Research on Abnormal Data Identification and Early Warning Model for Continuous Automatic Monitoring and Control of Ecological Environment

Jing Yu[1], Jing Bian[1], Xintao Ye[1], Lei Zhu[1], Yanchun Chang[2]
[1]Ecological and Environmental Monitoring Center of Zhejiang Province, Hangzhou 310012, Zhejiang, China
[2]Huzhou Ecological and Environmental Monitoring Center of Zhejiang Province, Huzhou 313000, Zhejiang, China
Corresponding author: Yanchun Chang (e-mail: 13868290995@163.com)

ABSTRACT In the context of digital transformation in ecological environment governance, continuous automatic monitoring systems have become key technical support for pollution tracing, quality assessment, and risk warning. However, monitoring data are easily affected by instrument failures, environmental interference, transmission anomalies, and other factors, which restricts data availability and the scientific validity of early-warning decisions. To address the problems of insufficient robustness, delayed warning, and weak generalization in traditional anomaly identification methods, this paper proposes an integrated solution based on multi-feature fusion, hierarchical recognition, and dynamic warning. A four-dimensional anomaly feature system covering data integrity, temporal stability, physical constraints, and spatial correlation is first constructed. Sliding-window statistics and multi-scale feature extraction are used to represent anomaly patterns. A hierarchical recognition model combining improved Isolation Forest and LSTM is then designed, where the improved IForest rapidly screens significant anomalies and the LSTM captures temporal dependencies for precise classification. The proposed method improves anomaly recognition and warning reliability, and provides technical references for environmental sensing networks, electromagnetic interference-resistant monitoring, and intelligent signal processing.

INDEX TERMS ecological environment, abnormal data identification, early warning model, sensing networks, signal processing

I. INTRODUCTION

A. RESEARCH BACKGROUND

WITH the deepening of the "dual carbon" target and the modernization strategy of ecological environment governance, continuous automatic monitoring systems (such as air quality automatic monitoring stations, water quality automatic monitoring buoys, soil moisture monitoring equipment) have been deployed on a national scale. As of 2024, China has established an ecological environment automatic monitoring network covering 338 prefecture level cities and over 1500 counties, forming an integrated monitoring pattern of "sky ground". The daily output of monitoring data exceeds 10TB, providing massive data support for environmental quality assessment, pollution tracing, and emergency response [1,2]. However, continuous automatic monitoring systems operate in complex outdoor environments for a long time, and the data acquisition and transmission process is susceptible to multiple factors: instrument level (sensor drift, pipeline blockage, power failure), environmental level (extreme temperature, heavy rainfall, electromagnetic interference), transmission level (network interruption, data packet loss, protocol anomalies), etc., resulting in various types of outliers such as missing, abrupt changes, drift, and outliers in the monitoring data [3,4].

B. RESEARCH SIGNIFICANCE

1) Theoretical Significance

Propose an improved hierarchical recognition model that integrates isolated forests with LSTM to address the challenges of traditional machine learning methods in capturing temporal dependencies and high training costs for deep learning models, providing a new paradigm for anomaly recognition of temporal monitoring data.

Establish a dynamic matching mechanism between anomaly types and warning thresholds, improve the logical link from anomaly identification to risk warning, and fill the theoretical gap of the disconnect between "identification warning" in existing research [5].

2) Practical Significance

Optimize environmental warning decision-making: achieve real-time identification and rapid warning of abnormal data, provide accurate guidance for handling sudden pollution incidents and instrument failures, shorten response time, and reduce environmental risk losses.

Support monitoring network operation and maintenance: By analyzing abnormal data types, locate instrument failures, transmission defects, and other issues, provide data support for monitoring equipment operation and network optimization, and improve the stability of monitoring network operation.

C. RESEARCH CONTENT AND METHODS

1. Construction of abnormal feature system for ecological environment monitoring data: Systematically analyze the causes and manifestations of abnormal data, construct a four-dimensional abnormal feature system covering integrity, temporal stability, physical constraints, and spatial correlation, and design feature extraction methods.

2. Hierarchical anomaly recognition model design: Based on the improved Isolation Forest (IForest) and LSTM network, a hierarchical model of "coarse recognition fine classification" is designed to quickly screen for significant anomalies and accurately classify anomaly types.

3. Establishment of dynamic warning mechanism: Combining anomaly types, severity, and environmental quality standards, establish a dynamic threshold warning model to achieve a closed-loop process from anomaly identification to risk warning.

4. Empirical verification and optimization of the model: Based on actual monitoring data, verify the recognition accuracy, generalization ability, and warning effect of the model, and optimize the model parameters.

II. ANALYSIS OF ABNORMAL CHARACTERISTICS AND TYPES OF ECOLOGICAL ENVIRONMENT MONITORING DATA

A. ABNORMAL DATA GENERATION MECHANISM

The abnormal occurrence of continuous automatic monitoring data of ecological environment is the result of multiple factors such as instruments, environment, transmission, and human factors. The specific mechanism of occurrence is as follows:

Instrument factors: sensor aging drift, detection light path pollution, pipeline blockage, power supply fluctuations, calibration abnormalities, etc., leading to data being high, low, or fluctuating abnormally;

Environmental factors: extreme temperatures (below $-20°C$ or above $40°C$), heavy rainfall, strong electromagnetic interference, sandstorm weather, etc., interfere with the normal operation of sensors, resulting in sudden changes or distorted data;

Transmission factors: network interruption, data loss, abnormal transmission protocol, server failure, etc., leading to data loss, duplication, or confusion;

Human factors: construction around monitoring stations, human interference with sampling ports, non-standard calibration operations, etc., leading to local data anomalies [6-9].

B. CLASSIFICATION OF ABNORMAL DATA TYPES

Based on the data representation and causes, ecological environment monitoring data anomalies are classified into six categories:

Missing anomalies: Blank values or invalid values (such as "-999" or "NaN") that occur during data collection or transmission, divided into two categories: single point missing and continuous missing;

Drift anomaly: Data slowly shifts over time (such as SO_2 concentration monitoring values consistently being 10% - 20% higher), mainly caused by sensor aging and untimely calibration;

Outlier anomaly: A single data point significantly deviates from the overall data distribution, but there is no obvious anomaly in the preceding and following data, often caused by accidental environmental interference;

Repetitive anomaly: The same data appears in multiple consecutive monitoring cycles, caused by transmission protocol failure or instrument jamming;

Logical anomaly: Data that violates physical and chemical laws or monitoring logic (such as negative pollutant concentration, unreasonable positive correlation between O_3 concentration and NO_2 concentration) [10-12].

Logical anomaly: Data that violates physical and chemical laws or monitoring logic (e.g., negative pollutant concentration, unreasonable positive correlation between O_3 and NO_2) (essential anomaly, primary judgment criterion).

Hierarchical Priority (from high to low): Logical anomalies > Sudden Change Anomalies > Drift Anomalies > Outlier Anomalies > Duplication Anomalies > Missing Anomalies. When an anomaly fits multiple definitions, the primary criterion (logical validity) takes precedence.

C. CONSTRUCTION OF FOUR-DIMENSIONAL ANOMALY FEATURE SYSTEM

To comprehensively characterize abnormal data patterns, a four-dimensional anomaly feature system covering integrity, temporal stability, physical constraints, and spatial correlation is constructed:

1) Integrity Characteristics

Used to describe missing and duplicated data, including: Missing rate: the proportion of missing data within the sliding window (with a window size of 24 monitoring cycles, i.e. 24 hours);

Continuous missing duration: the number of monitoring cycles for continuous missing data;

Repetition rate: the proportion of duplicate data within the sliding window;

Continuous repetition duration: the number of monitoring cycles for continuous repeated data [13].

2) Characteristics of Temporal Stability

Used to capture the temporal variation patterns of data, including: Absolute rate of change: A robust formula is adopted to avoid zero-value denominator issues:

$$\text{Absolute Rate of Change} = \frac{|x_t - x_{t-1}|}{x_{t-1} + \epsilon}.$$

Where $\epsilon = 0.001$ (set based on monitoring data precision, much smaller than the minimum measurable value of $0.01\,\mu\text{g/m}^3$). When $x_{t-1} = 0$, the rate reflects the relative difference between current data and the minimum measurable value, which is consistent with practical monitoring needs;

Sliding standard deviation: the standard deviation of data within the sliding window;

Trend slope: The linear regression slope of data within a sliding window;

Coefficient of variation: the ratio of the standard deviation to the mean of data within a sliding window [14].

3) Physical Constraints Characteristics

Design of physical and chemical properties based on environmental monitoring data, including: Reasonability of value: Whether the data is within the reasonable range of instrument measurement and environment (such as $PM_{2.5}$) The reasonable concentration range is 0-1000 $\mu\text{g/m}^3$;

Indicator correlation: Pearson correlation coefficient between relevant monitoring indicators (such as NO_2 and O_3 concentrations, which are usually negatively correlated);

Rationality of concentration gradient: Whether the change in data between adjacent monitoring cycles is within a reasonable gradient (such as a single change in SO_2 concentration not exceeding $50\,\mu\text{g/m}^3$) [15,16].

Note: Physical constraints refer to objective physical-chemical laws and instrument measurement ranges that monitoring data must comply with.

4) Spatial Correlation Characteristics

Design spatial correlation using data from adjacent monitoring stations, including: Spatial consistency: the mean absolute deviation between the current site data and the contemporaneous data of the three nearest surrounding sites;

Consistency of spatial change trend: the average difference in trend slope between the current site and surrounding site data;

Spatial anomaly synchronization rate: the proportion of anomalies in the same period data of surrounding stations [17].

D. FEATURE EXTRACTION METHODS

Design corresponding extraction methods for different types of features:

Statistical features (such as missing rate, sliding standard deviation, coefficient of variation): extracted using the sliding window method, with the window size dynamically adjusted according to the monitoring period (24 for hourly monitoring and 120 for minute monitoring);

Trend features (such as trend slope and spatial consistency): extracted through linear regression model fitting;

Correlation features (such as indicator correlation and spatial consistency): calculated using methods such as Pearson correlation coefficient and mean absolute deviation;

Logical features (such as reasonable values and indicator correlation): extracted based on preset thresholds and domain knowledge judgments [18].

III. DESIGN OF ABNORMAL DATA RECOGNITION MODEL

A. OVERALL MODEL ARCHITECTURE

Design a hierarchical anomaly recognition model of "coarse recognition fine classification" to address the limitations of traditional methods. The model is divided into three core modules: data preprocessing module, improved IForest coarse recognition module, and LSTM fine classification module. The data preprocessing module implements data cleaning, standardization, and feature extraction; Improve the IForest module to quickly screen for significant abnormal data and reduce the computational complexity of subsequent models;

The LSTM module further classifies the coarse recognition results, accurately identifies abnormal types, and improves recognition accuracy [19].

B. DATA PREPROCESSING

Missing value handling: Single point missing data is filled using linear interpolation, while continuous missing data exceeding 4 cycles is filled using K-nearest neighbor (KNN) interpolation;

Invalid value handling: Delete obvious erroneous data (such as negative pollutant concentration or extreme values beyond the instrument measurement range);

Duplicate value processing: Retain the first occurrence of data and delete subsequent duplicate data.

C. IMPROVE THE IFOREST COARSE RECOGNITION MODULE

Suspected anomaly screening: Set a three-level judgment standard of "significant anomaly normal suspected anomaly", where the average path length is significantly below the threshold as a significant anomaly, significantly above the threshold as normal, and close to the threshold as a suspected anomaly to avoid overlooking minor anomalies.

Optimize model parameters through grid search method: set the number of isolated trees to 100, the sample subset size to 256, the sliding window size to 24 (hourly monitoring), and the weighting coefficients to 0.35 for temporal features, 0.3 for spatial features, 0.2 for physical features, and 0.15 for integrity features [20].

Suspected anomaly screening: A three-level judgment standard with quantitative boundaries is set, based on the statistical distribution of the average path length of samples:

Let the average path length of sample x be L(x), the mean of normal sample path lengths be μ, and the standard deviation be σ:

Significant anomaly: $L(x) < \mu - 1.2\sigma$;

Normal: $L(x) > \mu + 0.8\sigma$;

Suspected anomaly: $\mu - 1.2\sigma \leq L(x) \leq \mu + 0.8\sigma$.

Calibration Process: μ and σ are calculated from normal samples in the training set (sample size $\geq$ 10,000). For the experimental data, $\mu = 8.6$ and $\sigma = 1.3$, leading to: Significant anomaly threshold: $8.6 - 1.2 \times 1.3 = 6.94$;

Suspected anomaly upper limit: $8.6 + 0.8 \times 1.3 = 9.64$.

The coefficients (1.2, 0.8) are calibrated through 3 rounds of cross-validation to cover 95% of normal sample distribution.

D. LSTM FINE CLASSIFICATION MODULE

The structure of the LSTM fine classification model is as follows:

Input layer: 15 feature variables with four-dimensional anomalous feature system as input dimensions;

Hidden layer: Set up 3 layers of LSTM hidden layers, with 64, 32, and 16 neurons in each layer, and use ReLU as the activation function;

Dropout layer: Add a dropout layer between the LSTM layer and the fully connected layer, with a dropout rate of 0.2, to prevent overfitting;

Fully connected layer: set 1 fully connected layer with 8 neurons;

Output layer: Using Softmax activation function, output 7 types of results (6 types of abnormal+normal).

Model training optimization:

Loss function: Cross entropy loss function with class balance weights (abnormal sample weight = 2.5, normal sample weight = 1);

Optimizer: Adam optimizer with learning rate 0.001 and decay rate 0.0001;

E. CLASS IMBALANCE HANDLING

Oversampling: SMOTE algorithm (k = 5 nearest neighbors, synthetic sample ratio 2:1) is used to expand abnormal samples from 8,503 to 21,258, making the ratio of normal to abnormal samples close to 2:1; Parameter Calibration: SMOTE parameters are calibrated based on data distribution characteristics to avoid overfitting.

IV. CONSTRUCTION OF DYNAMIC WARNING MODEL

A. EARLY WARNING INDICATOR SYSTEM

Based on the results of anomaly recognition and environmental monitoring requirements, construct a three-dimensional warning indicator system of "anomaly type severity impact range":

1) Abnormal Type Indicators

According to the six types of abnormal data, the warning priority is divided into: sudden pollution induced mutation anomaly (priority 1)> drift anomaly caused by instrument failure (priority 2)> logical anomaly (priority 3)> outlier anomaly (priority 4)> duplicate anomaly (priority 5)> missing anomaly (priority 6).

2) Severity Indicators

Based on the degree of deviation of abnormal data from the normal range, it is divided into three levels:

Mild abnormality: Data deviates from the normal range by 10% -20%, or a single indicator shows short-term abnormalities;

Moderate abnormality: Data deviates from the normal range by 20% -50%, or multiple indicators are synchronized abnormally;

Severe anomaly: The data deviates from the normal range by more than 50%, or is abnormal for multiple consecutive cycles.

3) Impact Scope Indicators

Based on spatial correlation features, it is divided into three levels:

Local impact: Only the current site is abnormal, and there are no abnormalities in surrounding sites;

Regional impact: The current site is experiencing synchronization issues with 1-2 surrounding sites;

Widespread impact: The current site is experiencing synchronization anomalies with three or more surrounding sites.

B. DETERMINATION OF DYNAMIC WARNING THRESHOLD

Using a combination weighting method combining Analytic Hierarchy Process (AHP) and Entropy Weight Method to determine the weights of each warning indicator, and then determining the warning threshold based on the weighted sum result:

Expert Selection Criteria: 5 experts participated, meeting three conditions:

(1) $\geq$ 8 years of experience in ecological environment monitoring;

(2) Senior Engineer or above title;

(3) Participated in $\geq$ 3 environmental early warning system projects;

AHP Consistency Check: A judgment matrix is constructed using the 1-9 scale method. The consistency ratio (CR) of the target layer is 0.062, and the average CR of the indicator layer is 0.078 (both < 0.1), meeting consistency requirements;

Weight Fusion: AHP subjective weights and Entropy Weight objective weights are integrated to determine the final weights.

Classify warning levels based on comprehensive scores:

Blue warning (Level IV): $1.0 \leq S < 2.0$, suggestive warning, need to pay attention to data changes;

Yellow warning (Level III): $2.0 \leq S < 3.0$, general warning, need to investigate the cause of the abnormality;

Orange warning (Level II): $3.0 \leq S < 4.0$, important warning, requiring immediate verification and disposal;

Red warning (Level I): $S \geq 4.0$, emergency warning, emergency response needs to be activated.

C. WARNING RESPONSE MECHANISM

Establish a closed-loop response mechanism of "identification grading push disposal feedback":

1. Exception recognition: The model identifies and classifies abnormal data in real-time; 2. Warning classification: Determine the warning level based on the comprehensive score; 3. Information push: Push warning information to operation and maintenance personnel and environmental management personnel through SMS, platform pop ups, etc., including abnormal types, severity, impact scope, and disposal suggestions;

4. Emergency response: Operations personnel handle emergencies based on warning information (such as instrument calibration and network repair), while environmental management personnel handle sudden pollution incidents;

5. Result feedback: After the disposal is completed, the results will be fed back to the system, and the model will adjust the threshold and parameters based on the feedback.

V. EMPIRICAL VERIFICATION AND RESULT ANALYSIS

A. EXPERIMENTAL DATA PREPARATION

Invite 3 experts in the field of ecological environment monitoring to manually annotate the data based on monitoring station operation and maintenance records, weather data, and pollution event records. The annotation results include normal data (54857 records) and 6 types of abnormal data (8503 records). The distribution of abnormal data is shown in Table 1:

Experimental hardware environment: Intel Core i7-12700K processor, 32GB memory, NVIDIA RTX 3090 graphics card;

Experimental software environment: Python 3.9, TensorFlow 2.8, Scikit learn 1.2, Pandas 1.5, NumPy 1.24.

B. EVALUATION INDICATORS

Adopting 5 commonly used evaluation indicators in the field of anomaly recognition:

Accuracy: The proportion of correctly identified samples to the total number of samples;

Precision: The proportion of true cases to predicted positive cases;

Recall: the proportion of true cases to actual positive cases;

F1 Score: The harmonic mean of precision and recall;

False positive rate (FAR): The proportion of false positive cases to the predicted positive cases.

C. PERFORMANCE VERIFICATION OF ANOMALY RECOGNITION MODEL

Overall Performance of the Model The overall performance indicators of the proposed hierarchical anomaly recognition model are shown in Table 2:

The results showed that the accuracy of the model reached 96.8%, the F1 score reached 95.0%, the false alarm rate was only 4.2%, and the average recognition time was 3.8 ms/item, meeting the requirements of real-time monitoring.

Analysis: After class imbalance handling, the recall rate of abnormal data increased from 94.3% to 95.7%, and the difference in F1 scores between normal and abnormal data narrowed from 2.2 to 1.1, effectively reducing model bias.

1) Comparison with Traditional Methods

The proposed model was compared with traditional methods (3 σ criterion, standard IForest, single LSTM), and the results are shown in Table 3:

The comparison results show that the proposed model outperforms traditional methods in accuracy, precision, recall, and F1 score. The false positive rate is reduced by 12.6 percentage points compared to the 3 σ criterion, 6.3 percentage points compared to the standard IForest, and 2.0 percentage points compared to a single LSTM; Although the recognition time is slightly longer than the 3 σ criterion and standard IForest, it is much shorter than a single LSTM, balancing recognition accuracy and real-time performance.

2) Recognition Performance of Different Types of Anomalies

The recognition performance of the proposed model for different types of anomalies is shown in Table 4:

The results showed that the model had the highest recognition accuracy for missing and duplicate anomalies (accuracy>98%), while the recognition accuracy for outlier anomalies and drift anomalies was relatively low.

However, the F1 scores exceeded 94%, indicating that it could effectively identify various types of abnormal data.

D. PERFORMANCE VERIFICATION OF WARNING MODEL

1) Warning Accuracy and Response Time

Based on 100 actual abnormal cases (including instrument failures, sudden pollution, transmission anomalies, etc.), the performance of the warning model was validated, and the results are shown in Table 5:

The results showed that the overall accuracy of the warning model reached 93.8%, with an average response time of 4.3 minutes, meeting the needs of emergency response; The response time for both red and orange warnings is within 5 minutes, providing quick guidance for handling sudden pollution incidents and instrument malfunctions.

2) Typical Case Analysis

Based on the $PM_{2.5}$ abnormal concentration of a monitoring station in an industrial zone on August 15, 2023 as an example for case analysis:

Abnormal performance: 8:00-10:00 $PM_{2.5}$ concentration suddenly increased from $65\,\mu g/m^3$ to $289\,\mu g/m^3$, and then remained consistently high;

Recognition results: Improved IForest coarse recognition identified significant anomalies, LSTM subdivision classi-

TABLE 1. Distribution of Abnormal Data Types

Abnormal Type	Missing Anomaly	Sudden Change Anomaly	Drift Anomaly	Outlier Anomaly	Duplication Anomaly	Logical Anomaly	Total
Quantity (items)	2861	1532	1387	1246	873	604	8503
Proportion (%)	33.6	18.0	16.3	14.7	10.3	7.1	100

TABLE 2. Overall Performance Metrics of the Model

Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	False Alarm Rate (%)	Average Recognition Time (ms)
96.8	95.7	94.3	95.0	4.2	3.8

TABLE 3. Performance Comparison of Different Methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	False Alarm Rate (%)
3σ Criterion	81.5	83.2	79.6	81.4	16.8
Standard IForest	88.1	89.5	86.7	88.1	10.5
Single LSTM	94.2	93.8	92.5	93.1	6.2
Proposed Model	96.8	95.7	94.3	95.0	4.2

TABLE 4. Recognition Performance of Different Abnormal Types

Abnormal Type	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Missing Anomaly	98.5	97.8	96.9	97.3
Sudden Change Anomaly	97.2	96.5	95.8	96.1
Drift Anomaly	96.3	95.2	94.1	94.6
Outlier Anomaly	95.7	94.8	93.5	94.1
Duplication Anomaly	98.1	97.5	96.7	97.1
Logical Anomaly	97.6	96.8	95.3	96.0
Normal Data	98.3	97.6	96.8	97.2

TABLE 5. Performance Verification Results of the Early Warning Model

Warning Level	Number of Cases	Warning Accuracy (%)	Average Response Time (min)
Red Warning	12	91.7	3.2
Orange Warning	28	92.9	4.1
Yellow Warning	35	94.3	4.5
Blue Warning	25	96.0	5.0
Overall	100	93.8	4.3

fied as "mutation anomalies+drift anomalies", determined as anomalies caused by sudden pollution;

Warning result: Priority 1 for abnormal types, severe severity, and regional impact, with a comprehensive score of 4.2, triggering a red warning;

Disposal result: Upon receiving the warning, environmental management personnel immediately checked and found that a nearby chemical plant had violated emissions regulations. They promptly took measures to stop production and rectify the situation to avoid the spread of pollution;

Verification result: The warning was accurate, the response time was 3.5 minutes, the disposal was timely, and there was no large-scale environmental impact.

E. MODEL GENERALIZATION ABILITY VERIFICATION

To verify the generalization ability of the model, monitoring data from 8 water quality automatic monitoring stations in another province (including 5 indicators such as pH, COD, and ammonia nitrogen) were selected for testing:

The results showed that the accuracy of the model in water quality monitoring data reached 95.3%, F1 score reached 93.9%, and false alarm rate was 5.7%. Although slightly lower than the recognition performance of air quality monitoring data, it still maintained a high level, proving that the model has good generalization ability.

VI. MODEL OPTIMIZATION AND APPLICATION SUGGESTIONS

A. MODEL PARAMETER OPTIMIZATION

Based on empirical results, further optimize the model parameters:

Improve the sliding window size of IForest: adjust according to the monitoring scenario, set the hourly monitoring to 24 and the minute monitoring to 120; LSTM network layers: For scenarios with more complex features such as water quality monitoring data, an additional LSTM hidden layer can be added;

Warning threshold: For high pollution areas such as industrial zones, the warning threshold can be lowered by 0.2 to improve sensitivity; In areas with less pollution such as suburbs, the threshold can be increased by 0.2 to reduce the false alarm rate.

B. LIGHTWEIGHT OPTIMIZATION FOR EDGE DEPLOYMENT

LSTM Network Pruning: Reduce 3 hidden layers to 2 layers, adjust neurons from 64-32-16 to 48-24; Parameter Quantization: Adopt INT8 quantization technology, compressing model volume by 62%; IForest Structure Optimization: Reduce the number of isolated trees from 100 to 60. The lightweight model maintains 95.6% accuracy (only 1.2% loss) and meets the deployment requirements of soil moisture monitoring equipment.

C. PRACTICAL APPLICATION SUGGESTIONS

1) Data Quality Control

Establish data preprocessing standards: unify data cleaning and standardization methods to ensure the quality of input model data;

Regularly calibrate feature weights: recalculate feature weights every 6 months based on monitoring equipment aging, environmental changes, and other factors.

2) System Deployment Plan

Distributed deployment: Distributed deployment is adopted in provincial monitoring networks, with one model node deployed in each region to improve data processing efficiency;

Real time data update: The model updates training data once per hour, dynamically adjusts parameters, and adapts to changes in data distribution.

3) Suggestions for Operation and Maintenance Management

Establish an abnormal handling ledger: record warning information, handling process, and handling results to form a closed-loop management;

Regular model evaluation: Evaluate the model performance once every quarter and optimize parameters and structure based on the evaluation results.

VII. CONCLUSION AND PROSPECT

A. RESEARCH CONCLUSION

A four-dimensional anomaly feature system covering integrity, temporal stability, physical constraints, and spatial correlation has been constructed with 15 variables to comprehensively characterize industrial discharge patterns and lay the foundation for improving recognition accuracy.

The model has shown good recognition performance and generalization ability in air quality and water quality monitoring data, and can be widely applied to various types of ecological environment automatic monitoring systems.

B. RESEARCH PROSPECTS

Future research can be further deepened in the following areas:

1. Multi source data fusion: Introducing meteorological data, pollution source emission data, and satellite remote sensing data to enrich the anomaly feature system and improve the accuracy of anomaly recognition in complex scenarios.
2. Model lightweight: for edge computing device deployment requirements, model compression, quantification and other technologies are used to achieve model lightweight and reduce computing resource consumption.
3. Intelligent decision fusion: Integrating anomaly recognition and warning models with pollution tracing, operation and maintenance scheduling decision models to build an integrated ecological environment monitoring intelligent decision system.
4. Multi scenario adaptation: Optimize the feature system and model parameters for other monitoring scenarios such as soil, noise, and radiation, and expand the application scope of the model.

AUTHOR CONTRIBUTIONS

Conceptualization –Yujing, Bianjing, Ye xintao, Zhu lei and Chang yanchun; methodology – Yujing, Bianjing, Ye xintao, Zhu lei and Chang yanchun; investigation – Yujing, Bianjing, Ye xintao, Zhu lei and Chang yanchun; writing-original draft preparation – Yujing, Bianjing, Ye xintao, Zhu lei and Chang yanchun. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] W. Alahmad, S. Quqa, F. Ubertini, and C. Gentilini, “Anomaly identification in historical modular structures using satellite data: Methodology and application to Portico di San Luca, Bologna.Journal of Building Engineering,” vol.118,no.114893,2026, doi: 10.1016/J.JOBE.2025.114893.
[2] Y. Chen, G. Ban, and T. Ding, “Abnormal data recognition method for wind turbines based on alpha channel fusion,” Applied Energy, vol. 396, Art. no. 126261, 2025, doi: 10.1016/J.APENERGY.2025.126261.
[3] M. Zhang, X. Hu, and G. Dogan, “Recognizing Normal and Unusual Activities for Parkinson’s Disease using Multimodal Sensor Data,” International Journal of Activity and Behavior Computing, vol. 2025, no. 2, pp. 1-16, 2025, doi: 10.60401/IJABC.114.
[4] J. Li, Y. Lv, Z. Zhou, Z. Du, Q. Wei, and K. Xu, “Identification and Correction of Abnormal, Incomplete Power Load Data in Electricity Spot Market Databases,” Energies, vol. 18, no. 1, pp. 176, 2025, doi: 10.3390/EN18010176.
[5] Y. Chen, K. Wang, M. Zhao, Y. Xiong, C. Li, and J. Liu, “Identification and reconstruction of anomalous data in dam monitoring considering temporal correlation,” Smart Materials and Structures, vol. 32, no. 11, 2023, doi: 10.1088/1361-665X/ACF970.
[6] Z. Kang, S. Lin, A. Wang, Y. He, and L. Chen, “Identification of Abnormal Data for Synchronous Monitoring of Transformer DC Bias Based on Multiple Criteria,” Sensors (Basel, Switzerland), vol. 23, no. 10, 2023, doi:10.3390/S23104959.

[7] B. Lin, “Research on abnormal data recognition method of optical network based on WiFi triangular location,” International Journal of Autonomous and Adaptive Communications Systems, vol. 16, no. 4, pp. 357, 2023, doi: 10.1504/IJAACS.2023.132506.

[8] F. Li, Y. Yang, H. Xu, and F. Zuo, “Multi-node abnormal data identification in distribution networks based on deep learning,” Journal of Physics: Conference Series, vol. 2358, no. 1, 2022, doi: 10.1088/1742-6596/2358/1/012013.

[9] D. Bin, T. Yu, Y. Ling, W. Chen, L. Lu, and H. Chen, “Power distribution abnormal data identification and processing system based on correlation classification,” Journal of Physics: Conference Series, vol. 2290, no.1, 2022, doi: 10.1088/1742-6596/2290/1/012027.

[10] S. Liu, Z. Ye, L. Yi, and M. Li, “Identification method of abnormal data of electric energy measurement based on BP neural network algorithm,” Journal of Physics: Conference Series, vol. 2260, no. 1, 2022, doi: 10.1088/1742-6596/2260/1/012040.

[11] Q. Tan, X. Mu, M. Fu, H. Yuan, J. Sun, G. Liang, et al., “A new sensor fault diagnosis method for gas leakage monitoring based on the naive Bayes and probabilistic neural network classifier,” Measurement, pp. 194, 2022, doi: 10.1016/J.MEASUREMENT.2022.111037.

[12] C. Yan, Z. Luo, Z. Lin, H. He, Y. Luo, and J. Chen, “Shear Wave Elastography-Assisted Ultrasound Breast Image Analysis and Identification of Abnormal Data,” Journal of Healthcare Engineering, vol. 2022, Art. no. 5499354, 2022, doi: 10.1155/2022/5499354.

[13] Y. Qiu, L. Jing, and S. Li, “Bridge anomaly data identification method based on statistical feature mixture and data augmentation through forwarding difference,” IOP Conference SeriesEarth and Environmental Sci -ence, vol. 791, no. 1, 2021, doi: 10.1088/1755-1315/791/1/012030.

[14] E. Aref-Eshghi, J. Oake, M. Godwin, K. Aubrey-Bassler, P. Duke, M. Mahdavian, et al., “Identification of Dyslipidemic Patients Attending Primary Care Clinics Using Electronic Medical Record (EMR) Data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Database,” Journal of Medical Systems, vol. 41, no. 3, pp. 45, 2017, doi: 10.1007/s10916-017-0694-7.

[15] L. Xing, K. Demertzis, and J. Yang, “Identifying data streams anomalies by evolving spiking restricted Boltzmann machines,” Neural Computing and Applications, vol. 32, no. 11, pp. 1-15, 2019, doi: 10.1007/s00521-019-04288-5.

[16] J. P. Kim, T. G. Ahn, J. Kim, and H. T. Yoon, “Autonomous Train Speed-Limit Estimation Algorithm by Recognition Abnormal Situation Based on of Infrastructure Sensor Data,” The Journal of Korean Institute of Communications and Information Sciences, vol. 43, no. 4, pp. 708-714, 2018, doi: 10.7840/kics.2018.43.4.708.

[17] Q. Qiu, Z. Wang, J. Tian, Y. Tu, X. Cui, C. Hu, et al., “Correction of Fused Rainfall Data Based on Identification and Exclusion of Anomalous Rainfall Station Data,” Water, vol. 15, 2023, doi: 10.3390/W15142541.

[18] Z. Zhang, “Application and Research of Data Mining Technology in Communication Network Environment,” Advanced Materials Research, vol. 3326, no. 989-994, pp. 3814-3817, 2014, doi: 10.4028/www.scientific.net/AMR.989-994.3814.

[19] S. Huang, G. Zhao, J. Guo, W. Zhu, and F. Yuan, “Design of power quality abnormal data identification and verification system,” International Conference on Mechanical Engineering, Measurement Control, and Instrumentation. Bellingham, WA, USA: SPIE; 2021:119301Z, doi: 10.1117/12.2611725.

[20] H. Cui, “A DBSCAN Anomaly Data Identification Method Driven by Data Distribution,” Department of Information and Control Engineering, Shenyang Institute of Science and Technology, China, 2025, doi: 10.54254/2755-2721/2025.20861.