KNN algorithms, which has a clear relative advantage. The KNN algorithm has

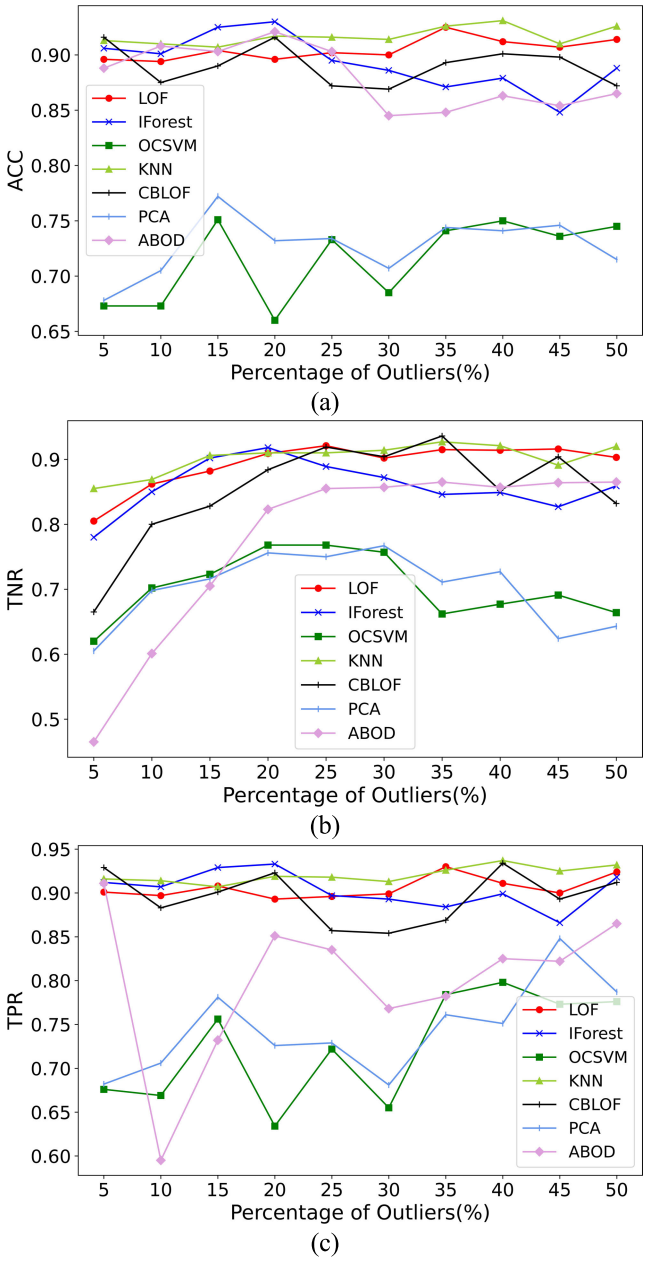

FIGURE 8. Comparison chart of the results after using the tuning parameters. (a) ACC. (b) TNR. (c) TPR.

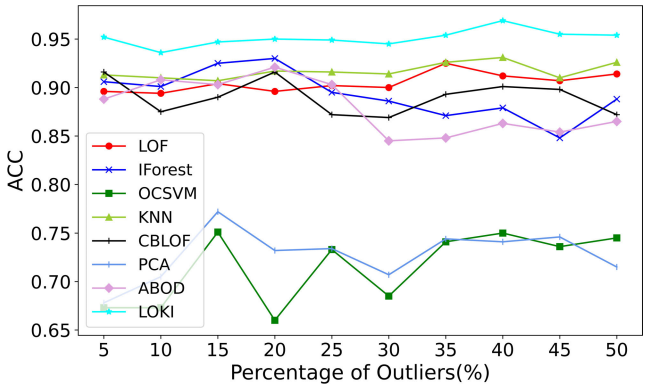

FIGURE 9. Comparison diagram of algorithm ACC with different outlier ratios.

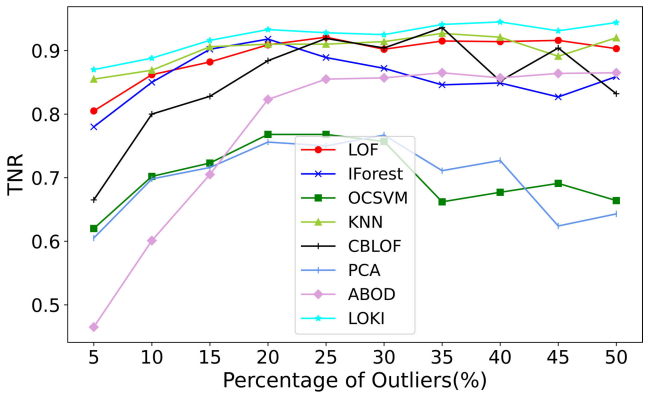

FIGURE 10. Comparison diagram showing the algorithm TNR under different outlier ratios.

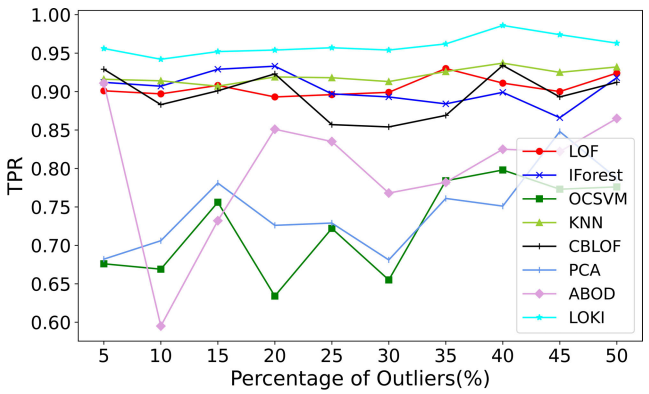

FIGURE 11. Comparison diagram of the algorithm TPR under different outlier ratios.

a high ACC, with the biggest difference being 4.5% between the KNN method and the LOKI algorithm with an outlier ratio of 45%. In general, the accuracy of the KNN algorithm is 3.4% lower than that of the LOKI algorithm.

Figure 10 compares the TNR of the eight algorithms with various outlier percentages. The TNR of the LOKI algorithm is greater than that of the CBLOF algorithm, 6.9% higher on average, which is a clear advantage, as shown in the comparison diagram. The detection rates of the LOF, IForest, and KNN algorithms are consistently lower than those of the LOKI algorithm, with the value of the LOF algorithm being 2.9% lower on average, that of the IForest algorithm being 6.3% lower on average, and that of the KNN algorithm being 2% lower on average. The LOKI algorithm remains stable when the fraction of outliers changes, but the OCSVM, PCA, and ABOD algorithms vary more. The TNR is the most crucial evaluation indication for businesses, since they do not want to pass on any outliers to their customers.

The TPR of the eight algorithms is compared in Figure 11 for different outlier proportions. With a percentage of outlier points of 5% to 10%, the ABOD algorithm has the largest difference in TPR with a 31.6% decrease. The KNN algorithm is closest to the LOKI algorithm and is relatively 3.9% lower.

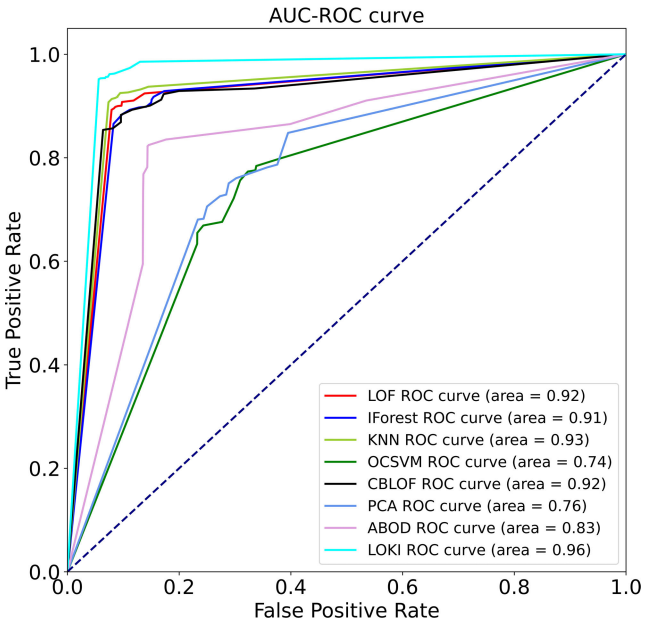

FIGURE 12. Comparison diagram of AUC-ROC curves for all outlier ratios.

Figure 12 depicts the AUC-ROC curves of 8 algorithms under all outlier ratios. Each AUC-ROC curve is plotted using 10 points, and the horizontal and vertical coordinates of each point are the FPR and TPR for each algorithm at each outlier ratio. The AUC values of the LOKI, LOF, IForest, KNN, and CBLOF algorithms are 0.96, 0.92, 0.91, 0.93, and 0.92, respectively, according to Figure 12. These values are superior to those of the OCSVM, PCA, and ABOD algorithms. The KNN algorithm's AUC value is the closest to that of the LOKI algorithm, but it is still 0.03 lower, indicating the LOKI algorithm's superior detection ability.

The LOKI algorithm has strong outlier detection and stability and can perform well under diverse outlier ratios, according to the four assessment indices listed above.

3) FRAMEWORK FUNCTIONAL COMPARISON

Hendrickx *et al.* [43]. proposed an anomaly detection framework for fleet-based condition monitoring, which is divided into four parts, namely, machine comparison, fleet clustering, anomaly detection, and visualization. The first part compares the similarities between the behaviors of two machines. The second part groups machines with similar behaviors using the clustering algorithm and the chosen measure. The third part uses the discovered clusters to assign an anomaly score to each machine. The fourth part helps to guide domain experts in analyzing specific deviating machines by visualizing the results of the other parts. However, the framework does not enable more granularity in locating exceptions, high-performance outlier detection, and feedback anomalies. Lee *et al.* [44]. developed a real-time health monitoring framework for predicting possible flight performance anomalies. The framework includes a training phase and a monitoring phase. The initial extraction preprocessing and Savitzky–Golay filtering of the flight data recorder are performed in the training phase to synchronize the sampling frequency and reduce the random noise in the sensor signal. The preprocessed flight features are then reduced by feature subset selection to select features that are highly correlated with the dynamic flight characteristics. The selected features are then used to train model classes to predict common patterns in flight performance during the takeoff and ascent phases. The monitoring phase simulates the flight data recorder dataset and introduces its real time data into the trained model to validate the detection capability of the proposed framework in real-time situations. Anomalous flight performances are detected when the predicted feature values violate the safety boundaries. However, the framework is incapable of achieving high-performance anomaly detection and feedback. Enrico *et al.* [45]. proposed an online remote fault detection system for underwater gliders to identify undesirable behaviors on the horizon. The system is tested using a deployment dataset of undesirable vehicle behaviors. Once the effectiveness of the system is determined, a trained anomaly detection scheme can be used online from a remote-controlled center to notify the pilot of a possible failure of the underwater glider after each surfacing and maintenance connection. The system does not allow for more granular detection of anomalies and does not provide an analysis on anomalies. Wada *et al.* [46]. proposed an adaptive-model-based anomaly detection system for daily life activities that adapts to new data corresponding to changes in human behavioral habits over time. A forgetting factor data-driven filtering approach was proposed to help the system adapt to the current behavioral habits of individuals while discarding features that are not relevant to old habits. The forgetting factor allows the system to identify outdated activity data that should be discarded while incorporating data representing changes in human behavior routines for adaptation. A total of two forgetting factor approaches are proposed in the paper: the data aging-based forgetting factor and the data difference-based forgetting factor. A set of anomaly detection models is then used for behavior modeling. The system cannot locate anomalous data at a fine-grained level and also does not provide an analysis or feedback on the anomalies. A comparison of the functions of each framework is shown in Table 4.

The above analysis compares the functionality of existing anomaly detection frameworks, each of which is lacking in terms of completeness. The mechanism proposed in this paper is functionally complete and is capable of locating outliers with fine granularity, achieving high performance outlier detection, analyzing the anomalies, and providing feedback on the detection and analysis results.

V. CONCLUSION

This work provides an unsupervised outlier detection mechanism for tea traceability to improve the quality of tea traceability data in order to address the challenges caused by poor data quality. The LOKI algorithm is proposed to improve the accuracy of outlier detection. It is suggested that the

TABLE 4. Framework function comparison.

	fine-grained localization of abnormal data	outlier detection with high performance	anomaly analysis	anomaly feedback
Kilian et al.	×	×	√	×
Hyunseong et al.	×	√	√	×
Enrico et al.	×	√	×	√
Salisu et al.	×	√	×	×
This paper	√	√	√	√

features of tea traceability data can be combined according to their correlations in order to determine the reasons for the occurrence of outliers with distinct characteristics so that targeted improvement actions can be implemented. An unsupervised anomaly detection algorithm based on DBSCAN was proposed with a parameter modification mechanism to optimize the algorithm parameters. The experimental results reveal that the proposed outlier detection mechanism for tea traceability data is well-functioning and can locate outliers at a finer granularity. The LOKI algorithm's is excellent and reliable in regard to outlier detection. When the quantity of outliers in the dataset is unknown, the suggested parameter adjustment approach can assist the outlier detection algorithm in selecting the best parameters.

The results of this study have the potential to encourage knowledge sharing in the tea supply chain. The described technology can assure the accuracy of tea traceability data and allow tea enterprises to fully comprehend production and operation issues and make timely, targeted adjustments. The following are some of the future research goals: (1) The proposed unsupervised outlier detection mechanism for tea traceability data needs to be applied to specific tea production in subsequent studies, and the mechanism needs to be further improved; (2) because timely monitoring can aid in the discovery of outliers, it is vital to investigate online outlier detection mechanisms; (3) because the LOKI algorithm has a high time cost, more research into how to increase the method's operating efficiency is required; and (4) more research into how to increase the effectiveness of the parameter modification approach and broaden its use in the field of unsupervised outlier detection is required.

REFERENCES

[1] Y. Yang, T. Wei, and M. Li, "Research and practice of agricultural product quality traceability social intercourse application," in *Proc. E3S Web Conf.*, vol. 235, 2021, pp. 1–5.

[2] L. Li, K. P. Paudel, and J. Guo, "Understanding Chinese farmers' participation behavior regarding vegetable traceability systems," *Food Control*, vol. 130, Dec. 2021, Art. no. 108325.

[3] Z. Wang, L. Wang, F. Xiao, Q. Chen, L. Lu, and J. Hong, "A traditional Chinese medicine traceability system based on lightweight blockchain," *J. Med. Internet Res.*, vol. 23, no. 6, Jun. 2021, Art. no. e25946.

[4] S. Islam, J. M. Cullen, and L. Manning, "Visualising food traceability systems: A novel system architecture for mapping material and information flow," *Trends Food Sci. Technol.*, vol. 112, pp. 708–719, Jun. 2021.

[5] A. Iftekhar and X. Cui, "Blockchain-based traceability system that ensures food safety measures to protect consumer safety and COVID-19 free supply chains," *Foods*, vol. 10, no. 6, p. 1289, Jun. 2021.

[6] J. Wang, J.-M. Wang, and Y.-J. Zhang, "Agricultural product quality traceability system based on the hybrid mode," in *Proc. 4th Annu. Int. Conf. Netw. Inf. Syst. Comput. (ICNISC)*, Apr. 2018, pp. 392–395.

[7] V. Bellon-Maurel, M. D. Short, P. Roux, M. Schulz, and G. M. Peters, "Streamlining life cycle inventory data generation in agriculture using traceability data and information and communication technologies—Part I: Concepts and technical basis," *J. Cleaner Prod.*, vol. 69, pp. 60–66, Apr. 2014.

[8] S. Maleki, S. Maleki, and N. R. Jennings, "Unsupervised anomaly detection with LSTM autoencoders using statistical data-filtering," *Appl. Soft Comput.*, vol. 108, Sep. 2021, Art. no. 107443.

[9] S. Fu, S. Zhong, L. Lin, and M. Zhao, "A re-optimized deep auto-encoder for gas turbine unsupervised anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 101, May 2021, Art. no. 104199.

[10] X. Du, J. Yu, Z. Chu, L. Jin, and J. Chen, "Graph autoencoder-based unsupervised outlier detection," *Inf. Sci.*, vol. 608, pp. 532–550, Aug. 2022.

[11] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. Chen, "ECOD: Unsupervised outlier detection using empirical cumulative distribution functions," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 16, 2022, doi: 10.1109/TKDE.2022.3159580.

[12] X. Meng, S. Wang, Z. Liang, D. Yao, J. Zhou, and Y. Zhang, "Semi-supervised anomaly detection in dynamic communication networks," *Inf. Sci.*, vol. 571, pp. 527–542, Sep. 2021.

[13] M. Shao and N. Gu, "Anomaly detection algorithm based on semi-supervised collaborative strategy," *J. Phys., Conf. Ser.*, vol. 1944, no. 1, Jun. 2021, Art. no. 012017.

[14] C. Piciarelli, P. Mishra, and G. L. Foresti, "Supervised anomaly detection with highly imbalanced datasets using capsule networks," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 35, no. 8, Jun. 2021, Art. no. 2152010.

[15] S. Vargaftik, I. Keslassy, A. Orda, and Y. Ben-Itzhak, "RADE: Resource-efficient supervised anomaly detection using decision tree-based ensemble methods," *Mach. Learn.*, vol. 110, no. 10, pp. 2835–2866, Oct. 2021.

[16] L. Xiang, W. Penghe, and L. Jingxu, "Abnormal state detection of wind turbines based on CNN-LSTM," *J. Vib. Shock*, vol. 40, no. 22, pp. 11–17, 2021.

[17] L. Chen, W. Buhong, T. Jiwei, and G. Rongxiao, "Anomaly detection method for UAV sensor data based on LSTM-OCSVM," *J. Chin. Comput. Syst.*, vol. 42, no. 4, pp. 700–705, 2021.

[18] N. A. Andriyanov, V. E. Dementiev, and A. G. Tashlinskiy, "Detection of objects in the images: From likelihood relationships towards scalable and efficient neural networks," *Comput. Opt.*, vol. 46, no. 1, pp. 139–159, Feb. 2022.

[19] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, no. 10, p. 2809, May 2020.

[20] L. Liu, M. Hu, C. Kang, and X. Li, "Unsupervised anomaly detection for network data streams in industrial control systems," *Information*, vol. 11, no. 2, p. 105, Feb. 2020.

[21] A. Mikhailova, N. M. Adams, C. A. Hallsworth, F. D.-H. Lau, and D. N. Jones, "Unsupervised deep-learning-powered anomaly detection for instrumented infrastructure," *Proc. Inst. Civil Eng.-Smart Infrastruct. Construct.*, vol. 172, no. 4, pp. 135–147, Dec. 2019.

[22] Y. Zhou, H. Ren, Z. Li, and W. Pedrycz, "An anomaly detection framework for time series data: An interval-based approach," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107153.

[23] Á. G. Faura, D. Štepec, M. Cankar, and M. Humar, "Application of unsupervised anomaly detection techniques to moisture content data from wood constructions," *Forests*, vol. 12, no. 2, p. 194, Feb. 2021.

[24] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "Water leak detection using self-supervised time series classification," *Inf. Sci.*, vol. 574, pp. 528–541, Oct. 2021.

[25] P. Wu, C. A. Harris, G. Salavasidis, A. Lorenzo-Lopez, I. Kamarudzaman, A. B. Phillips, G. Thomas, and E. Anderlini, "Unsupervised anomaly detection for underwater gliders using generative adversarial networks," *Eng. Appl. Artif. Intell.*, vol. 104, Sep. 2021, Art. no. 104379.

[26] S. Park, J. Kang, J. Kim, S. Lee, and M. Sohn, "Unsupervised and non-parametric learning-based anomaly detection system using vibration sensor data," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 4417–4435, Feb. 2019.

[27] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.

[28] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012.

[29] T. T. Dang, H. Y. T. Ngan, and W. Liu, "Distance-based k-nearest neighbors outlier detection method in large-scale traffic data," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 507–510.

[30] W. Moxian, D. Xiaoou, W. Hongzhi, and L. Jianzhong, "Correlation-based method for tracing multi-dimensional time series data anomalies," *J. Frontiers Comput. Sci. Technol.*, vol. 15, no. 11, pp. 1–10, 2020, doi: 10.3778/J.ISSN.1673-9418.2008100.

[31] H. Burgan, B. Vaheddoost, and H. Aksoy, "Frequency analysis of monthly runoff in intermittent rivers," in *Proc. World Environ. Water Resour. Congr.*, 2017, pp. 327–334.

[32] F. Chebana and T. B. M. J. Ouarda, "Multivariate non-stationary hydrological frequency analysis," *J. Hydrol.*, vol. 593, Feb. 2021, Art. no. 125907.

[33] X. Zhang, B. Han, J. Wang, Z. Zhang, and Z. Yan, "A novel transfer-learning method based on selective normalization for fault diagnosis with limited labeled data," *Meas. Sci. Technol.*, vol. 32, no. 10, Oct. 2021, Art. no. 105116.

[34] Z. Qi, H. Yupeng, J. Cun, Z. Peng, and L. Xueqing, "Edge computing application: Real-time anomaly detection algorithm for sensing data," *J. Comput. Res. Develop.*, vol. 55, no. 3, pp. 524–536, 2018.

[35] C. Liu, J. Yang, and J. Wu, "Web intrusion detection system combined with feature analysis and SVM optimization," *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, pp. 1–9, Dec. 2020.

[36] Z. Dong and L. Peng, "VDBSCAN: Varied density based clustering algorithm," *Comput. Eng. Appl.*, vol. 45, no. 11, pp. 137–141, 2009.

[37] W. Ma, D. Tran, and D. Sharma, "A study on the feature selection of network traffic for intrusion detection purpose," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, Jun. 2008, pp. 245–247.

[38] T. T. Khuat and M. H. Le, "Evaluation of sampling-based ensembles of classifiers on imbalanced data for software defect prediction problems," *Social Netw. Comput. Sci.*, vol. 1, no. 2, pp. 1–16, Mar. 2020.

[39] T. Morris, A. Srivastava, B. Reaves, W. Gao, K. Pavurapu, and R. Reddi, "A control system testbed to validate critical infrastructure protection concepts," *Int. J. Crit. Infrastruct. Protection*, vol. 4, no. 2, pp. 88–103, 2011.

[40] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognit. Lett.*, vol. 24, nos. 9–10, pp. 1641–1650, Jun. 2003.

[41] K. Pearson, "On lines and planes of closest fit to points in space," Tech. Rep., 1900.

[42] H.-P. Kriegel, M. S. Hubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Las Vegas, NV, USA, Aug. 2008, pp. 444–452.

[43] K. Hendrickx, W. Meert, Y. Mollet, J. Gyselinck, B. Cornelis, K. Gryllias, and J. Davis, "A general anomaly detection framework for fleet-based condition monitoring of machines," *Mech. Syst. Signal Process.*, vol. 139, May 2020, Art. no. 106585.

[44] H. Lee, G. Li, A. Rai, and A. Chattopadhyay, "Real-time anomaly detection framework using a support vector regression for the safety monitoring of commercial aircraft," *Adv. Eng. Informat.*, vol. 44, Apr. 2020, Art. no. 101071.

[45] E. Anderlini, G. Salavasidis, C. A. Harris, P. Wu, A. Lorenzo, A. B. Phillips, and G. Thomas, "A remote anomaly detection system for Slocum underwater gliders," *Ocean Eng.*, vol. 236, Sep. 2021, Art. no. 109531.

[46] S. W. Yahaya, A. Lotfi, and M. Mahmud, "Towards a data-driven adaptive anomaly detection system for human activity," *Pattern Recognit. Lett.*, vol. 145, pp. 200–207, May 2021.

HONGGANG YANG is currently pursuing the M.S. degree with the School of Information and Computer Science, Anhui Agricultural University. His research interests include blockchain technology and machine learning.

SHAOWEN LI was born in Hefei, China, in 1962. He is currently a Professor with the School of Information and Computer Science, Anhui Agricultural University. Since the mid-1990s, he has been engaged in research on intelligent agricultural information technology, and has been invited to visit universities and research institutions in the USA, Germany, Japan, and Israel, for many times. His research interests include artificial intelligence and agricultural expert systems.

LIJING TU is currently pursuing the Ph.D. degree with the School of Information and Computer Science, Anhui Agricultural University. Her research interest includes personalized computing.

RONGRONG MA received the M.S. degree in agricultural engineering and information technology from the School of Information and Computer Science, Anhui Agricultural University. Her research interest includes research on blockchain's privacy information protection method.

YIN CHEN received the M.S. degree in agricultural engineering and information technology from the School of Information and Computer Science, Anhui Agricultural University. Her research interest includes blockchain identity authentication.