IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 December 2023, accepted 23 January 2024, date of publication 8 February 2024, date of current version 14 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359302

RESEARCH ARTICLE

Unsupervised Real Time and Early Anomalies Detection Method for Sewer Networks Systems

CHUNMING QIU[1,2], GUOXIANG SHAO[1,2], ZHENYU ZHANG[1,2], CHICHUN ZHOU[1,2], YUEJIE HOU[1,2], ENMING ZHAO[1], XIAO GUO[3], AND XIAOLIN GUAN[1,2]

[1]School of Engineering, Dali University, Dali, Yunnan 671003, China
[2]Air-Space-Ground Integrated Intelligence and Big Data Application Engineering Research Center, Yunnan Provincial Department of Education, Dali, Yunnan 671003, China
[3]Institute of Unmanned System, Beihang University, Beijing 100191, China

Corresponding authors: Zhenyu Zhang (zhangzhenyu@dali.edu.cn) and Chichun Zhou (zhouchichun@dali.edu.cn)

This work was supported in part by the Intelligent Monitoring Project of the Pollution Control of Erhai Lake; in part by the National Natural Science Funds of China under Grant 62106033, Grant 62065001, and Grant 42367066; in part by the Yunnan Fundamental Research Projects under Grant 202001AU070020 and Grant 202301BA070001-037; in part by the Doctoral Programs of Dali University under Grant KYBS201910; and in part by the Scientific Research Fund Project of Yunnan Provincial Department of Education under Grant 2022Y868 and Grant 2023Y1042.

ABSTRACT Sewer networks (SNs) are susceptible to various factors that can lead to failures, resulting in economic losses and environmental pollution. Data-driven approaches based on sewage flow monitoring enhance the awareness and maintenance capabilities of SNs. However, the current research lacks early warning systems for flow anomalies. This presents a challenge for the application of supervised methods, primarily due to the scarcity of anomalous flow datasets. Even with the availability of such datasets, the effectiveness of these methods may vary due to environmental differences, since SNs are situated in diverse environments. Therefore, effectively achieving early warnings for anomalies in unlabeled flow data is a challenge that must be addressed in the field of flow monitoring. To address this challenge, we propose a detection method for effectively warning of anomalies in flow data. Since anomalies typically result in significant deviations from normal data, early warnings can be achieved by comparing the differences between current and historical data. The key to this early warning lies in establishing an adaptive threshold for detecting abnormal data changes. Our detection method employs an unsupervised bagging-based multi-anomaly detection algorithm to detect such abnormal data changes. Experiments conducted on Erhai Lake SNs flow data demonstrate that our method can predict anomalies 5-15 minutes in advance with a precision of 80.00%, a recall of 66.67%, and an F1 score of 0.73. Our approach not only achieves cost-effective and timely anomalies detection but also overcomes the challenges associated with limited dataset availability, making it applicable to various other industries.

INDEX TERMS Unsupervised, anomaly detection, bagging, sewer networks, data-driven, early warning.

I. INTRODUCTION

Sewer networks (SNs) are an integral part of the urban wastewater system (UWS), and their failure can severely affect the operation of the entire system [1], [2]. Maintenance and repair costs associated with SNs worldwide amount to trillions of dollars annually, and these costs are expected to increase as the frequency of failures rises [3], [4], [5], [6]. In order to effectively assist people in managing their assets and reducing the costs and consequences of SNs failures, various management approaches have been proposed. These approaches can be broadly categorized into three main groups: 1) fault detection, 2) deterioration modeling, and 3) sensor-based data-driven methods.

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

Fault detection methods mainly detect physical faults in the SNs, such as deformations and breaks [7], [8], [9]. Typical methods include closed circuit television (CCTV) [10], [11], [12], sonar [13], [14], [15] and radar [16], [17], [18]. These methods can effectively detect fault anomalies in the SNs, and in recent years have achieved high detection accuracy when combined with artificial intelligence (AI) technology [19], [20], but they are usually used to detect faults that have already occurred and cannot effectively compensate for the losses caused by faults. Therefore, considering that corrosion is a major cause of physical failures in SNs, deterioration models have been developed to predict corrosion in SNs and make timely decisions to reduce the impact of corrosion on SNs [21], [22]. Currently, deterioration models are available as deterministic [23], [24], statistical [25], [26] and probabilistic models [27], [28], etc. These models can effectively predict the corrosion of SNs, but they are primarily used to predict the deterioration of SNs under natural conditions. When affected by sudden external factors, such as the discharge of untreated chemical waste liquids, contamination of the soil around the SNs. These factors can easily cause the method to fail. Therefore, in order to effectively achieve early detection of abnormal in the SNs and to effectively overcome the interference of various external mutating factors, sensor-based data-driven methods have been proposed. People place various types of sensors in the SNs to obtain monitoring data, and the monitoring data is combined with AI technology to effectively achieve real-time monitoring and intelligent fault diagnosis of the SNs [2], [29], [30], [31]. Compared with fault detection and deterioration modelling approaches, the sensor-based data-driven methods offers better monitoring efficiency, does not need to rely too much on the expertise of system experts, is easy to implement, and is now widely used to monitor anomalies in SNs [2], [32].

Currently, among sensor-based data-driven methods, sewage flow monitoring has proven to be an effective means of managing SNs. It accomplishes this by reflecting the condition of SNs through flow data, thereby enhancing awareness and facilitating maintenance. Sewage flow detection significantly contributes to urban safety and livability [33], [34], as it allows for the timely identification of SNs leaks and aids in the maintenance of drainage systems [35], [36], [37]. Furthermore, compared to other data-driven methods such as water quality monitoring [38], sewage flow monitoring uses cost-effective flow meters for long-term SNs monitoring [39]. Sewage flow monitoring plays a vital role in SNs management. However, it is worth noting that limited systematic attention has been devoted to flow feedback anomalies in sewage flow monitoring studies, resulting in a lack of early warning systems for these anomalies [34]. Additionally, the absence of anomalous flow datasets has posed challenges in applying existing monitoring methods. Even when anomalous flow datasets are available, SNs are located in various environments, which can lead to biases in flow distribution [40]. These environmental factors include differences in precipitation between the rainy and dry seasons, different types of industrial and residential areas in various regions with different water demands and water use habits. As shown in Figure 1, it can be observed that the difference in precipitation between the dry and rainy seasons results in varying water levels in the two different seasons. The anomalies that exceed the thresholds in the rainy season (b) are at a similar distance from the normal data points in the dry season (a) in the t-Distributed Stochastic Neighbor Embedding (t-SNE) example plot (c). This suggests that there are differences in the definition of anomalies in different environmental flow data, and that SNs data from different environmental conditions need to be segmented to avoid environmental biases that can lead to method failure. In this case, using supervised methods would necessitate constructing datasets for SNs under various environmental conditions. However, labeling flow data from SNs in different periods and environments poses challenges such as time-consuming manual labeling of data, high costs, and low feasibility. Therefore, the effective early warning of unlabeled SNs flow data under different environmental conditions is a challenging problem that needs to be addressed in current flow monitoring.

In this study, we propose a detection method that effectively provides early warnings for anomalies in flow data. Since anomalies often result in significant deviations from normal data, early warning can be achieved by comparing the differences between current and historical data. The key to achieving early warning lies in applying an adaptive threshold to detect abnormal data changes. Our detection method employs an unsupervised bagging-based multi-anomaly detection algorithm to identify abnormal data changes. This algorithm is created by combining three anomaly detection algorithms: one class SVM, isolation forest, and local outlier factor, using a bagging technique inspired by previous work [41], [42]. Previous results have shown that bagging multiple algorithms enhances precision due to the comprehensive field of view and multiple perspectives it provides. Our detection method, based on the comparison of current and historical data, combined with the unsupervised bagging-based multi-anomaly detection algorithm, effectively mitigates the flow data discrepancies caused by environmental variations in SNs and addresses the issue of unlabeled data, allowing for cost-effective early warning of flow anomalies and providing an appropriate response time for SNs monitoring and maintenance. Importantly, our detection method, free from domain bias, shows promise for application in other industrial anomaly detection scenarios.

This study is organized as follows. In Section II, we present the study data, and the data preprocessing steps. The research task and the main method are introduced in Section III. In Section IV, the experiments are conducted and the results are shown and analyzed. Conclusions and discussions are given in Section V.

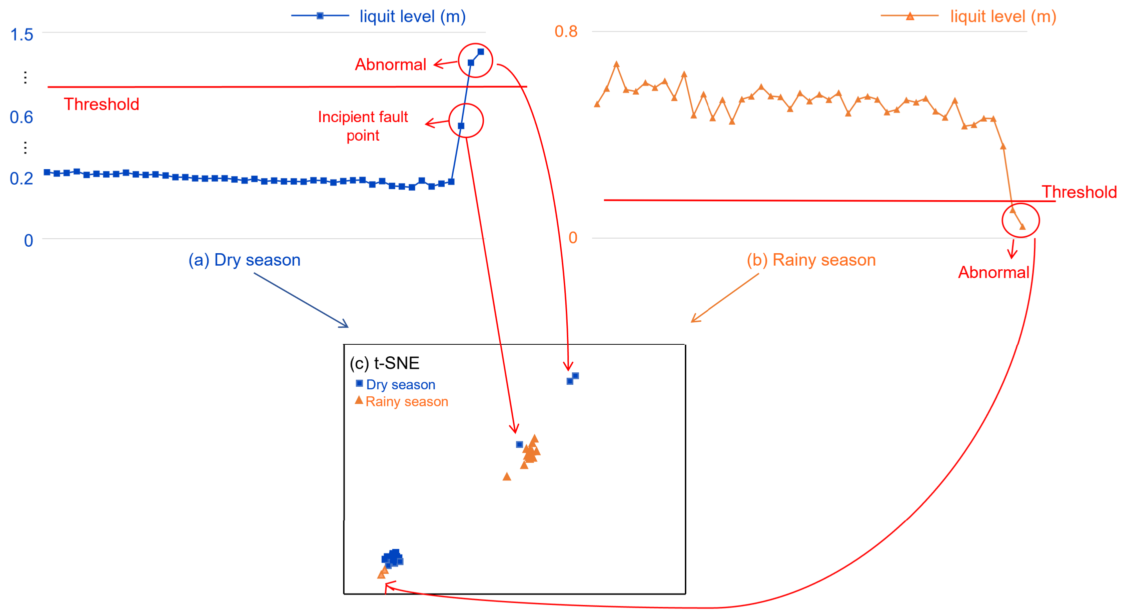


FIGURE 1. The figure displays time series data for the liquid level in the same SNs during different seasons. In (a), it presents the time series data for the liquid level during the dry season, while (b) illustrates the time series data for the liquid level during the rainy season. By combining these two sequences from distinct periods, an example t-SNE plot (c) is generated, representing the flow data for both seasons. In (a), points that deviate from the normal data but do not exceed the threshold are referred to as 'initial fault points' by us, while the remaining points that exceed the threshold are classified as anomalies. The thresholds used for determining anomalies in the figure are established based on the characteristics of SNs' diameter and operation.

II. TYPES OF ANOMALIES AND RESEARCH DATA

In this section, we introduced the data used in this research and the types of anomalies.

A. TYPES OF ANOMALIES AND DATA INTRODUCTION

The government department of Dali, China, initiated the Intelligent Monitoring Project for the Pollution Control of Erhai Lake, a renowned lake in Yunnan, China. The data used in this study was provided by this project. The project's objective is to achieve real-time monitoring of the SNs surrounding Erhai Lake by installing flow meters. These flow meters are employed to identify anomalies, including damages, blockages, and overflows. Among the flow meters used, we selected the ultrasonic Doppler flowmeter as the primary measurement equipment. This choice was suitable for the large-diameter SNs around Erhai Lake and could be easily installed without any modifications to the SNs. This ease of installation supports long-term SNs monitoring [34]. The flow data collected by the ultrasonic Doppler flowmeter includes sampling time (day/hour), instantaneous flow rate (m^3/h), liquid level (m), and flow velocity (m/s). The data is sampled at 5-minute intervals.

After analyzing the data, we classified the anomalies into three types: (1) The instantaneous flow and liquid level abruptly drop to zero, primarily due to damaged flow monitoring equipment, signal transmission interruptions, damaged or clogged pipes, etc. (2) The leakage of sewage generated by damaged pipes and the reduction in flow caused by blocked pipes are the main causes of the sudden decrease in instantaneous flow and liquid level. (3) Groundwater, surface water, or rainwater leaks, resulting from pipe breakage, lead to a sudden rise in liquid level and instantaneous flow. In this work, we do not distinguish between these anomalies. Illustrative examples of the anomalies are shown in Table 1, where the instantaneous flow is characterized by 'IF,' and the anomalies are indexed as 'abnormal'.

B. DATA PREPROCESSING

In this section, we describe the establishment of a sliding window in the detection method.

The data preprocessing results are shown in Figure 2. First, we establish a large sliding window, denoted as $T_M(D)$, to capture flow data information for D consecutive time periods. Within this large sliding window $T_M(D)$, we create a smaller sub-window, denoted as $t_K(N)$, with a length of N time periods to obtain data samples that serve as inputs for the model. The model detects these data samples and provides the prediction result for the last $t_{(D-N)}(N)$ data samples captured. The core idea of this detection process is to compare the data samples acquired by $t_{(D-N)}(N)$ with historical data samples. If a change is detected, it may indicate the presence of an anomaly.

The window $t_M(P)$ is known as the data label, which signifies the state for P future time periods following the large window $T_M(D)$. If an anomaly is predicted to occur within the next P time periods, the data label is set to "*abnormal*". It's important to note that the data label window $t_M(P)$ is primarily used to evaluate the model's prediction precision.

TABLE 1. The anomaly examples were selected from three periods: 9/7, 9/12, and 4/12. These samples respectively demonstrate an abrupt change in instantaneous flow (IF) and liquid level to zero, a sudden increase in IF and liquid level, and a sudden reduction in IF and liquid level.

Type	Day	Hour	IF	liquid level	Label
The IF and liquid level suddenly change to zero	9/7	11:20	1.288	0.18	normal
	9/7	11:25	20.929	0.173	normal
	9/7	11:30	4.326	0.168	normal
	9/7	11:35	43.969	0.164	normal
	9/7	**11:40**	**0**	**0**	**abnormal**
	9/7	**11:45**	**0**	**0**	**abnormal**
Sudden increase of IF and liquid level	9/12	7:05	193.759	0.291	normal
	9/12	7:10	153.54	0.281	normal
	9/12	7:15	141.475	0.265	normal
	9/12	7:20	132.941	0.255	normal
	9/12	**7:25**	**]929.935**	**1.237**	**abnormal**
	9/12	**7:30**	**1478.788**	**1.626**	**abnormal**
Sudden decrease of IF and liquid level	4/12	9:00	2103.394	0.575	normal
	4/12	9:05	2015.826	0.58	normal
	4/12	9:10	2004.826	0.575	normal
	4/12	9:15	2007.798	0.581	normal
	4/12	**9:20**	**755.699**	**0.361**	**abnormal**
	4/12	**9:25**	**705.991**	**0.357**	**abnormal**

Once the data samples captured by the large window $T_M(D)$ and the sub-window $t_K(N)$ are input into the model, the detection method enters an offline phase. In this phase, the model analyzes and compares all the data samples to provide prediction results. When new data arrives, the detection method switches back to the online phase. During this phase, both sliding windows $T_M(D)$ and $T_M(P)$ advance by a single time step increment, acquiring new data for the next detection cycle.

The large sliding window $T_M(D)$, the data sample window $t_K(N)$, and the data labelling window $t_M(P)$, respectively:

$$T_M(D) = [t_1(N), t_2(N), t_3(N), \cdots, t_K(N), \cdots, t_{(D-N)}(N)],\quad K = 1, \cdots, D-N \tag{1}$$

$$t_K(N) = [f_1(i), f_2(i), f_3(i), \cdots, f_j(i), \cdots, f_N(i)],\quad i = 1, \cdots, m; j = 1, \cdots, N \tag{2}$$

$$t_M(P) = \begin{cases} normal, & d < [f_{D+1}(i), \cdots, f_{D+P}(i)] < h \\ abnormal, & Otherwise \end{cases} \tag{3}$$

In Equation 1, D represents the length of the large sliding window, and M stands for the number of large sliding windows. In Equation 2, N denotes the number of samples contained in the data sample window samples, K is the number of data samples, and the parameter i signifies the dimension of each sample, including the dimensions of instantaneous flow, liquid level, and flow rate. Finally, since the data is not labeled, we introduce upper bound d and lower bound h as thresholds in Equation 3 to determine whether the data is normal or not.

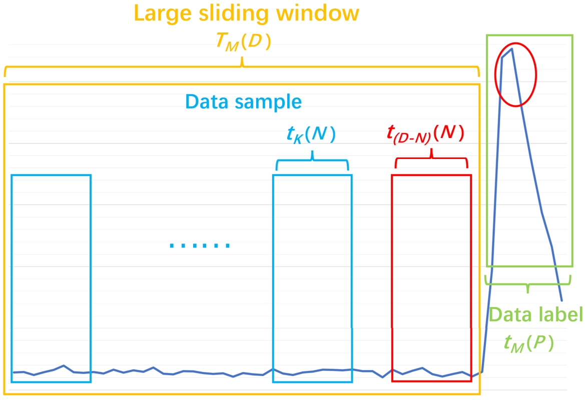


FIGURE 2. This schematic diagram represents data preprocessing. We use the flow data captured in the first window as input. The data label in the second window indicates the status of the flow data for the upcoming P time periods. The second window depicted in the diagram shows an abnormal situation characterized by a sudden increase in flow data. Consequently, the data label is marked as 'abnormal,' signifying the presence of an anomaly during this P time period.

III. TASK DESCRIPTION AND THE MAIN METHOD

In this section, we provide a comprehensive description of our research task and outline the main methodology used. The research task is divided into three main parts, as shown in Figure 3. First, two sliding windows are set up to obtain the data samples to be detected and the data labels. The data labels are only used to assess the accuracy of our method's predictions. Next, we input the data samples into the bagging-based multi-anomaly detection algorithm for detection and output the prediction results of the final intercepted data samples, i.e., the samples intercepted by the window shown in red. The idea behind the detection method is to compare the current data sample with the historical data samples, and if a change occurs, it indicates that an anomaly may occur either in the present or in the future. To assess the precision of the prediction, we compare the prediction results with the data labels captured by the window. Finally, we move the sliding window to ensure continuous monitoring and detection of new data as it becomes available.

A. EARLY AND REAL-TIME ANOMALIES DETECTION TASKS

In this section, we explain the difference between the tasks of early anomalies detection and the tasks of real-time anomalies detection.

In the context of our study, we have defined two different tasks: early anomalies detection and real-time anomalies detection, based on the input samples provided to the model. As shown in Figure 4, the sliding window that has been set up keeps moving and intercepts different data samples. The figure illustrates that the data samples captured by the sliding window at t_{k+n}, while not exceeding the threshold, have deviated from the historical data (t_0 - t_k). When our detection method identifies these deviations, it issues an early warning, performing an early anomalies

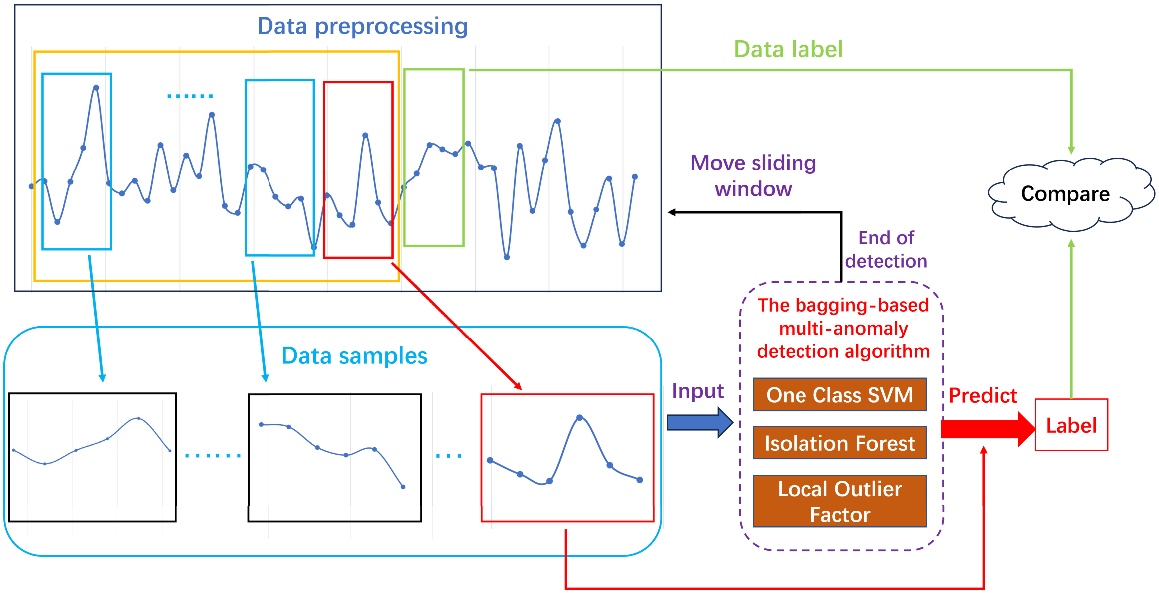


FIGURE 3. Schematic diagram of research task.

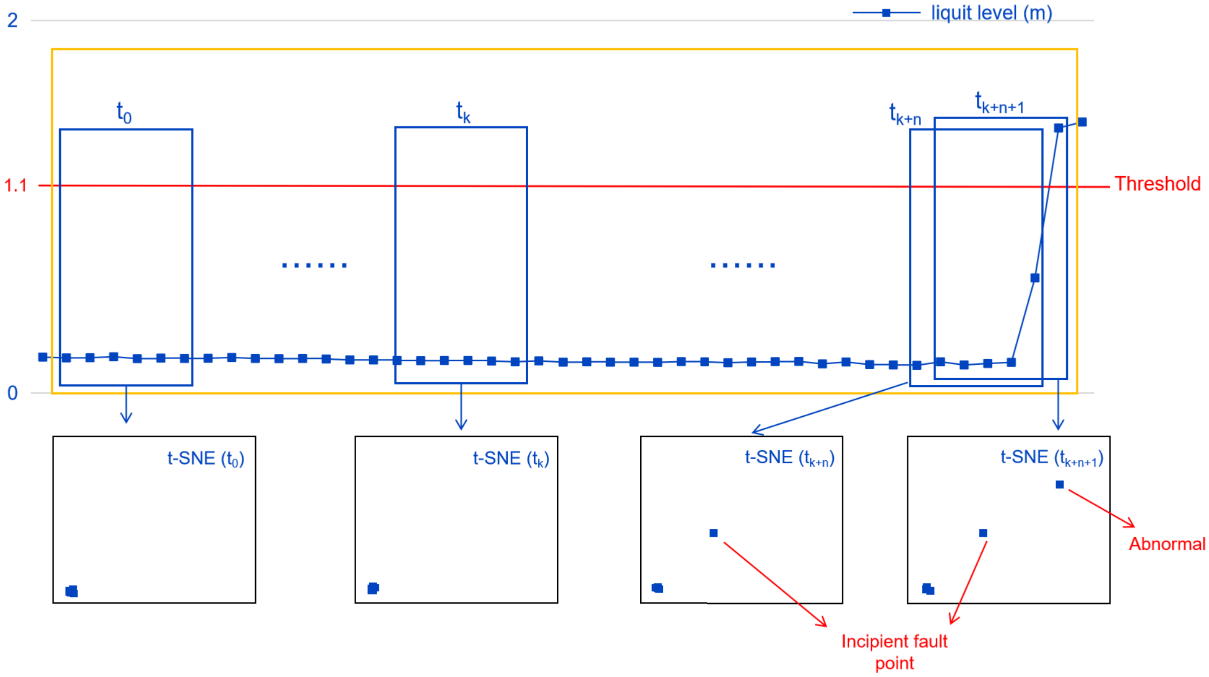


FIGURE 4. Diagram describing the early and real-time anomalies detection tasks.

detection task. These points that deviate from the normal data but do not exceed the threshold are defined by us as incipient fault points. As the sliding window progresses and the t_{k+n+1} intercepted data samples contain anomalous points exceeding the threshold, the prediction result of our method is considered an alarm result, and the detection method performs the real-time anomalies detection task.

B. THE MAIN METHOD

The bagging-based multi-anomaly detection algorithm used in this study is built upon prior research [41], [42]. Previous studies have shown that integrating the outputs of multiple models can enhance the reliability of classification results.

We implemented the bagging-based multi-anomaly detection method using three anomaly detection algorithms: one-class SVM [43], isolation forest [44], [45], and local outlier

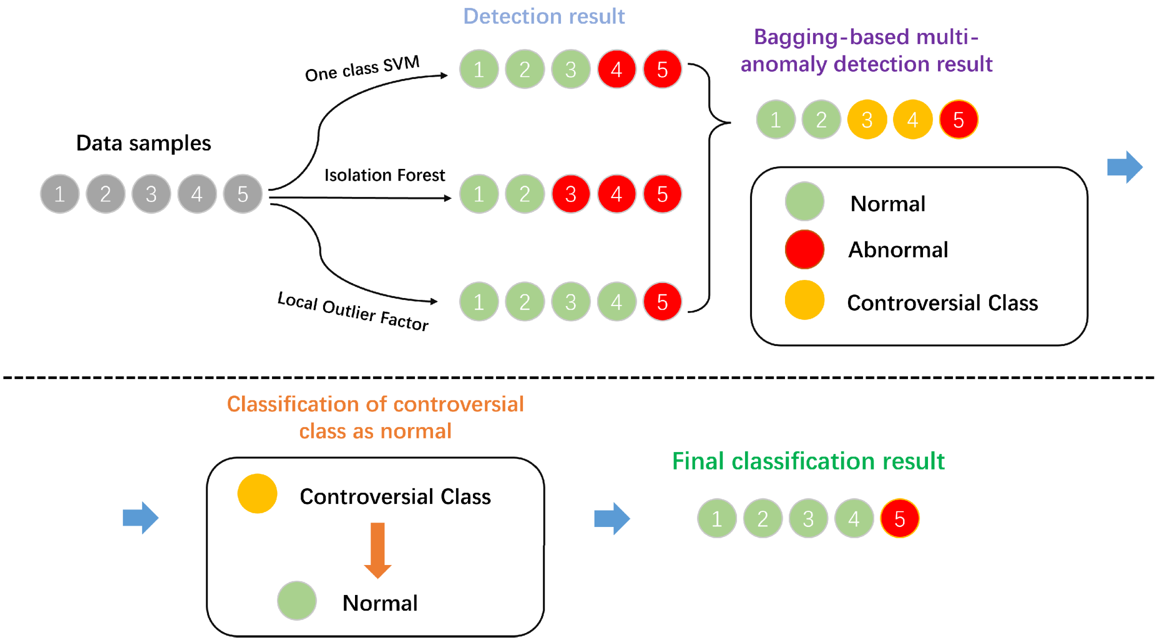


FIGURE 5. **An illustration of the bagging-based multi-anomaly detection algorithm. There are 5 samples numbered 1 to 5. The three anomaly detection algorithms classify these samples into two groups, shown in green and red in the graph. Then, a voting method is applied to determine the classification result. The samples that are classified differently are depicted in yellow. Finally, these controversial samples are considered as normal to obtain the final prediction results, i.e., 1, 2, 3, and 4 are classified as normal, while 5 is classified as abnormal.**

factor [46]. A brief overview of these three methods is provided in the appendix. Figure 5 illustrates the fundamental concept behind the bagging-based multi-anomaly algorithm. In this approach, a sub-sample is classified as abnormal only if all three algorithms concur on its abnormality. While this strategy yields high-confidence predictions, it may lead to lower recall due to subsamples with conflicting votes being classified as normal. However, in practical situations, it is more valuable to make high-confidence predictions for anomalies.

IV. RESULTS AND ANALYSIS

Since the early warning of flow anomalies is the primary focus of this research, this section will concentrate on evaluating the performance of the detection methods in the task of early anomalies detection.

The data are collected by flowmeters placed at different locations in the SNs. Since sewage flow is affected by the location of the SNs and the environment, it results in different patterns of data changes. Consequently, during our sewage flow anomaly detection experiments, we divided the data from different monitoring points to reduce the impact of inconsistent data patterns. We set the size of the data sample window and the data label window to N=8 and P=3, respectively. The length of the large sliding window is set to D=228, and the number of information 'i' for each sample is set to 3, which includes the instantaneous flow, liquid level, and flow velocity. In the bagging-based multi-anomaly detection algorithm, three anomaly detection algorithms are implemented based on scikit-learn. The kernel function of the one-class SVM is selected as the Gaussian kernel (rbf), and the floating-point number 'nu' is set to 0.08. The number of local domain samples 'n-neighbors' in the Local Outlier Factor is set to 20, and the proportion of abnormal sample contamination is set to 0.27. The proportion of anomalous contamination in the Isolation Forest is set to 0.17, the maximum number of features 'max-features' is set to 1, and the number of subtrees is set to 10.

Experimental results demonstrate that our detection method can predict anomalies in flow data 5-15 minutes in advance with a prediction precision of 80.00%, a recall of 66.67%, and an F1 score of 0.73. This performance surpasses that of the traditional thresholding method in the task of early anomalies detection.

A. WARNING TIME AND RESULT OF THE METHOD

The effectiveness of early warning is determined by the lead time provided by the method. The earlier the warning signal is issued, the more effective it is in preventing losses caused by anomalies. To evaluate the early warning performance of our method, we assessed all correctly predicted samples when conducting the early anomalies detection task. Our analysis revealed that the optimal early warning time for the method is 15 minutes, with an effective range of 5 to 15 minutes. As depicted in Figure 6, we can observe that the number of predicted samples gradually increases with the duration of P, which represents the prediction time. However, when P exceeds 3, i.e., extends beyond 15 minutes, the number

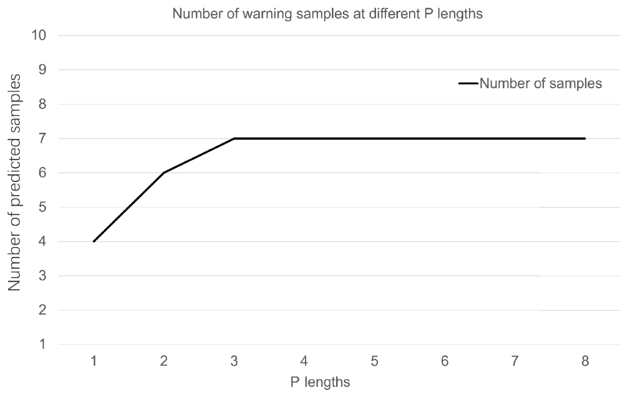


FIGURE 6. The graph illustrates the number of samples predicted at various warning times. The horizontal axis represents the prediction time, while the vertical axis shows the number of predicted samples.

TABLE 2. Comparison of the results of three methods in early anomalies detection tasks.

Method	Prediction time	Recall(%)	Precision(%)	F_1
Static threshold	15 minutes in advance	0.00	0.00	0.20
	10 minutes in advance	0.00	0.00	0.00
	5 minutes in advance	0.00	0.00	0.00
	Anomalies occur	100.00	100.00	1.00
Variable threshold	15 minutes in advance	0.00	0.00	0.00
	10 minutes in advance	0.00	0.00	0.00
	5 minutes in advance	65.41	12.24	0.21
	Anomalies occur	100.00	15.38	0.27
Our method	15 minutes in advance	16.67	50.00	0.25
	10 minutes in advance	33.33	66.67	0.44
	5 minutes in advance	66.67	80.00	0.73
	Anomalies occur	100.00	100.00	1.00

of predicted samples no longer increases. We calculated the precision, recall, and F1 scores for our methods in the context of the early anomalies detection task, as presented in Table 2.

B. COMPARING THE RESULTS OF THE PROPOSED METHOD WITH THE THRESHOLD METHOD IN EARLY ANOMALIES DETECTION TASK

To demonstrate the effectiveness of our approach and the challenges of anomaly alerting, we compare the performance of a bagging-based multi-anomaly detection algorithm with that of a commonly used anomaly detection method (i.e., the thresholding method) in an early anomalies detection task. We used both static and variable thresholding methods [47] in our thresholding approach. A brief overview of all the thresholding methods used is given in the Appendix. Table 2 shows the results of the comparison of the methods in the early anomalies detection task.

In this study, our approach is to compare the difference between current and historical data and then establish an effective adaptive threshold to detect abnormal data changes in order to implement early warning. In Table 2, we can see that static thresholds are more difficult to implement in SNs early warning, and can hardly be used to detect potential changes in data from normal to abnormal, and the method is usually used for alerting tasks. Variable thresholds are generated based on the data and can therefore detect some potential changes in the data from normal to abnormal, thus enabling early warning, but since in SNs flow data fluctuates due to factors such as residential water consumption, these fluctuations are likely to lead to false alarms of variable thresholds, thus reducing the precision of the warning. Our method can be considered as an enhanced adaptive threshold. It effectively detects abnormal data changes and improves robustness to data fluctuations. In addition, our method dynamically determines the threshold based on the data itself and continuously updates it with the arrival of new data. Compared to traditional thresholding techniques, our method performs better, is not affected by environmental biases, and is applicable to different SNs environments.

TABLE 3. Comparison of the results of five methods in a real-time anomalies detection task.

Performance index/Method	Static threshold	Static threshold and deadband	Variable threshold	Variable threshold and deadband	Our method
FAR	0.0039	**0.0039**	0.3864	0.3864	0.1897
MAR	0.2204	0.1667	0.3656	0.2847	**0.1505**

C. COMPARING THE RESULTS OF THE PROPOSED METHOD WITH THE THRESHOLD METHOD IN REAL-TIME ANOMALIES DETECTION TASK

We conducted a comprehensive comparison of our method with static thresholding and variable thresholding, incorporating dead zones to enhance the performance in a real-time anomalies detection task. We adopted FAR and MAR as the evaluation metrics, with lower values indicating superior model performance. In Table 3, it can be observed that the introduction of dead zones effectively reduced the MAR for both static and dynamic thresholding methods. Our method exhibits a higher FAR compared to static thresholding methods that incorporate dead zones, primarily because of the presence of harmful data and densely anomalous data [48]. Furthermore, our method demonstrated a lower MAR when compared to the static thresholding method with the introduction of dead zones. This is primarily due to the inadequacy of threshold settings in static thresholding, resulting in some anomalies being under-considered. In light of these results, further research could further improve detection methods and reduce the impact of dense anomalous or harmful data to increase the effectiveness and precision of detection methods.

D. ANALYSIS OF THE EFFECTIVENESS OF BAGGING

In early anomalies detection tasks, single anomaly detection algorithms typically yield subpar performance. This phenomenon primarily arises due to the complexity of sewage flow data dimensions and the diversity of anomalies, as discussed in previous studies [49], [50]. To address this, we introduce a bagging-based multi-anomaly detection

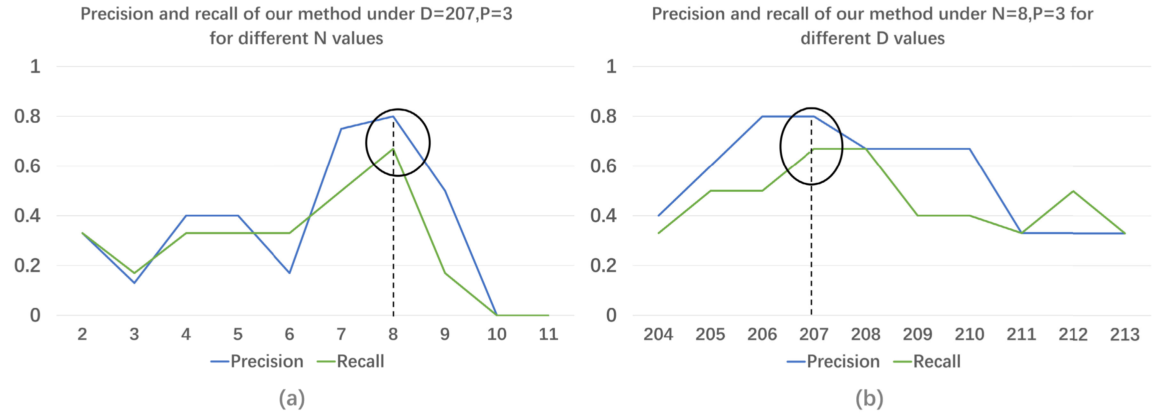


FIGURE 7. The impact of different window lengths, *D* and *N*, on the model's prediction results is shown in Figure 7. In Figure 7(a), the model's prediction precision and recall are presented for varying values of *N*, with *P* fixed at 3 and *D* at 207. Notably, the model achieves the best results when *N* is set to 8. Figure 7(b) displays the model's prediction precision and recall for different values of *D*, while keeping *P* at 3 and *N* at 8. The optimal model prediction results are obtained when *D* is set to 207.

TABLE 4. Comparison of the early warning results of the bagging-based multi-anomaly detection algorithm and the single anomaly detection algorithm.

Method	Recall(%)	Precision(%)	F_1
One class SVM	66.67	13.79	0.23
Isolation forest	83.33	7.69	0.14
Local outlier factor	83.33	5.88	0.11
Our method	66.67	80.00	0.73

algorithm employing a voting mechanism. To assess the method's effectiveness, we compare its results with those of single anomaly detection algorithms, as presented in Table 4. It's evident that there is a substantial improvement in the final prediction results of our model when utilizing the voting method, in contrast to using a single anomaly detection algorithm. However, it's worth noting that, as indicated by the recall metric, our method classifies the prediction results of certain contentious samples as normal. Consequently, this approach does not lead to a significant improvement in recall compared to the single anomaly algorithm; in fact, our method either maintains the lowest recall or even exhibits a lower recall compared to the single anomaly detection algorithm. Nevertheless, in the context of anomaly alerting, the paramount focus is unquestionably on promptly obtaining high-confidence predictions.

E. EFFECT OF DIFFERENT WINDOW LENGTHS ON PREDICTION PRECISION

By setting a suitable large sliding window length D and data sample window length N, the model can effectively obtain the changes in the data, thus improving the prediction precision of the model. We fixed the condition of data label window length P=3 and adjusted the lengths of D and N respectively to observe the prediction precision of our method, and the results are shown in Figure 7. Figure 7(a)(b) shows that when D=207, N=8, the prediction of the detection method achieves the best results. By selecting an appropriate length for the large sliding window (D) and the data sample window (N), the model can effectively capture changes in the data, resulting in improved prediction precision. While keeping the data label window length fixed at P=3, we adjusted the values of D and N to evaluate the prediction precision of our method. The results are presented in Figure 7. Figure 7(a)(b) illustrates that the detection method achieves the best predictions when D=207 and N=8.

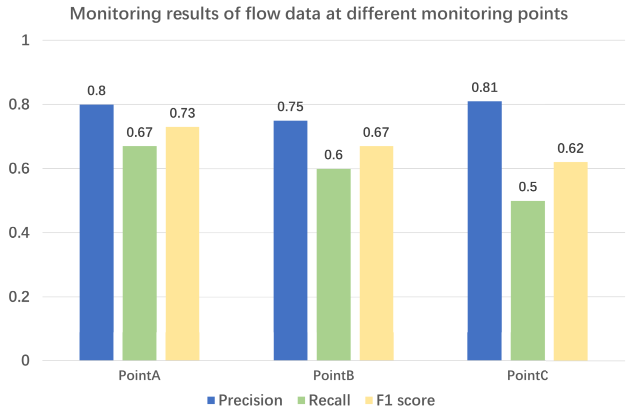


FIGURE 8. Results of the detection method for the early anomalies detection task on different monitoring points.

F. ROBUSTNESS EXPERIMENT OF THE METHOD

Because SNs are situated in various environments, they exhibit different patterns of flow data changes. To validate the effectiveness and robustness of our detection method, we conducted early anomalies detection tasks on data from various monitoring points. The results are illustrated in Figure 8, indicating that our method remains effective in detecting anomalies across different monitoring points while maintaining high precision.

TABLE 5. Performance of different anomaly detection algorithms in early anomalies detection tasks.

Method	Recall(%)	Precision(%)	F_1
One class SVM	66.67	13.79	0.23
Isolation forest	83.33	7.69	0.14
Local outlier factor	83.33	5.88	0.11
Connectivity-based Outlier Factor	66.67	4.23	0.08
SVDD	0	0	0
DBSCAN	0	0	0

TABLE 6. The results are based on the combination of one class SVM, local outlier factor, connectivity-based outlier factor based on the bagging voting mechanism.

Method	Recall(%)	Precision(%)	F_1
One class SVM	66.67	13.79	0.23
Local outlier factor	83.33	5.88	0.11
Connectivity-based outlier factor	66.67	4.23	0.08
Bagging	50.00	66.67	0.57

V. CONCLUSION AND OUTLOOKS

In this study, we present a method for early warning of sewage flow anomalies in SNs. This method is implemented within a sliding window monitoring framework and leverages a bagging-based multi-anomaly detection algorithm based on unsupervised bagging. Early warnings are generated by comparing current data to historical data, enabling the detection of abnormal changes and the provision of timely alerts. Notably, our detection method doesn't rely on data labeling and effectively addresses data bias resulting from environmental variations within SNs. In experiments conducted using sewage flow data from Erhai Lake SNs, our method effectively detects anomalies. It is data-agnostic and free from domain bias, with the potential for future applications in anomaly detection tasks across various industries.

Currently, our method has certain limitations, including relatively short warning times and the opportunity for further improvement in precision and recall. To address these challenges, we intend to explore advanced deep learning techniques that enhance feature distinctions between current and historical data, leading to higher precision, recall, and earlier warnings. Additionally, we can delve into the specific classification of these anomalies as a research problem to be addressed in the future.

APPENDIX

A. ANOMALY DETECTION ALGORITHM SELECTION

We have selected several conventional machine learning anomaly detection algorithms for the task of early anomalies detection, as presented in Table 5. It is evident that the algorithms in the first four categories of methods in the table show some promise in early anomalies detection. However, the precision of these algorithms still falls short of the desired ideal. To enhance their performance, we applied the bagging method to combine each of these four classes of methods. The final results of the two best-performing combinations are shown in Table 4 and Table 6, respectively.

Table 4 showcases the best-performing prediction results in this study. In contrast, Table 6 combines the local outlier factor and the connectivity-based outlier factor, which belong to the same category of algorithms. Consequently, their prediction results exhibit similarity, leading to no significant improvement after applying the bagging voting mechanism. On the other hand, Table 4 combines three algorithms with different anomaly detection principles, enabling diverse angles and perspectives in obtaining prediction results. This approach yields higher precision through the bagging voting mechanism.

B. A BRIEF REVIEW OF THE ONE CLASS SVM

One class SVM. The algorithm was given by Schölkopf et al. [43] to deal with binary classification. The algorithm works by dividing the positive examples of a given sample into a specific region, and returns $+1$ if the sample is within that region, indicating a positive sample. Otherwise, it returns -1, indicating a negative sample. The algorithm for one-class SVM is outlined below: If one-dimensional training data $X_i \in R^n (i = 1, 2, 3, \ldots., n)$ is set, there is: $\min_{\omega,i,\xi,\rho} (1/2)\omega^T \omega - \rho + (1/vn) \sum_{i=1}^{n} \xi_i$ and $\omega^T \Phi(x_i) \geq \rho - \xi_i$ and $xi_i \geq 0, i = 1, 2, \ldots., n$. Then solving by transformation yields: $\min_a (1/2)\alpha^T Q_{ij} \alpha$ and $0 \leq \alpha_i \leq (1/vn), i = 1, 2, .., n$ and $e^T \alpha = 1$. Finally, the decision and sign functions are: $f(x) = sign(\sum_{i=1}^{n} K(x, x_i - \rho))$, $g(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) - \rho$.

C. A BRIEF REVIEW OF THE ISOLATION FOREST

Isolation forest. This algorithm conception is an unsupervised learning proposed by Fei et al. [44] which can detect dataset quickly. Anomaly detection can be performed without building any data model and any label description. Anomalies can be found simply by estimating the data in turn.

The isolation forest algorithm operates by placing the data on isolated trees [45] and then separating the data based on their characteristics. Data with similar characteristics are grouped together while dissimilar data are separated. Since anomalies have significant differences compared to normal data, they are likely to be separated early in the process, making it more likely that the first separated data is anomalous data. This makes the isolation forest algorithm effective in quickly detecting anomalies without the need for a labeled dataset or complex modeling.

D. A BRIEF REVIEW OF THE LOCAL OUTLIER FACTOR

Local outlier factor. The algorithm, introduced by Breunig et al. [46], calculates a numerical score to reflect the degree of anomalousness of a sample. It does so by examining the average density of the sample points around each individual sample point. If the density of a particular point is found to be less than the density of the surrounding sample points by a factor of greater than 1, then it is likely to be an outlier. If the opposite is true, then the point is likely to be a normal data point. This algorithm is particularly effective in situations where the data points are not uniformly distributed and are instead grouped in clusters of varying densities.

E. A BRIEF REVIEW OF THE STATIC THRESHOLD

Static threshold. The algorithm mainly sets a fixed threshold to diagnose whether the data is abnormal or not. This method is one of the most commonly used methods in alarm systems, and the method is achieved as follows:

$$Result = \begin{cases} 1, & d < S < h \\ -1, & Otherwise \end{cases} \tag{4}$$

h and d are the upper and lower bounds of the threshold value, respectively, and S is the data. When the value of the data is within the threshold boundary, 1 is returned to indicate that the data is normal, otherwise -1 is returned to indicate an exception.

F. A BRIEF REVIEW OF THE VARIABLE THRESHOLD

Variable threshold. The main idea of the algorithm is that the threshold value changes with time and is achieved in the following steps: $\overline{v} = \gamma v(k-1) + (1-\gamma)v(k)$, $\overline{m} = \gamma m(k-1) + (1-\gamma)m(k)$, $T(k) = \overline{m}(k) \mp \alpha\overline{v}(k)$. Where $m(k)$ is the mean of the data intercepted by the sliding window, γ is the momentum factor and takes values in the range [0,1], $v(k)$ is the variance of the data intercepted by the sliding window, k is the window length (number of samples in each window), α is an adjustable factor, and $T(k)$ is variable threshold value in the kth window [47], [48].

G. A BRIEF REVIEW OF THE DEADBAND

Deadband Method. This method is designed to improve the ability of the algorithm to resist harmful data in the alarm system, for clearing another limit of alarms. Static threshold and deadband refer to the establishment of a deadband based on the percentage of the high and lower limits of the established threshold [51], [52]. Variable threshold and deadband have the same deadband setting, but the deadband changes with the threshold value. The deadband can be achieved as follows [47]: $deadband\ width(DB) = H(1-db)$ $for\ high\ limit$, $deadband\ width(DB) = D(1+db) for\ lower\ limit$. Where H and D is the threshold value, and db is deadband value.

H. EVALUATION METRIC

The performance of the bagging-based multi-anomaly detection method proposed in this study is evaluated by the precision [53], recall [54], f1 score [55], MAR [56], and FAR [57] of the marked anomalies.

Precision describes how many of the two classifiers are true positive examples from the perspective of prediction results. The main research of precision is the authenticity of classification as anomalies to judge the quality of the bagging-based multi-anomaly detection method proposed in this study.

Recall describes how many real positive examples are selected by the second classifier in the test set from the perspective of reality. It's divided into anomalies types in this studies, which are really anomalies.

F1 score is a measurement index of classification problems, referred to as F_1. It's the harmonic average of precision and recall. The maximum is 1 and the minimum is 0. In general, the larger the index, the better the performance of the model. The calculation method of F_1 is as follows: $F_1 = 2 \cdot precision \cdot recall/(precision + recall)$.

MAR and FAR are two indicators commonly used to evaluate alarm systems, in the alarm task, which when these two indicators are lower, the better the performance of the model, in the confusion matrix, FP represents negative samples predicted as positive samples, TN represents negative samples predicted as negative samples, FN represents positive samples predicted as negative samples, TP represents positive samples predicted as positive samples, and the two indicators are calculated as follows: $FAR = FP/(FP + TN)$, $MAR = FN/(TP + FN)$.

REFERENCES

[1] Y. Li, J. Bräunig, P. K. Thai, M. Rebosura, J. F. Mueller, and Z. Yuan, "Formation and fate of perfluoroalkyl acids (PFAAs) in a laboratory-scale urban wastewater system," *Water Res.*, vol. 216, Jun. 2022, Art. no. 118295.

[2] Y. Liu, P. Ramin, X. Flores-Alsina, and K. V. Gernaey, "Transforming data into actionable knowledge for fault detection, diagnosis and prognosis in urban wastewater systems with AI techniques: A mini-review," *Process Saf. Environ. Protection*, vol. 172, pp. 501–512, Apr. 2023.

[3] K. Thiyagarajan, S. Kodagoda, R. Ranasinghe, D. Vitanage, and G. Iori, "Robust sensor suite combined with predictive analytics enabled anomaly detection model for smart monitoring of concrete sewer pipe surface moisture conditions," *IEEE Sensors J.*, vol. 20, no. 15, pp. 8232–8243, Aug. 2020.

[4] I. Pikaar, K. R. Sharma, S. Hu, W. Gernjak, J. Keller, and Z. Yuan, "Reducing sewer corrosion through integrated urban water management," *Science*, vol. 345, no. 6198, pp. 812–814, Aug. 2014.

[5] K. Thiyagarajan, "Robust sensor technologies combined with smart predictive analytics for hostile sewer infrastructures," Ph.D. dissertation, Univ. Technol. Sydney, Ultimo, NSW, Australia, Jul. 2018.

[6] P. F. Boulos and A. T. Walker, "Fixing the future of wastewater systems with smart water network modeling," *J. AWWA*, vol. 107, no. 4, pp. 72–80, Apr. 2015.

[7] J. B. Haurum and T. B. Moeslund, "A survey on image-based automation of CCTV and SSET sewer inspections," *Autom. Construct.*, vol. 111, Mar. 2020, Art. no. 103061.

[8] A. Alshami, M. Elsayed, S. R. Mohandes, A. F. Kineber, T. Zayed, A. Alyanbaawi, and M. M. Hamed, "Performance assessment of sewer networks under different blockage situations using Internet-of-Things-based technologies," *Sustainability*, vol. 14, no. 21, p. 14036, Oct. 2022.

[9] E. Okwori, M. Viklander, and A. Hedström, "Spatial heterogeneity assessment of factors affecting sewer pipe blockages and predictions," *Water Res.*, vol. 194, Apr. 2021, Art. no. 116934.

[10] M. R. Bintang, K. Rossa Sungkono, and R. Sarno, "Time and cost optimization in feasibility test of CCTV project using CPM and PERT," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Jul. 2019, pp. 678–683.

[11] M. Wang, H. Luo, and J. C. P. Cheng, "Towards an automated condition assessment framework of underground sewer pipes based on closed-circuit television (CCTV) images," *Tunnelling Underground Space Technol.*, vol. 110, Apr. 2021, Art. no. 103840.

[12] M. Harshini, J. Philip, I. Haritha, and S. Patil, "Sewage pipeline fault detection using image processing," in *Proc. 6th Int. Conf. Electron., Commun. Aerosp. Technol.*, Dec. 2022, pp. 1181–1185.

[13] Y. Wang, P. Li, and J. Li, "The monitoring approaches and non-destructive testing technologies for sewer pipelines," *Water Sci. Technol.*, vol. 85, no. 10, pp. 3107–3121, Apr. 2022.

[14] J. Latif, M. Z. Shakir, N. Edwards, M. Jaszczykowski, N. Ramzan, and V. Edwards, "Review on condition monitoring techniques for water pipelines," *Measurement*, vol. 193, Apr. 2022, Art. no. 110895.

[15] Y. Yu, A. Safari, X. Niu, B. Drinkwater, and K. V. Horoshenkov, "Acoustic and ultrasonic techniques for defect detection and condition monitoring in water and sewerage pipes: A review," *Appl. Acoust.*, vol. 183, Dec. 2021, Art. no. 108282.

[16] H. Noshahri, M. van der Meijde, and L. O. Scholtenhuis, "GPR surveys in enclosed underground sewer pipe space," *Tunnelling Underground Space Technol.*, vol. 129, Nov. 2022, Art. no. 104689.

[17] H. Noshahri, L. L. O. Scholtenhuis, M. V. Delft, J. Franco, and E. Dertien, "Towards underground void detection with in-pipe ground penetrating radar," in *Proc. 3rd Asia Pacific Meeting Near Surf. Geosci. Eng.*, Nov. 2020, pp. 1–5.

[18] D. Wang, I. D. Moore, N. Hoult, and H. Lan, "Evaluation and comparison of different detection technologies on simulated voids near buried pipes," *Tunnelling Underground Space Technol.*, vol. 123, May 2022, Art. no. 104440.

[19] M. Li, M. Li, Q. Ren, H. Liu, and C. Liu, "Intelligent identification and classification of sewer pipeline network defects based on improved RegNetY network," *J. Civil Struct. Health Monitor.*, vol. 13, nos. 2–3, pp. 547–560, Dec. 2022.

[20] C. Xiong, S. Lian, and W. Chen, "An ensemble method for automatic real-time detection, evaluation and position of exposed subsea pipelines based on 3D real-time sonar system," *J. Civil Structural Health Monitor.*, vol. 13, nos. 2–3, pp. 485–504, Nov. 2022.

[21] M. J. Chae, W. Kim, and H. K. Hwang, "Digital pipeline image scanning and intelligent data management for sewer pipelines," ASCE, Apr. 2008, pp. 1–11.

[22] S. Foorginezhad, M. Mohseni-Dargah, K. Firoozirad, V. Aryai, A. Razmjou, R. Abbassi, V. Garaniya, A. Beheshti, and M. Asadnia, "Recent advances in sensing and assessment of corrosion in sewage pipelines," *Process Saf. Environ. Protection*, vol. 147, pp. 192–213, Mar. 2021.

[23] M. H. Abbas, R. Norman, and A. Charles, "Neural network modelling of high pressure CO_2 corrosion in pipeline steels," *Process Saf. Environ. Prot.*, vol. 119, pp. 36–45, Oct. 2018.

[24] W. Kaempfer and M. E. Berndt, "Estimation of service life of concrete pipes in sewer networks," *Durability Building Mater. Compon.*, vol. 8, pp. 36–45, May 1999.

[25] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer life span prediction: Comparison of methods and assessment of the sample impact on the results," *Water*, vol. 11, no. 12, p. 2657, Dec. 2019.

[26] G. Del Giudice, R. Padulano, and D. Siciliano, "Multivariate probability distribution for sewer system vulnerability assessment under data-limited conditions," *Water Sci. Technol.*, vol. 73, no. 4, pp. 751–760, Oct. 2015.

[27] M. Ahammed and R. E. Melchers, "Probabilistic analysis of underground pipelines subject to combined stresses and corrosion," *Eng. Struct.*, vol. 19, no. 12, pp. 988–994, Dec. 1997.

[28] V. Aryai, H. Baji, M. Mahmoodian, and C.-Q. Li, "Time-dependent finite element reliability assessment of cast-iron water pipes subjected to spatio-temporal correlated corrosion process," *Rel. Eng. Syst. Saf.*, vol. 197, May 2020, Art. no. 106802.

[29] H. Chen, Z. Liu, C. Alippi, B. Huang, and D. Liu, "Explainable intelligent fault diagnosis for nonlinear dynamic systems: From unsupervised to supervised learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 8, 2022, doi: 10.1109/TNNLS.2022.3201511.

[30] S. Russo, M. D. Besmer, F. Blumensaat, D. Bouffard, A. Disch, F. Hammes, A. Hess, M. Lürig, B. Matthews, C. Minaudo, E. Morgenroth, V. Tran-Khac, and K. Villez, "The value of human data annotation for machine learning based anomaly detection in environmental systems," *Water Res.*, vol. 206, Nov. 2021, Art. no. 117695.

[31] P. Kazemi, J. Giralt, C. Bengoa, A. Masoumian, and J.-P. Steyer, "Fault detection and diagnosis in water resource recovery facilities using incremental PCA," *Water Sci. Technol.*, vol. 82, no. 12, pp. 2711–2724, Aug. 2020.

[32] N. Li, X. Wang, Z. Li, F. Zhao, A. Nair, J. Zhang, and C. Liu, "Real-time identification and positioning of sewer blockage based on liquid level analysis in rural area," *Processes*, vol. 11, no. 1, p. 161, Jan. 2023.

[33] J. L. B. Krajewski, F. C. Meyer, and M. Lepot, *Metrology in Urban Drainage and Stormwater Management: Plug and Pray*. London, U.K.: IWA, Aug. 2021.

[34] B. Sun, S. Chen, Q. Liu, Y. Lu, C. Zhang, and H. Fang, "Review of sewage flow measuring instruments," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 2089–2098, Jun. 2021.

[35] C. Zhao, H. Liu, Y. Guo, and L. An, "Sewage system diagnosis based on online monitoring technology," *Water Int.*, vol. 48, no. 3, pp. 322–330, May 2023.

[36] F. Larrarte, M. Lepot, F. H. C. Meyer, J. L. Bertrand-Krajewski, D. Ivetic, D. Prodanovic, and B. Stegeman, "Water level and discharge measurements," in *Metrology in Urban Drainage and Stormwater Management: Plug and Pray*. London, U.K.: IWA, Aug. 2021.

[37] T. Giakoumis and N. Voulvoulis, "Combined sewer overflows: Relating event duration monitoring data to wastewater systems' capacity in England," *Environ. Sci., Water Res. Technol.*, vol. 9, no. 3, pp. 707–722, Mar. 2023.

[38] Y. Jiang, C. Li, L. Sun, D. Guo, Y. Zhang, and W. Wang, "A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks," *J. Cleaner Prod.*, vol. 318, Oct. 2021, Art. no. 128533.

[39] J. O. Ighalo, A. G. Adeniyi, and G. Marques, "Internet of Things for water quality monitoring and assessment: A comprehensive review," in *Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications*, Sep. 2020, pp. 245–259.

[40] J. Li, K. Sharma, W. Li, and Z. Yuan, "Swift hydraulic models for real-time control applications in sewer networks," *Water Res.*, vol. 213, Apr. 2022, Art. no. 118141.

[41] C. Zhou, Y. Gu, G. Fang, and Z. Lin, "Automatic morphological classification of galaxies: Convolutional autoencoder and bagging-based multiclustering model," *Astronomical J.*, vol. 163, no. 2, p. 86, Jan. 2022.

[42] Y.-J. Hou, Z.-X. Xie, and C.-C. Zhou, "An unsupervised deep-learning method for fingerprint classification: The CCAE network and the hybrid clustering strategy," 2021, *arXiv:2109.05526*.

[43] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[44] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

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

[46] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, May 2000.

[47] K. Aslansefat, M. Bahar Gogani, S. Kabir, M. A. Shoorehdeli, and M. Yari, "Performance evaluation and design for variable threshold alarm systems through semi-Markov process," *ISA Trans.*, vol. 97, pp. 282–295, Feb. 2020.

[48] M. Bahar-Gogani, K. Aslansefat, and M. A. Shoorehdeli, "A novel extended adaptive thresholding for industrial alarm systems," in *Proc. Iranian Conf. Electr. Eng. (ICEE)*, May 2017, pp. 759–765.

[49] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep isolation forest for anomaly detection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12591–12604, Dec. 2023.

[50] S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *J. Big Data*, vol. 7, no. 1, pp. 1–30, Jul. 2020.

[51] E. Naghoosi, I. Izadi, and T. Chen, "A study on the relation between alarm deadbands and optimal alarm limits," in *Proc. Amer. Control Conf.*, Jun. 2011, pp. 3627–3632.

[52] A. J. Hugo, "Estimation of alarm deadbands," *IFAC Proc. Volumes*, vol. 42, no. 8, pp. 663–667, 2009.

[53] X. Yin, Y. Chen, A. Bouferguene, H. Zaman, M. Al-Hussein, and L. Kurach, "A deep learning-based framework for an automated defect detection system for sewer pipes," *Autom. Construct.*, vol. 109, Jan. 2020, Art. no. 102967.

[54] S. S. Kumar, D. M. Abraham, M. R. Jahanshahi, T. Iseley, and J. Starr, "Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks," *Autom. Construct.*, vol. 91, pp. 273–283, Jul. 2018.

[55] Q. Xie, D. Li, J. Xu, Z. Yu, and J. Wang, "Automatic detection and classification of sewer defects via hierarchical deep learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1836–1847, Oct. 2019.

[56] J. Xu, J. Wang, I. Izadi, and T. Chen, "Performance assessment and design for univariate alarm systems based on FAR, MAR, and AAD," *IEEE Trans. Autom. Sci. Eng.*, vol. 9, no. 2, pp. 296–307, Apr. 2012.

[57] X. Xu, S. Li, X. Song, C. Wen, and D. Xu, "The optimal design of industrial alarm systems based on evidence theory," *Control Eng. Pract.*, vol. 46, pp. 142–156, Jan. 2016.

CHUNMING QIU received the B.E. degree in biomedical engineering from the Sichuan University of Light and Chemical Engineering, Yibin, China, in 2021. He is currently pursuing the M.E. degree in electronic information with Dali University, Dali, China.

His current research interests include environmental big data, machine learning, and deep learning applications in the field of environmental engineering.

GUOXIANG SHAO received the B.E. degree in automation from the Beijing University of Information Technology, in 2020, and the M.E. degree in electronic information from Dali University, in 2022.

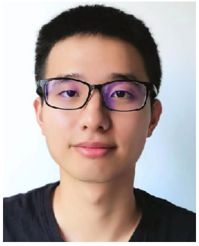
He is currently with the Yunnan Academy of Science and Technology, Kunming, China. His current research interests include information management of cities based on machine learning and deep learning.

ZHENYU ZHANG received the Ph.D. degree in environmental engineering from the Kunming University of Science and Technology, Yunnan, China, in 2020.

He is currently with Dali University. He is also the Director and a Senior Engineer with the Engineering Research Center for Air-Ground Integrated Intelligence and Big Data Application, Yunnan Provincial Department of Education. His research interest includes environmental information and prediction.

CHICHUN ZHOU received the Ph.D. degree in material physics and chemistry from Tianjin University, Tianjin, China, in 2018.

He is currently with Dali University. His research interests include mathematical physics, deep learning environment applications, and data mining.



YUEJIE HOU received the B.E. degree in communication engineering from the Zhongshan College, University of Electronic Science and Technology, in 2020, and the M.E. degree in electronic information from Dali University, in 2022.

He is currently working in Shenzhen, China. His research interests include AI applications in the intersection of environmental and biological sciences and large language model development.

ENMING ZHAO received the Ph.D. degree in mechanical design and theory from Harbin Engineering University, Harbin, China, in 2013.

He is currently with Dali University. His research interests include ecological environment informatization and machine vision technology.

XIAO GUO received the Ph.D. degree in control science and engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2013.

He is currently employed at the Institute of Unmanned Systems, Beihang University. His research interest includes machine learning.

XIAOLIN GUAN received the B.E. degree in electronic information from the Wenzhou Institute of Technology, Wenzhou, China, in 2021. He is currently pursuing the M.E. degree in electronic information with Dali University, Dali, China.

His current research interests include deep learning in environmental engineering and wastewater instrumentation flow meter measurement accuracy improvement applications.