2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) | 979-8-3315-2228-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/AINIT65432.2025.11035420

Anomaly Identification and Prediction Method for Network Timing Data Based on Improved Isolated Forest Approach

[1st]Xianlang hu
Jiangsu Automation Research Institute/Harbin Engineering University
Lianyungang, China
sya88070887@163.com

[2nd]Zhongliang Zhang
Jiangsu Automation Research Institute
Lianyungang, China
453901511@qq.com

[3rd]Ruini Wang*
Jiangsu Automation Research Institute
Lianyungang, China
wangruini1994@163.com

[4th]Chuntian Hu
Jiangsu Automation Research Institute
Lianyungang, China
spring_111@163.com

***Abstract*—Aiming at the problem of insufficient real-time and poor adaptability to dynamic changes in network timing data anomaly detection, this paper proposes an anomaly identification and prediction method based on temporal convolutional network and isolated forest. The method firstly uses time convolution network to extract features from network timing data, effectively capturing the timing dependence and long-term trend in the data. Subsequently, the isolated forest algorithm is used for fast anomaly identification and prediction of the extracted features, and the accuracy and efficiency of anomaly detection is improved by optimising the hyperparameters of the isolated forest. Experiments demonstrate that the method proposed in this paper can significantly improve the accuracy and real-time performance of anomaly detection compared with the traditional isolated forest method, and provide lightweight and adaptive anomaly warning support for dynamic network environments.**

Keywords—Isolated forests; temporal convolutional networks; network timing data; anomaly identification and prediction

I. INTRODUCTION

With the rapid development of cloud computing, IoT and edge computing, Linux OS has become the core support platform for server clusters, distributed systems and smart terminal devices by virtue of its open source, high stability and customisability. However, with the exponential growth in the scale of network services, network timing data of the Linux system presents features such as high dynamics, strong noise interference and complex timing correlation. In this context, the difficulty of detecting network attacks and system anomalies increases significantly. Traditional anomaly detection methods are difficult to balance real-time, accuracy and environment adaptability, resulting in Linux systems facing serious security risks. Therefore, how to construct lightweight and highly robust anomaly identification and prediction models has become a key challenge for securing dynamic network environments.

Currently, anomaly detection methods for network timing data are mainly classified into three categories: methods based on statistical models, rule engines and unsupervised learning. Statistical models detect deviations by fitting the distribution pattern of historical data, but their modelling ability for non-linear timing features is insufficient, especially in Linux multi-service concurrency scenarios prone to high false alarm rates. Rule engines rely on predefined attack feature libraries, and although they can quickly identify known threats, the lag in updating rule libraries leads to a drastic decrease in detection performance when facing zero-day attacks or complex environment changes[1]. In recent years, unsupervised learning methods have gained widespread attention due to the advantage of not requiring labelled data, among which isolation forest quickly isolates anomalies by randomly dividing the feature space, and demonstrates high real-time performance in Linux network traffic detection[2]. However, the core defect of Isolation Forest is that it assumes that the data are independently and identically distributed, and ignores the contextual dependencies in the network timing data, which results in insufficient sensitivity to the detection of slow-drifting anomalies, and is susceptible to the interference of Linux system noises , which generates a large number of false alarms.

To break through the above limitations, researchers try to introduce deep learning techniques to enhance the timing modelling capability. Long Short-Term Memory Network (LSTM) and Temporal Convolutional Network (TCN) are widely used in traffic prediction and classification tasks due to their advantages in long sequence modelling[3-4]. Among them, TCN can efficiently capture long-term dependencies by expanding causal convolution and residual connection structures, and significantly outperforms LSTM in terms of computational efficiency; however, TCN is mostly used in traffic prediction or pattern classification, and its potential in anomaly detection has not been fully explored. Meanwhile, although Isolated Forest performs well in static data analysis, its hyperparameters are usually fixed, which is difficult to adapt to the dynamically changing characteristics of Linux network traffic . Therefore, how to effectively integrate the temporal modelling capability of TCN with the fast detection advantage of isolated forest, and design an adaptive parameter optimization strategy for the Linux environment has become a core issue to improve the accuracy of anomaly detection.

In this paper, we propose a network timing data anomaly identification and prediction method for Linux system, which adopts a two-stage model of 'TCN feature extraction-improved isolated forest detection'(TCN-I_IF). Firstly, TCN is used to extract deep features of the original network timing data to capture long-term trends and local mutations. Second, the features are fed into the isolated forest model, and dynamic hyperparameter optimisation is proposed: incremental training based on sliding windows, feature importance weighting and Bayesian optimisation to search for optimal parameters. The method aims to adapt to the dynamic changes of Linux traffic in real time, improve the sensitivity and accuracy of anomaly detection, and reduce the cost of manual parameter adjustment.

II. Related Work

Network timing data anomaly detection, as a key technology to guarantee system stability, has received extensive attention in academia and industry in recent years. Isolation Forest has become one of the core algorithms in this field due to its unsupervised feature and linear time complexity. Traditional Isolation Forest achieves fast isolation of anomalies by randomly dividing the feature space, but its ability to capture time-dependent features is weak, especially when dealing with data with significant periodicity and suddenness such as Linux server traffic, which is prone to decrease in detection accuracy due to the high dimensionality of the features and the complexity of the patterns. In order to improve this problem, scholars have conducted a lot of research.Zhao et al. proposed a Linux log anomaly detection method based on an improved isolated forest algorithm [5]. The method adjusts the features and sample points of concern by introducing the dynamics of the concern mechanism. Literature [6] proposed an improved isolated forest algorithm through data fusion, feature extraction and modal decomposition, and experiments show that the AUC value of the new method is closer to 1. The paper [7] proposed an isolated forest weak target detection method based on improved multi-feature correlation, which effectively improves the detection performance of the target.Gong et al. proposed a traffic flow anomaly detection model based on the improved isolated forest and K-Means++ algorithm [8]. of traffic flow anomaly detection model [8], which significantly improves the detection accuracy by constructing the anomaly scoring model and sliding window threshold.

The evolution of feature extraction methods for time-series data provides new ideas for anomaly detection. Early studies relied on statistical features to construct feature sets, but manually designed features are difficult to capture nonlinear time-series patterns. The introduction of deep learning techniques significantly improves the feature characterisation ability. Literature [9] proposed a two-layer multi-head self-attention model for the difficult problem of tariff forecasting in wind power high-share systems by using load and wind power to construct new features, and by combining the attention mechanism with the time-sequence convolutional network. Literature [10] proposed a short-term cloud resource prediction model based on TCN-LSTM, and the experiments showed that the MAE and RMSE were significantly reduced, which verified the effectiveness of the model. Literature [11] proposed a PV power prediction model based on TCN, which has a dual focus mechanism, and experiments show that the prediction accuracy and computational efficiency are significantly improved.

In recent years, the research of combining deep learning feature extraction with traditional machine learning models has become a hot spot. Literature [12] proposed a method of anomaly detection and reconstruction of bridge monitoring data based on LSTM neural network, and the experiments showed that the MAPE and R2 indexes were excellent, and the detection and reconstruction effect was remarkable. Literature [13] proposed a CBT-Net flight trajectory prediction method, fusing meteorological and airspace control data, CNN-BiLSTM-Transformer model to optimise the prediction accuracy, and combining GNN self-supervised learning to improve the anomaly detection capability. However, the existing methods suffer from the problem of poor compatibility between TCN high-dimensional features and isolated forests and the problem of performance fluctuation caused by sudden changes in network traffic.

In summary, the isolated forest algorithm has made significant progress through feature selection and dynamisation improvement, and the introduction of TCN provides a new path for time-series feature extraction, but the combination of the two still has challenges in terms of lightweighting, real-time and environment adaptability. Existing research focuses on single-module optimisation, and lacks a systematic solution from feature space alignment to collaborative model updating, especially in Linux network monitoring scenarios, where algorithmic complexity needs to be further balanced with the limitations of edge computing resources. In this paper, we address the above issues and propose a fusion architecture of improved TCN feature extraction and adaptive isolated forests, aiming to achieve the goal of high-precision, low-latency and dynamically robust anomaly detection.

III. Dynamic Prediction and Optimization Method of PRP Network Congestion Based On GAT

This chapter proposes a two-stage anomaly detection framework based on temporal convolutional networks (TCNs) with improved isolated forests, whose core process includes temporal feature extraction based on TCNs, anomaly scoring modelling with improved isolated forests, and a dynamic hyper-parameter optimisation mechanism.

A. TCN temporal feature extraction model

For network timing data (e.g., sequences such as CPU occupancy, TCP retransmission rate, etc., collected by Linux system), TCN achieves modelling of long and short-term dependencies by Dilated Causal Convolution and Residual Connection Residual Connection to achieve the modelling of long and short term dependencies.

Define the input sequence as $X \in \mathrm{R}^{T\times d}$ (d is the feature dimension), the lth layer convolution kernel weight $W^{(l)} \in \mathrm{R}^{k\times d\times m}$(k is the size of the convolution kernel, m is the number of output channels), and the expansion factor is $2^{(l-1)}$. The output feature $H^{(l)}$ of layer l is computed as:

$$H^{(l)} = \sigma(W^{(l)} *_d H^{(l-1)}) \tag{1}$$

where $*_d$ denotes the expansion convolution operation and σ is the ReLU activation function. The inflationary convolution expands the receptive field by interval sampling, and its output position t depends only on the historical data from$t - 2^{l-1}(k-1)$to t in the input, satisfying the causality constraint.

To further mitigate the gradient vanishing problem in deep networks, TCN introduces a residual block structure with the output:

$$H_{\text{out}} = H^{(l)} + W_{\text{skip}} \cdot H^{(l-1)} \tag{2}$$

where $W_{\text{skip}} \in \mathrm{R}^{d \times m}$ is a jump-connected linear transformation matrix for adjusting the feature dimension. By stacking multiple residual blocks, the TCN finally outputs high-dimensional temporal features $F \in \mathrm{R}^{T \times m}$, where m is the feature embedding dimension.

Compared with traditional recurrent neural networks (RNNs), the parallel convolutional structure of TCNs significantly improves the computational efficiency, while the inflated convolution mechanism effectively captures periodic fluctuations and sudden anomalous patterns in network traffic. For example, a sudden increase in the number of TCP connections in Linux systems is usually accompanied by high activation values in the inflated convolutional layer, while long-term memory leaks are characterised by the low-frequency components of deep residual blocks.

B. Improved anomaly detection algorithm for isolated forests

Traditional Isolation Forest constructs binary trees by randomly dividing the feature space and calculates anomaly scores using the property of shorter paths for anomalous samples, but it faces the following problems in dynamic network environments:

1. static subtrees cannot adapt to data distribution drift;
2. feature importance is not differentiated, leading to insufficient sensitivity of key indicators;
3. the hyperparameters rely on empirical settings, making it difficult to balance efficiency and accuracy.

Aiming at the above problems, this paper proposes the following improvement mechanisms:

In order to adapt to the dynamic changes of Linux network traffic, the incremental training strategy based on sliding window is designed. Define the current time window as $W_{\text{t}} = \{f_{t-\tau+1}, \ldots, f_t\}$ (τ is the size of the window), and the set of isolated forest subtrees as $\{\mathcal{T}_k\}_{k=1}^{K}$. Perform the following operations at every interval of Δt time step:

1. Subtree elimination: weighted random elimination of subtrees by weight $w_k = \exp(-\lambda \cdot age_k)$, where age_k is the survival time of subtree $\mathcal{T}_k$ and λ is the decay coefficient.
2. Subtree generation: randomly sample $\lfloor \rho K \rfloor$ subsets from the current window W_{t},ρ ∈ (0,1) is the elimination ratio), construct a new subtree set $\{\mathcal{T}_{\text{new}}^{(j)}\}_{j=1}^{\lfloor \rho K \rfloor}$.
3. Model update: new subtrees are merged into the original set, keeping the total number of subtrees K constant.

This mechanism enables the model to continuously learn the latest state of the network environment by dynamically updating the set of subtrees, avoiding degradation of detection performance due to data distribution drift.

Aiming at the difference in the contribution of different metrics (e.g., ICMP error rate, process switching frequency) to anomaly detection in Linux systems, a feature importance weighting method based on mutual information is proposed. The feature weight vector $\beta \in \mathrm{R}^m$ is defined and its jth dimension weight is calculated as:

$$\beta_j = \frac{I(F_j; Y) + \epsilon}{\sum_{k=1}^{m} (I(F_k; Y) + \epsilon)} \tag{3}$$

where I(·) is the mutual information, Y is the sequence of historical anomaly labels, and ϵ is a smoothing factor to avoid zero weights. The probability of feature j being selected as a delineation attribute in the subtree construction process of the isolated forest is adjusted to:

$$p_j = \frac{\beta_j}{\sum_{k=1}^{m} \beta_k} \tag{4}$$

This weighting strategy makes the model pay more attention to features that are strongly correlated with abnormal events, e.g., a sudden increase in the ICMP error rate may directly trigger the adjustment of the delineation boundary.

The key hyperparameters of isolated forest (number of subtrees K, sampling ratio ϕ) directly affect the detection efficiency and accuracy. In this paper, Bayesian Optimization is used to achieve adaptive adjustment of hyperparameters. The optimisation objective is defined as the F1-score on the validation set:

$$\max_{K,\phi} \mathrm{E}_{\mathcal{D}_{\text{val}}}[F1(K, \phi)] \quad s.t. \;\; K \in [100,500], \phi \in [0.2,0.8] \tag{5}$$

The objective function is modelled by Gaussian Process (GP) and the next set of candidate parameters are selected based on Expected Improvement (EI) criterion:

$$(K_{\text{next}}, \phi_{\text{next}}) = \arg\max\{\mathrm{E}[\max(F1_{\text{new}} - F1_{\text{best}}, 0)]\} \tag{6}$$

The method quickly approximates the optimal hyperparameter combination within a finite number of iterations, avoiding the computational overhead of grid search.

C. Dynamic Thresholds and Real-Time Early Warning

The feature F_{t} extracted by TCN is directly inputted into the Improved Isolated Forest to calculate the real-time anomaly score s_{t}. To adapt to the non-stationary characteristics of network traffic, Dynamic Quantile Estimation (DQE) is used to adaptively adjust the warning threshold s_{t} :

1. Initial threshold setting: 95 per cent quartile determination of η_0 based on historical data.
2. Update the rules online:

$$\eta_{t+1} = \alpha\eta_t + (1-\alpha)\cdot Q_{0.95}(s_{t-\tau:t}) \quad (7)$$

where $\alpha \in (0,1)$ is the smoothing factor and $Q_{0.95}(.)$ is the 95% quantile within the window τ. The anomaly warning signal $\text{Alarm}_t = 1$ is triggered when $s_t > \eta_t$, otherwise $\text{Alarm}_t = 0$.

IV. Experimental Demonstration

A. Experimental environment and experimental data

The experiments were deployed on a Linux-based system (Ubuntu 20.04 LTS) with hardware configuration of Intel Xeon Gold 6248R CPU (3.0GHz, 48 cores), NVIDIA A100 GPU (40GB video memory) and 128GB RAM. The software environment includes Python 3.8, PyTorch 1.12 (supporting TCN training) and Scikit-learn 1.2 (isolated forest implementation). All experiments were run in Docker containers to ensure environment consistency.

The experiment uses three datasets covering static attacks and dynamic system anomalies:

Public datasets: (1) The KDD Cup 1999 network traffic dataset (with 4 types of anomaly attacks) from UCI and the server performance dataset from Kaggle (CPU, memory, and network traffic timing data collected from AWS EC2 instances) are used. (2) UNSW-NB15 is a large dataset for Network Intrusion Detection System (NIDS) research, created by the Network Security Laboratory at the University of New South Wales, Australia. It contains 175,341 network connection records covering normal traffic and multiple attack types (e.g. DoS, scanning, etc.).

Self-collected data(SC-date): A monitoring tool (Prometheus + Grafana) is deployed on a cluster of local Linux servers (10 nodes) to collect network traffic (TCP/UDP packet rate, number of connections), system metrics and anomaly tags (manual injection of DDoS attacks, hardware failures, etc.) for 30 consecutive days. Comparative Analysis of Experiments.

In this paper, Accuracy, Recall, F1-score and ROC-AUC are used as experimental metrics to comprehensively evaluate the model performance. Accuracy reflects the overall prediction effect, Recall measures the ability to identify positive class samples, F1-score combines precision and robustness, and ROC-AUC evaluates the classification performance of the model under different thresholds. These metrics assess the model performance from multiple dimensions, which is especially suitable for binary classification problems and unbalanced data processing to ensure the scientific validity and reliability of the experimental results.

B. Comparative analysis of experiments

The following are the experimental comparison results based on Accuracy, Recall, F1-score and ROC-AUC to evaluate the performance of Decision Tree, Isolated Forest, and Hyper-parameter Optimised Isolated Forest (I_IF) with the proposed method TCN-I_IF in this paper, respectively:

As shown in Figs. 1, 2, and 3, the TCN-I_IF method proposed in this paper demonstrates significant advantages in Accuracy, Recall, and F1-score comparisons.The accuracy of TCN-I_IF is higher than that of Decision Trees, Isolated Forests, and I_IF, which suggests that it performs the most robustly in the overall classification task. In terms of recall, TCN-I_IF is higher than decision trees, isolated forests, and I_IF, indicating that it is more sensitive to the detection of abnormal samples and can effectively reduce underreporting. Meanwhile, the F1-score of TCN-I_IF is higher than that of Decision Tree, Isolated Forest, and I_IF, which further proves its excellent performance in balancing false alarms and missed alarms. These results indicate that TCN-I_IF is more robust in dealing with complex data distributions and unbalanced samples.

As shown in Table 1, TCN-I_IF also performs outstandingly in the ROC-AUC comparison, with significantly higher values than decision trees, isolated forests, and I_IF.As ROC-AUC is an important measure of the global discriminative ability of a model, the high scores of TCN-I_IF indicate that it has a stronger ability to distinguish between normal and abnormal samples. This result further validates the superiority of TCN-I_IF in dealing with complex data and high-dimensional features, making it an efficient solution in anomaly detection tasks.

TABLE 1. ROC-AUC Comparison

Decision tree	Isolation Forest	I_IF	TCN-I_IF	Decision tree
KDD Cup 1999	0.893	0.922	0.945	0.967
UNSW-NB15	0.89	0.926	0.947	0.968
SC-date	0.896	0.925	0.946	0.968

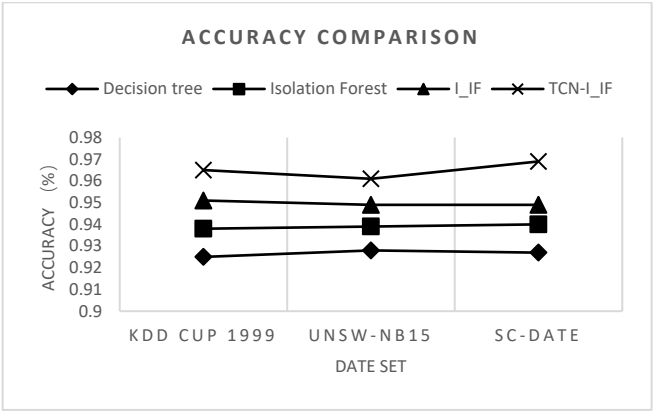

Fig.1. Accuracy Comparison

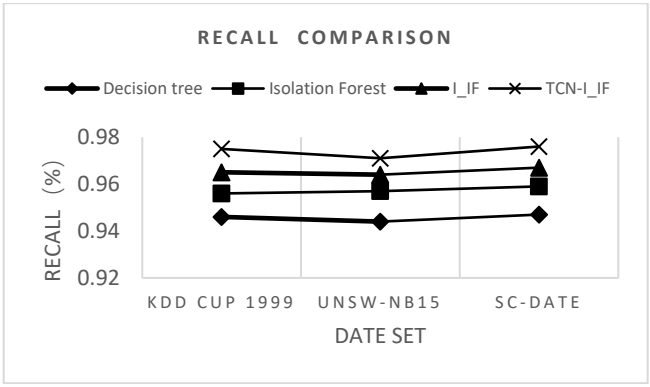

Fig.2. Recall Comparison

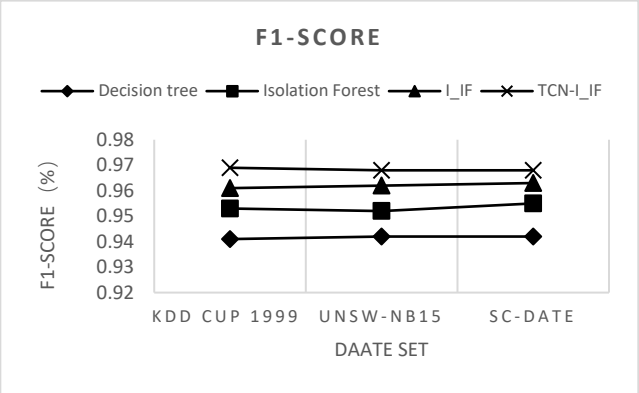

Fig.3. F1-score Comparison

The TCN-I_IF method proposed in this paper significantly outperforms decision trees, isolated forests and hyper-parameter-optimised isolated forests in terms of accuracy, recall, F1-score and ROC-AUC, indicating that it has significant advantages in processing high-dimensional and complex data, reducing omissions and false alarms as well as global discrimination, and is suitable for high-precision anomaly detection scenarios.

V. Conclusion

In this paper, an anomaly identification and prediction method based on TCN and improved isolated forest is proposed. The TCN extracts the long-term dependent features of time series data and combines with isolated forest to achieve efficient anomaly detection, which significantly improves the accuracy and real-time detection. The experimental results show that compared with the traditional isolated forest method, this paper's method significantly improves the accuracy, recall and ROC-AUC, and provides lightweight and adaptive anomaly warning support for the dynamic network environment.

Acknowledgment

This project is supported by Jiangsu Science and Technology Project LYG065212024006.

References

[1] YANG Bin,MA Tinghuai,HUANG Xuejian,et al. Time series anomaly detection based on spatio-temporal feature fusion and sequence reconstruction[J/OL]. Computer Science and Exploration,1-19[2025-03-21].http://kns.cnki.net/kcms/detail/11.5602.tp.20250320.1615.010.html.

[2] YANG Bin,LI Jian,HU Huiwen,et al. Network anomaly detection algorithm based on unsupervised adversarial model[J]. Henan Science,2025,43(03):330-336.

[3] LIU Zhouzhou,JIN Cong,JIANG Guangyi,et al. Deep learning-based traffic anomaly detection in wireless sensor networks[J/OL]. Journal of Jilin University (Engineering Edition),1-8[2025-03-21].
https://doi.org/10.13229/j.cnki.jdxbgxb.20250045.

[4] WEI Jingping,DU Mengdi,WANG Guo. Deep learning based automatic detection method of anomalous data stream intrusion in power communication networks[J]. Automation Application,2025,66(04):247-248+252.DOI:10.19769/j.zdhy.2025.04.068.

[5] H. Zhao,H. Li. Linux log anomaly detection method based on improved isolated forest algorithm[J]. Command Control and Simulation,2024, 46(05):114-118.

[6] WANG Hongyu,LI Xue'an. Research on power marketing anomaly data identification method based on improved isolated forest algorithm[J]. Electrotechnology,2024,(22):29-31.DOI:10.19768/j.cnki.dgjs.2024.22.009.

[7] Li Yuxiao,Hu Jurong,Xing Yanxia,et al. Sea surface small target detection method based on improved isolated forest[J]. Radar Science and Technology,2024,22(06):628-636.

[8] MONG Xiao-Ying,DONG Pei-Xin. Traffic flow anomaly data detection model based on improved isolated forest algorithm[J]. Journal of Chongqing Jiaotong University (Natural Science Edition),2024,43(05) :61-69+90.

[9] Zikai Li, Bo Yang, Zhongtang Zhou, et al. A multi-attention TCN tariff prediction method considering high percentage of wind power fluctuations[J]. Electricity Measurement and Instrumentation,2025, 62(03):138-146.DOI:10.19753/j.issn1001-1390.2025.03.017.

[10] Ki-Li Chen,Navy Li,Xiaolan Xie. A short-term cloud resource prediction model based on time convolution and long-short-term memory network[J]. Science, Technology and Engineering,2025,25(07):2856-2864.

[11] Zhang Lei,Ji Yuanyuan,Li Na,et al. A photovoltaic power prediction model based on TCN with dual attention mechanism[J/OL]. Computer Technology and Development,1-8[2025-03-21].https://doi.org/10.20165 /j.cnki.ISSN1673-629X.2025.0045.

[12] WU Yu,HOU Chuanchuan. Anomaly detection and reconstruction of bridge monitoring data based on LSTM neural network[J/OL]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition),1-7[2025-03-21].http://kns.cnki.net/kcms/detail/42. 1824.U.20250319.1735.018.html.

[13] YK Xu, J Wang, ZX Li, et al. Flight trajectory prediction and anomaly detection method based on multi-source data fusion and deep learning[J/OL]. Journal of Qingdao University (Engineering and Technology Edition),1-9[2025-03-21].http://kns.cnki.net/kcms/detail/37. 1268.TS.20250317.1147.002.html.