Wiley
International Journal of Intelligent Systems
Volume 2024, Article ID 5459452, 18 pages
https://doi.org/10.1155/2024/5459452

WILEY

Research Article

An Efficient Anomaly Detection Method for Industrial Control Systems: Deep Convolutional Autoencoding Transformer Network

Wenli Shang,[1,2] **Jiawei Qiu**,[1,2] **Haotian Shi**,[1,2] **Shuang Wang**,[3] **Lei Ding**,[2,4] **and Yanjun Xiao**[5]

[1]*The School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China*
[2]*The Key Laboratory of On-Chip Communication and Sensor Chip of Guangdong Higher Education Institutes, Guangzhou University, Guangzhou 510006, China*
[3]*The Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China, Tianjin 300300, China*
[4]*The School of Cyber Security, Guangzhou University, Guangzhou 510006, China*
[5]*The Parallel Laboratory, NSFOCUS Technologies Group Co., Ltd., Beijing 100089, China*

Correspondence should be addressed to Shuang Wang; s-wang@cauc.edu.cn and Lei Ding; dloftjcu@163.com

Received 23 October 2023; Revised 8 April 2024; Accepted 7 May 2024; Published 29 May 2024

Academic Editor: Yu-an Tan

Industrial control systems (ICSs), as critical national infrastructures, are increasingly susceptible to sophisticated security threats. To address this challenge, our study introduces the CAE-T, a deep convolutional autoencoding transformer network designed for efficient anomaly detection and real-time fault monitoring in ICS. The CAE-T utilizes unsupervised deep learning, employing a convolutional autoencoder for spatial feature extraction from multidimensional time-series data, and combines this with a transformer architecture to capture long-term temporal dependencies. The design of the model facilitates rapid training and inference, while its dual-component approach, utilizing an optimization function based on support vector data description (SVDD), enhances detection accuracy. This integration synergistically combines spatiotemporal feature extraction, significantly improving the robustness and precision of anomaly detection in ICS environments. The CAE-T model demonstrated notable performance enhancements across three industrial control system datasets. Notably, the CAE-T model achieved approximately a 70.8% increase in $F1$ score and a 9.2% rise in AUC on the WADI dataset. On the SWaT dataset, the model showed improvements of approximately 2.8% in $F1$ score and 5% in AUC. The power system dataset saw more modest gains, with an approximately 0.1% uptick in $F1$ score and a 1% increase in AUC. These improvements validate the CAE-T model's efficacy and robustness in anomaly detection across various scenarios.

1. Introduction

Industrial control systems (ICSs) are pivotal technologies that support national critical infrastructure, extensively utilized in key domains such as military, aerospace, and energy. As ICSs evolve towards increased openness and intelligence, the associated security risks become more prominent [1], particularly given the rise in targeted cyber-attacks in recent years [2–4]. In this context, effective anomaly detection techniques that monitor and identify data deviations in real time to prevent potential malfunctions and attacks are crucial. Current anomaly detection techniques in ICS face several challenges.

1.1. Challenges. Firstly, data in industrial control systems (ICSs) exhibit characteristics that are distinct from those in traditional information systems. These data are often generated in real-time by multiple sensors distributed across the entire system, with each sensor capturing different aspects,

leading to high dimensionality and heterogeneity [5–7]. The high dimensionality of ICS data, meaning that each data point contains a vast array of features, poses challenges to traditional anomaly detection methods [8], which typically struggle with large feature spaces. Furthermore, the heterogeneity of these data indicates that the features originate from various types of sensors and devices, each providing vastly different information [9]. This necessitates that anomaly detection algorithms understand and process these diverse data types. Additionally, the temporal dependency present in ICS data imposes additional demands on anomaly detection algorithms [10]. Abnormal behaviors may manifest not only in the anomalous readings of a single sensor but also in the relationships between readings from multiple sensors [11]. Therefore, it is essential for algorithms to capture and analyze these complex patterns in time-series data.

Moreover, in the ICS environment, the effectiveness of supervised learning methods is limited due to their reliance on large volumes of accurately labeled data [12]. The rarity of abnormal events and the complexity of labeling tasks result in a shortage of anomalous samples in datasets, posing a significant challenge to learning algorithms that rely on labeled data [13]. Furthermore, the infrequency of abnormal events in ICS often leads to class imbalance issues, which further limits the effectiveness of supervised learning methods [14]. While semisupervised learning methods rely less on labeled data, they still encounter challenges in addressing the high dimensionality, heterogeneity, and temporal dependencies of ICS data. These methods frequently struggle to accurately represent the full spectrum of ICS data behaviors, particularly in depicting complex normal operation modes [15].

1.2. Solution Strategy. Given these challenges, unsupervised anomaly detection has emerged as a more suitable approach in the ICS environment. Unlike supervised methods, unsupervised techniques do not rely on labeled data, making them well-suited to the reality of scarce and often unbalanced labeled data in ICS. Unsupervised methods, capable of learning from unlabeled data, are particularly adept at handling the unique complexities of ICS data, which include high dimensionality, heterogeneity, and temporal dependency. Although reconstruction-based techniques such as the convolutional autoencoder (CAE) [16] and denoising autoencoder (DAE) [17] have been employed, they often overgeneralize by fitting both normal and abnormal inputs. Recurrent models such as RNN [18] and LSTM [19] face challenges in capturing long-term trends due to their architecture and tend to have slow run times. Furthermore, two-step methodologies, commonly employed in time-series anomaly detection [20–22], may lead to suboptimal performance when training separate models for different tasks.

1.3. The Novelty of the Proposed Approach. Addressing these limitations, this paper proposes a novel unsupervised learning strategy for anomaly detection in ICS. This strategy aims to develop an effective unsupervised deep learning technique for detecting anomalies in multidimensional time-series data, termed the deep convolutional autoencoder-transformer (CAE-T) network, to address the aforementioned challenges. The CAE-T network comprises two subnetworks: a convolutional representation learning network and a temporal information extraction network. Compared to existing methods, this approach employs a deep convolutional autoencoder as the spatial semantic extraction module, combined with a transformer model with positional encoding for prediction. By utilizing the transformer model, the detection speed is accelerated compared to recurrent methods, as inference can be parallelized on GPUs. Additionally, the transformer model has the added advantage of accurately encoding large sequences, with its training and inference time being largely unaffected by sequence length.

1.4. Contribution

(1) This paper proposes a novel anomaly detection model that leverages spatiotemporal dependencies. The model accounts for the spatial and temporal characteristics of the data, encompassing the relationships between different sensors' readings at a specific time (spatial) and the evolution of these relationships over time (temporal). Specifically, the model uses convolutional neural networks to efficiently extract these spatiotemporal features and integrates them with a transformer-based module to capture the time-varying relationships among multiple sensor data. This approach not only facilitates robust modeling of normal patterns but also enhances the detection of abnormal events in comparison to real-time sensor data.

(2) The paper introduces an enhanced optimization function based on support vector data description (SVDD). This SVDD-based function improves the model's discriminative ability, thereby enhancing anomaly detection performance across multiple datasets.

(3) The paper introduces an integrated loss function that incorporates spatiotemporal information of the data for end-to-end model training. This approach takes into account both the spatial relationships between different sensors' readings and their temporal evolution, leading to enhanced anomaly detection performance.

1.5. Organization. The remainder of this paper is organized as follows: Section 2 provides an overview of related work. Section 3 outlines how the CAE-T model works for multivariate anomaly detection and diagnosis. Section 4 shows the performance evaluation of the proposed method. Section 5 shows the final summary and outlook.

2. Related Work

Anomaly detection techniques in industrial control systems (ICS) encounter unique challenges, primarily arising from the high dimensionality, heterogeneity, and complexity of time-series data. These characteristics render traditional anomaly detection methods less effective in the context of ICS. This paper examines the applicability and limitations of unsupervised anomaly detection techniques in addressing specific challenges inherent in the ICS environment. This section initially introduces traditional anomaly detection methods and then discusses the advantages and limitations of deep learning-based anomaly detection approaches in tackling these challenges.

2.1. Traditional Anomaly Detection. In ICS datasets, traditional anomaly detection methods encounter several challenges. These methods, typically designed for static and single data sources, struggle with the high dimensionality and heterogeneous nature of multisensor ICS data. For example, reconstruction-based methods such as autoencoders, while capable of constructing normal models from high-dimensional data, may incur increased reconstruction errors when applied to complex ICS data. Clustering analysis techniques, including GMM [23], k-means, and KDE [24], can handle heterogeneous data but might struggle to capture the temporal dynamics in the constantly changing ICS environment. Learning-based methods, such as SVM [25] and SVDD [26], face challenges in accurately distinguishing between normal and abnormal data in complex datasets.

For time-series data, models, such as AR and ARIMA [27], widely used in various domains, have limited applicability in ICS. These models face challenges such as high computational costs and limited capability in handling long-term trends, especially in multisensor multivariate time series data.

2.2. Deep Learning-Based Anomaly Detection. The discussion of reconstruction models, predictive models, and combinatorial models in deep learning-based anomaly detection is crucial for understanding their suitability in addressing the unique challenges of ICS.

2.2.1. Reconstruction Models. Reconstruction models strive to minimize reconstruction errors using various approaches. Autoencoders, widely used in anomaly detection, are designed to reconstruct given inputs. Trained exclusively on normal data, these models identify anomalies through significant reconstruction errors. LSTM autoencoder models detect anomalies by learning temporal features of input time series and utilizing reconstruction errors. Although effective for time-series data, these models' performance in detecting anomalies in multidimensional time series diminishes due to a lack of spatial correlation consideration. CAE [22] captures two-dimensional (2D) image structures and anomalies by transforming data into an image format. Audibert et al. [28] employed adversarial training between two networks for effective anomaly detection in time-series data. DAGMM [29] integrates deep autoencoders with Gaussian mixture models, using unsupervised learning to acquire low-dimensional representations of input data and identify anomalies by comparing new samples with the learned distribution.

2.2.2. Predictive Models. The predictive model is a widely used tool in anomaly detection. Its basic idea is to predict future output values and to compare the predicted values with the actual values. If the predicted value deviates significantly from the actual value, the data point is considered as an anomaly. Commonly used prediction models include a variety of architectures: RNNs and LSTMs, which excel at capturing temporal dependencies, and CNNs, adept at processing spatial structures in data such as images or videos. Recently, transformer-based models have been employed for their ability to capture long-range dependencies in data. For instance, Erba et al. [30] utilized LSTM to predict values for subsequent time periods and detect anomalies by minimizing the mean square error between predicted and actual values. Li et al. [31] recently employed transformer models to capture temporal dependencies and patterns within the data. By integrating a transformer encoder-decoder architecture with a prediction branch, this approach reconstructs and predicts future values of the time series. Anomalies are detected by comparing the reconstruction and prediction errors, thus identifying deviations from normal behavior. Additionally, there are anomaly detection models based on generative adversarial networks (MAD-GAN) [32], which use the predictive power of GAN generators to detect anomalies and distinguish between false and real data. The continued development of these models is expected to play a significant role in the future of anomaly detection.

2.2.3. Combinatorial Models. In addition to single models, combinatorial models are increasingly gaining attention in the field of unsupervised anomaly detection. Su et al. [29] utilize a deep autoencoder architecture in combination with an RNN to capture temporal dependencies and complex patterns in data. They employ a reconstruction-based approach in which the model is trained to accurately reconstruct normal data. Anomalies are identified based on reconstruction errors, with larger errors indicating deviations from normal behavior. This algorithm provides a versatile and effective solution for detecting anomalies in diverse and dynamic time series data, making it suitable for a range of anomaly detection applications. Conversely, Ullah et al. [33] proposed a composite model employing a single encoder LSTM and multiple decoder LSTMs to perform

tasks like reconstructing input sequences and predicting future sequences. In certain studies, the ConvLSTM model has been utilized within the composite LSTM model as an intermediary unit between the reconstruction and prediction branches. Currently, this composite model is used for feature extraction from video data for action recognition tasks. Additionally, other researchers [22] have proposed a convolutional neural network-based model for multivariate deep spatiotemporal anomaly detection (CAE-M). This model employs a convolutional architecture to capture spatial-temporal variations, featuring a network structure with an encoder and two decoder branches for reconstructing past sequences and predicting future sequences, respectively.

In this study, we introduce a deep learning combinatorial model that synergistically combines reconstruction and prediction strategies. Utilizing transformers to capture temporal dependencies and CNNs for spatial feature extraction, our model offers a comprehensive approach to anomaly detection. A key feature of our model is the integration of a unique optimization function, significantly enhancing the precision and robustness of anomaly detection. Comparative evaluations indicate that our methodology surpasses existing combinatorial models, particularly in capturing intricate spatiotemporal patterns, accommodating high-dimensional data, and enhancing the accuracy and resilience of anomaly detection.

3. Method

3.1. Notation. In this paper, we consider multisensor time series data, which contains a multisensor time series $X = \{x_1, x_2, \ldots, x_n\}$. As shown in Figure 1, we now define the two problems in anomaly detection and diagnosis. (a) Unsupervised anomaly detection and diagnosis in multivariate time series data. (b) Discerning distinctive system signature matrices between normal and abnormal states.

(a) In unsupervised anomaly detection and diagnosis, normal data is used for training. During testing, we compute an anomaly score for each data window. By comparing these scores against a predetermined threshold value (T), we are able to identify windows exhibiting anomalies, thereby signifying their presence within the local time range.

(b) We convert system signature data from the supervisory control and data acquisition (SCADA) system into an $N \times L$ matrix $\mathbf{W}$ with N representing the number of sensors and L representing the sliding window length. This approach, referenced in prior work [22], effectively retains local temporal information of the multidimensional sensors. Through this transformation, we convert the original data into a fixed-length sliding window sequence for subsequent processing.

3.2. Overview. The CAE-T framework shown in Figure 2 comprises two key components: the convolutional representation learning network (CRLN) and the temporal information extraction network (TIEN), playing a central role in capturing spatiotemporal dependencies. The CRLN utilizes an advanced deep convolutional autoencoder specifically designed to transform complex multidimensional temporal data into a compact, low-dimensional representation. This transformation process not only simplifies the data but also focuses on preserving and emphasizing the rich semantic and temporal information in the data. The complexity of the CRLN lies in its ability to recognize and encode subtle variations in temporal data. The temporal information extraction network (TIEN), as another key component, aims to utilize the convolutional representation provided by the CRLN. TIEN enhances the framework's understanding and analysis of the dynamics of time series by combining the temporal features encoded by the CRLN with additional contextual data in a complex manner. The design focus of TIEN is to capture subtle variations in time and to integrate them with the output of the CRLN to comprehensively interpret spatial and temporal patterns. Subsequently, TIEN introduces an innovative dual-input mechanism that processes both the convolutional representation from the CRLN and the original data simultaneously. This mechanism significantly enhances the model's efficiency and accuracy in capturing spatiotemporal information, allowing TIEN to more deeply understand and analyze the dynamics of time series. Specifically, this dual-input mechanism enables TIEN to utilize not only the advanced temporal features provided by the CRLN but also to directly process the raw data, offering a more comprehensive data perspective. Through this design, the CAE-T framework more effectively identifies and processes complex and subtle spatiotemporal dependencies in multidimensional temporal data.

These two components together form a synergistic mechanism of the model. The advanced encoding capabilities of the CRLN combined with the dynamic temporal analysis of TIEN enable the CAE-T framework to capture, interpret, and understand the complex spatiotemporal dependencies in multisensor time series signals. This enhanced capability ensures that the model not only recognizes patterns and anomalies with higher accuracy but also gains deep insights into the latent structure of the temporal data. The innovation of CAE-T is significant in two aspects: first, the integration of the penalty term from the SVDD algorithm. This measure aims to prevent the autoencoder from overfitting and to avoid similar treatment of normal and anomalous data, thereby improving anomaly detection performance. Secondly, before the data enters TIEN, the convolutional representation learned by the encoder undergoes a linear transformation, inspired by the transformer encoder, decoder, and position encoder, aimed at enhancing the extraction of temporal information. Our approach enhances the anomaly detection of multidimensional temporal data and accelerates the training process for extracting temporal information, thanks to its synergistic effect with the convolutional autoencoder. We adopted a composite model to capture the spatiotemporal dependencies in multisensor time series signals. Furthermore, to simplify the end-to-end training process, we introduced a weighted comprehensive objective function. During the inference

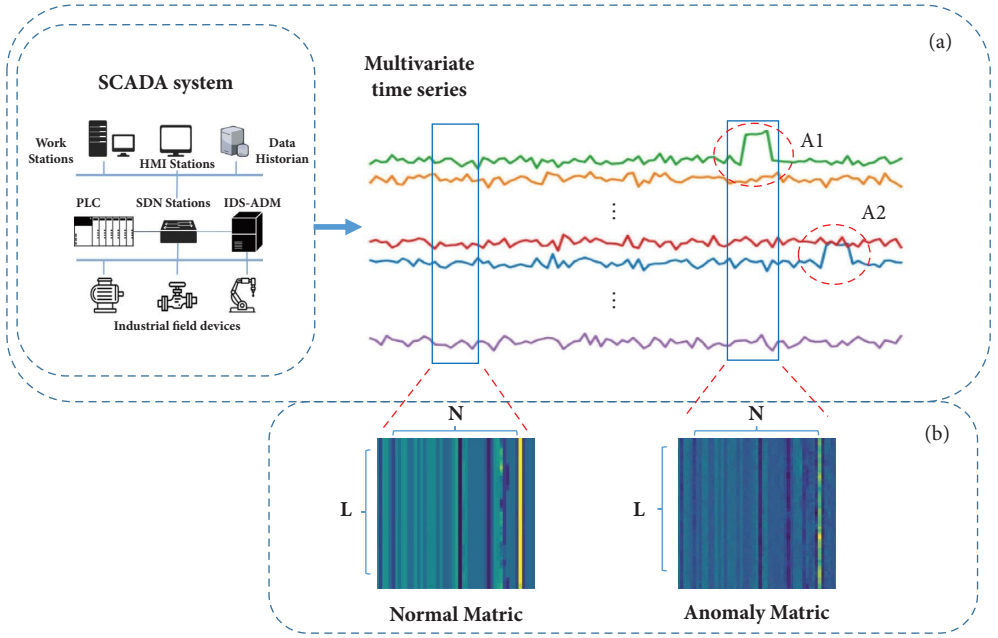


FIGURE 1: (a) Unsupervised anomaly detection and diagnosis in multivariate time series data, which illustrates two anomalies, i.e., A1 and A2 marked a by red dash circle, in multivariate time series data. (b) Different system signature matrices between normal and abnormal periods.

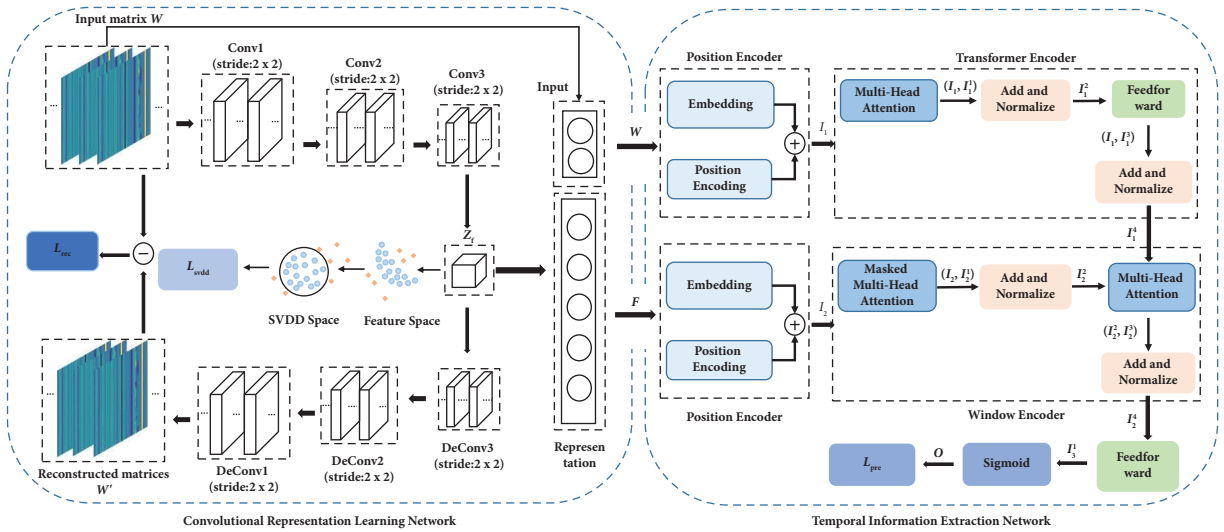


FIGURE 2: CAE-T, the architecture of the proposed real-time anomaly detection framework.

process, the model computes the loss function within the composite framework to accurately detect anomalies, effectively addressing multidimensional temporal data's anomaly detection challenges.

3.3. Convolutional Representation Learning Network. In the convolutional representation learning network, a fusion mechanism is used to integrate multiple variable signals. The process comprises two key components: (1) feature extraction from multivariate signals using a convolutional autoencoder and (2) calculating the reconstruction error, typically using mean squared error (MSE). Reconstruction loss quantifies the similarity between the reconstructed and original inputs, thereby providing a measure of the reconstruction's fidelity. To mitigate the risk of overgeneralization by the convolutional autoencoder when processing anomalous inputs, this paper proposes a specialized loss function. This method effectively captures the potential spatial semantic representation of the input signal, while reducing the likelihood of the autoencoder overfitting to anomalous data.

3.3.1. Convolutional Autoencoder Networks. Autoencoders, a type of artificial neural network, are designed to learn efficient encodings of input data, typically used for data dimension reduction. Specifically, an

autoencoder is trained to replicate its input at its output. The network can be seen as comprising two parts: an encoder function that converts input data into a compressed representation and a decoder function that reverts this representation to its original form. The central concept is to learn a representation (encoding) for the data, aimed at dimensionality reduction, which is more compact than the original data yet retains sufficient information for minimal-loss reconstruction of the original data.

In our work, a deep convolutional autoencoder is utilized to extract low-dimensional features from the input matrix $W \in \mathbb{R}^{N\times L}$, where N represents the N-dimensional time series input and L the time window length.

As depicted in (1), the encoder maps the input matrix W to a hidden representation Z_f by applying multiple convolutional and pooling layers. Each convolutional layer is followed by a maximum pooling layer, which reduces the dimensions of the corresponding layers. The maximum pooling layer selects the highest value within each region of the feature map, thus generating an output feature map with reduced dimensions, determined by the size of the pooling kernel. Conversely, the decoder, as shown in (2), performs reverse mapping, transforming the hidden representation Z_{f} back into the original input space, thus resulting in a reconstruction. This process involves expanding the compact representation into a wider reconstruction matrix using transposed convolutional layers, which increase the layers' width and height, functioning similarly to convolutional layers but in reverse. The discrepancy between the original input vector W and the reconstructed vector, termed the reconstruction loss L_{rec}, is typically quantified using the mean square error (MSE), as depicted in (3). The MSE measures the distance between the reconstructed data $\mathbf{W}'$ and the original input $\mathbf{W}$, thus indicating the reconstruction's effectiveness.

$$Z_{\mathrm{f}} = \mathrm{Encode}(W), \tag{1}$$

$$W' = \mathrm{Decode}(Z_{\mathrm{f}}), \tag{2}$$

$$L_{\mathrm{rec}} = \left\|W - W'\right\|_2^2. \tag{3}$$

3.3.2. Deep SVDD for Anomaly Detection. Distinguishing between normal and abnormal data during anomaly detection with autoencoders can be challenging. This challenge arises due to the autoencoder's tendency to overfit, noise in the training data, and redundancy in large-scale datasets. Such issues hinder the autoencoder's feature extraction capabilities, thereby compromising the model's robustness. This paper discusses the deep SVDD method, which enhances the optimization function by adding a penalty term to the original loss. The training process of deep SVDD consists of two stages. The first stage involves pretraining the autoencoder to initialize network parameters and to learn implicit data features. In the second stage, the network is trained using an objective function tailored for anomaly detection, which aims to maximize differences between data representations. The anomaly detection objective function is presented in (4).

$$\min_{R,w} R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\max\left\{0, \left\|\phi(x_i; w) - c\right\|^2 - R^2\right\} + \frac{\lambda}{2}\sum_{\iota=1}^{l}\left\|w^{\iota}\right\|_F^2 \tag{4}$$

where ϕ denotes a neural network with l hidden layers and the weight parameter $w = \{w^1, w^2, w^3, \ldots, w^{\iota}\}$, w^{ι} represents the l^{th} layer network parameter, $\iota \in \{1, 2, \ldots, l\}$, x_i denotes the i^{th} training data, n represents the total number of training samples, $R > 0$ denotes the radius of the hypersphere, and $\nu \in (0, 1]$ is a hyperparameter. λ is a regularization hyperparameter controlling the trade-off between the network's complexity and its performance on the training data. A higher value of λ penalizes more complex models, aiding in the prevention of overfitting by maintaining smaller weights. This objective function's hallmark is its ability to identify the smallest hypersphere that encapsulates the majority of normal data representations while excluding most anomalous data points. Minimizing the hypersphere's volume corresponds to minimizing the objective function. A penalty is assigned to data points falling outside the hypersphere and activated only when a data point's distance from the centroid exceeds the radius R. This design enhances the network's sensitivity to the boundaries between normal and anomalous data, thereby improving anomaly detection. In this study, a convolutional autoencoder is used to map the feature space to the SVDD space, with the representation's position relative to the hypersphere incorporated as part of the loss function, as detailed in (5).

$$\mathrm{L}_{\mathrm{svdd}} = R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\max\left\{0, \left\|\mathrm{Z}_{\mathrm{f}} - c\right\|_2^2 - R\right\}. \tag{5}$$

3.4. Temporal Information Extraction Network. Based on existing research as demonstrated in reference [22], our temporal information extraction network (TIEN) introduces its key innovation—a dual-input system. This system combines convolutional representation and direct raw data input, a unique approach that endows TIEN with the ability to utilize both the deep semantics and temporal insights of multidimensional temporal data extracted by CRLN's convolutional representation and to process raw data for immediate data insights. By integrating dual inputs, the system's depth of data understanding is significantly enhanced, enabling it to manage spatiotemporal information and detect subtle anomalies in raw data, in addition to handling advanced abstract features.

This blended input technique stands out by integrating deep learning's robust feature extraction capabilities with acute perception of raw data. Convolutional representation, for instance, captures extensive dependencies and complex patterns in time series data, while raw data input is crucial for identifying rapid or minor changes within specific time windows. Through this, TIEN achieves a more comprehensive data analysis, increasing the precision and speed of anomaly detection in multidimensional time data.

TIEN also adopts distinct strategies for the two input types. It uses high-level semantic information from convolutional representation for broad anomaly pattern analysis, while concentrating on local features and immediate anomaly detection when processing raw data. This balanced approach renders TIEN more flexible and effective in navigating complex spatiotemporal data.

In the subsequent sections, we will delve deeper into how TIEN's innovative dual-input mechanism is powered by cutting-edge deep learning technologies, showcasing its practical benefits and strengths in real-world time series anomaly detection scenarios.

The transformer, a deep learning model, is extensively used in tasks such as natural language processing and visual processing. Inspired by the literature [31], the transformer has been restructured to be applicable for temporal anomaly detection tasks. To achieve this, the transformer's structure was adapted and optimized accordingly. Initially, a convolutional layer was introduced as a preprocessing step for windowed data to capture local features of time series data. Subsequently, in the transformer's encoder section, improved modules, including position encoding, were added to better cater to time series data. This integration of position encoding and the transformer model enables the effective capturing of dependencies and key features within time series data. Additionally, the inclusion of the temporal anomaly score calculation as part of the loss function optimizes the model's performance in temporal anomaly detection tasks. Figure 2 illustrates the extraction of temporal information and the calculation of the temporal anomaly score.

The following section details how the transformer functions. First, the scaled dot product attention involving three matrices Q (query), K (key), and V (value) is defined.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{m}}\right)V. \tag{6}$$

In this context, the softmax function forms a convex combination of weights for the median of V, enabling the compression of the matrix V into smaller representative embedding. This simplifies the inference of downstream neural network operations. Unlike traditional attention operations, scaled dot product attention scales the weights by $\sqrt{m}$ to reduce weight variance, thereby stabilizing training. For the input matrices Q, K, and V, Q_i, K_i, and V_i are obtained by initially passing them through h feedforward layers, where $i \in \{1, 2, \ldots, h\}$. The scaled dot product attention is then applied as follows for the multiheaded self-attention calculation:

$$\text{MultiHeadAtt}(Q, K, V) = \text{Concat}(H_1, \ldots, H_h), \tag{7}$$

$$H_i = \text{Attention}(Q_i, K_i, V_i). \tag{8}$$

In our time series anomaly detection model, the multihead attention mechanism enables the model to simultaneously focus on various patterns and trends from different representational subspaces at different positions. This is crucial for identifying anomalous patterns in time series data, as anomalies may manifest as unusual relationships between multiple sensor readings or abnormal behaviors within specific time windows. Additionally, the position encoding plays a central role in our transformer model, providing the model with positional information about each time point in the sequence. Given the nature of time series data, the temporal positional information of each data point is particularly important for understanding the dynamic changes of the entire sequence. Position encoding integrates the positional information of each time point into its corresponding feature representation, enabling the model to capture temporal dependencies and long-term relationships, which is crucial for accurately identifying anomalous patterns in time series. The implementation formula for position encoding is as follows:

$$\text{PE}_{(\text{pos},2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right), \tag{9}$$

$$\text{PE}_{(\text{pos},2i+1)} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right). \tag{10}$$

In this context, "pos" represents the position index in the time series, "d_{model}" represents the data dimensions, and "2i" and "2i + 1" denote different dimensions. The model utilizes sine and cosine waveforms to differentiate between various time points. Our model comprises a transformer encoder, a window encoder, and a position encoder, as illustrated in Figure 2. The model's inference process is divided into two stages. First, to scale the input and balance the effects between different dimensions, the window data and feature representation are paired as input to derive a focus score F. The focus score F is broadcast to match the dimensions of W, followed by appropriate zero padding and splicing. The data is then position-encoded and used as input to the encoder denoted as I_1. The transformer encoder then performs the following operations:

$$I_1^1 = \text{MultiHeadAtt}(I_1, I_1, I_1) \tag{11}$$

$$I_1^2 = \text{LayerNorm}\left(I_1 + I_1^1\right), \tag{12}$$

$$I_1^3 = \text{FeedFword}\left(I_1^2\right), \tag{13}$$

$$I_1^4 = \text{LayerNorm}\left(I_1 + I_1^3\right). \tag{14}$$

Here, MultiHeadAtt (I_1, I_1, I_1) denotes the multiheaded self-attention operation of the input matrix I_1, and "+" denotes matrix addition. These operations utilize the input

time series window and the full series to generate attention weights, capturing the temporal trend in the input series. These operations enable the model to infer multiple time series windows in parallel, as the neural network does not rely on the output of previous timestamps at each timestamp, significantly improving the training time of the proposed method. In the window encoder, position encoding is applied to the input window W to obtain I_2. The self-attention in the window encoder is modified to mask data from subsequent positions. This modification prevents the decoder from accessing data points for future timestamps during training, as all data **W** and **F** are provided simultaneously for parallel training. The window encoder then performs the following operations:

$$I_2^1 = \text{Mask}\left(\text{MultiHeadAtt}\left(I_2, I_2, I_2\right)\right), \tag{15}$$

$$I_2^2 = \text{LayerNorm}\left(I_2 + I_2^1\right), \tag{16}$$

$$I_2^3 = \text{MultiHeadAtt}\left(I_1^4, I_1^4, I_2^2\right), \tag{17}$$

$$I_2^4 = \text{LayerNorm}\left(I_2^2 + I_2^3\right). \tag{18}$$

When the window encoder performs the attention operation, the complete sequence I_1^4 is used as the value and key, while the encoded input window serves as the query matrix. Window input masking is applied here to hide the window sequence for future timestamps within the same input batch. As the model receives the complete input sequence up to the t timestamp, it can encapsulate and exploit a larger context, as opposed to the limited context in the previous literature. Finally, the same decoder as the encoder is used, with Sigmoid activation to generate outputs in the range [0, 1], matching the normalized input window W. Thus, the temporal information extraction network utilizes inputs F and W to generate the output O.

$$I_3^1 = \text{FeedForward}\left(I_2^3\right), \tag{19}$$

$$O = \text{Sigmoid}\left(I_3^1\right), \tag{20}$$

$$L_{\text{pre}} = \|W - O\|_2^2. \tag{21}$$

3.5. Loss Function. As the models are trained separately, they are susceptible to becoming trapped in local optimum solutions. Therefore, an end-to-end hybrid model is proposed, which was achieved by minimizing the composite objective function. The CAE-T loss function comprises three components: a mean square error (reconstruction error) term, a maximum mean difference (regularization) term, and a prediction error term for nonlinear prediction tasks. The loss function is constructed in the following manner:

$$loss = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{svdd}} + \lambda_3 L_{\text{pre}}. \tag{22}$$

The parameters λ_1, λ_2, and λ_3 govern various aspects of the model, including the influence of the transformer anomaly score (λ_1), the weighting of the support vector data description (SVDD) enhancement λ_2), and the weighting of temporal anomaly significance (λ_3). The mean squared error term is primarily used to evaluate the model's ability to reconstruct input data, which is a key factor in anomaly detection. The SVDD term, by providing an additional regularization constraint, helps the model better differentiate between normal and abnormal data, preventing overfitting. Meanwhile, the prediction error term focuses on performance in nonlinear prediction tasks, enabling the model to effectively identify anomalous patterns in time series. The collaborative work of these three components allows the CAE-T model to handle anomaly detection tasks in multidimensional time data more comprehensively and effectively.

3.6. Inference. During the inference phase, the trained model is utilized to calculate the anomaly score and distinguish between normal and anomalous samples by setting a threshold. The anomaly score calculation, as represented by (17), comprises three elements: the reconstruction error, the distance of the representation vector Z_f, and the prediction error. By combining these errors, more abnormal judgment factors are incorporated, thus enhancing the ability to distinguish between normal and abnormal samples.

For threshold calculation, the POT (peak over threshold) method [34], an application of the extreme value theory for estimating extreme value distributions above a certain threshold, was used. Initially, an appropriate threshold is chosen, followed by the extraction of extreme observations above that threshold and the modeling of these extreme events. By estimating the parameters of the extreme value distribution, an estimate of the threshold is obtained.

This integrated approach is crucial in anomaly detection as it enables the full utilization of multiple error metrics, yielding more accurate anomaly scores. Concurrently, the use of the POT method renders the determination of the threshold value more scientific and reliable. Through this well-designed abnormality detection mechanism, the differences between normal and abnormal samples are identified more effectively, playing a crucial role in practical applications.

$$\text{Anomaly score} = L_{\text{rec}} + L_{\text{svdd}} + L_{\text{pre}}. \tag{23}$$

4. Experiment

In this section, we demonstrate the CAE-T model for anomaly detection experiments on a real industrial control dataset to evaluate the performance of the model for anomaly detection in industrial control systems, and compare it with other popular anomaly detection algorithms

to demonstrate the superiority of the model. In addition, to verify the effectiveness of the proposed improvement method, ablation experiments are conducted in this paper to compare and analyze the performance of the model in anomaly detection before and after the improvement, respectively. The results of these experiments can further validate the feasibility and effectiveness of our proposed CAE-T model and its improved method in practical applications.

4.1. Dataset. Experiments are carried out on three datasets, each derived from real industrial control scenarios. The SWaT (https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/#swat) dataset [35] uses real-time data from a modern water treatment plant, undergoing various attacks throughout the water treatment process, including types like single-point single-stage and multipoint single-stage types. The WADI (https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/#wadi) dataset [36] captures data from 123 sensors and actuators, showcasing the different states of the water treatment and distribution. Power System) (https://www.ece.uah.edu/%7Ethm0009/icsdatasets/binaryAllNaturalPlusNormalVsAttacks.7z) dataset [37] provides operational data from power plants, aiding research in energy systems.

This paper evaluates CAE-T by implementing anomaly detection on the above datasets using the CAE-T model and juxtaposing it with other advanced algorithms. In this anomaly detection implementation, there's a marked disparity in the ratio of normal to abnormal samples, with fewer abnormal samples making the data more unbalanced. Training uses normal samples, while testing uses a blend of untrained normal and abnormal samples. The dataset details during training and testing are in Table 1.

4.2. Comparison Algorithm. To extensively evaluate the performance of the proposed CAE-T method, we compare it with several deep anomaly detection methods. Table 2 shows some descriptions of the compared algorithms:

4.3. Implementation Details. In this study, we introduce the CAE-T model, characterized by an encoder comprising three convolutional pooling layers and a decoder formed by corresponding inverse convolutional pooling layers. The detailed parameter settings are meticulously documented in Table 3. Additionally, we explored a variety of deep learning models for comparative analysis:

(i) MAD_GAN: based on generative adversarial networks, this model features a three-layer fully connected network, utilizing LeakyReLU and Sigmoid activation functions.

(ii) MSCRED: integrates CNN and ConvLSTM layers, aiming to capture the spatiotemporal dependencies in time series data.

(iii) USAD: uses a dual autoencoder structure for anomaly detection, using ReLU activation functions between layers.

(iv) OmniAnomaly: merges variational autoencoders with GRU, focusing on complex time series modeling.

(v) LSTM_AD: utilizes a two-layer LSTM network structure to capture temporal dependencies, outputting anomaly scores via a Sigmoid function.

(vi) TranAD: implements a Transformer architecture for time series anomaly detection, including positional encoding and transformer encoder-decoder.

(vii) CAE_M: A convolutional autoencoder designed for multivariate time series, characterized by feature extraction and data reconstruction using Sigmoid functions.

(viii) DAGMM: combines autoencoders with Gaussian Mixture Models for high-dimensional data anomaly detection, adopting Tanh and Sigmoid activation functions.

During the training phase, all models used the Adam optimizer with an initial learning rate of 0.01 and a meta learning rate of 0.02, adjusted through a step scheduler with a step size of 0.5. Common hyperparameter settings included: training epochs of 20, window size of 5 (adjusted according to the dataset dimensions for CAE_M, TranAD, and MSCRED), one layer of transformer encoders, two layers of feed-forward units in encoders, 64 hidden units, and a dropout rate of 0.1. Given the class imbalance in the datasets, as detailed in Table 1, precision, recall, and $F1$ scores were adopted as evaluation metrics. The data was partitioned into training and testing sets at a $6:4$ ratio, consistent with the methodologies in existing research [31], ensuring the training set contained only normal samples and did not overlap with the test set containing abnormal samples. This separation is crucial for unbiased hyperparameter tuning and the calculation of anomaly detection thresholds.

4.4. Evaluation Metrics. Before delving into the results, it is crucial to comprehend the evaluation metrics used in this study. The performance of our proposed method and other deep learning-based anomaly detection methods is assessed using the following metrics:

Precision: precision measures the proportion of correctly predicted positive observations out of the total predicted positives. Mathematically, it is represented as

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{24}$$

Recall: recall calculates the proportion of actual positives that are correctly identified. It is given by

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{25}$$

Table 1: The datasets in this paper during training and testing.

Dataset	Train	Test	Anomalies (%)	Dimensions	Permissions
SWaT	496800	449919	11.98	51	Public
WADI	1048571	172801	5.99	123	Public
Power system	22706	55662	71	129	Public

$F1$ Score: the $F1$ Score is the harmonic mean of precision and recall, providing a balance between the two. It is computed as

$$F1 = 2 \times \frac{P \times R}{P + R}. \tag{26}$$

AUC (area under the curve): For anomaly detection, the roc_auc_score function from the scikit-learn (sklearn) library was utilized to compute the AUC. This metric acts as an indicator of the model's ability to distinguish anomalies from normal instances. To calculate the AUC, true labels and predicted anomaly scores are provided as inputs to the roc_auc_score function. The function then generates the ROC curve and calculates the AUC by numerically integrating the area beneath it. A higher AUC value indicates superior anomaly detection performance, demonstrating better discrimination between anomalies and normal data points.

4.5. Results and Analysis. This section presents an exhaustive evaluation of the CAE-T model's performance across three distinct datasets, as detailed in Table 4. Utilizing key metrics, including precision (P), recall (R), $F1$ score, and the area under the receiver operating characteristic curve (AUC), the analysis provides critical insights:

On the SWaT dataset, the CAE-T model outperforms eight other evaluated methodologies, achieving a precision of 0.9764 and a recall of 0.9923. Conversely, the USAD model exhibits the lowest AUC score at 0.9472, indicating significant potential for improvement in its anomaly detection approach. Reconstruction-based methodologies, including DAGMM, MSCRED, and USAD, are notably affected by noisy data, which could lead to erroneous identification of anomalous inputs as normal. Time-series predictive methods, such as LSTM-AD and OmniAnomaly, show limited performance due to inadequate accounting for the data's spatial attributes. Integrated approaches like TranAD, MSCRED, and CAE-M experience suboptimal convergence due to their phased training, resulting in moderate outcomes.

The complexity of the WADI dataset presents significant challenges to most models, as reflected in generally lower $F1$ scores. However, CAE-M ($F1$: 0.4119) and OmniAnomaly ($F1$: 0.4260) demonstrate superior performance, attributable to their effective dimensionality reduction and analysis of feature interrelationships. Conversely, models like TranAD, which overlook interfeature correlations, show subpar efficacy.

In the context of the power system dataset, characterized by a higher proportion of anomalies and simpler, less noisy patterns, and most algorithms perform admirably. Methods underpinned by convolutional neural networks, including CAE-M, MSCRED, and OmniAnomaly, demonstrate considerable efficacy. However, predictive models, namely, TranAD ($F1$, 0.4524) and LSTM-AD ($F1$, 0.6584), display a deficiency in capturing spatial interrelations within multidimensional data.

These observations underscore that although the CAE-T model shows promising results, there is still room for improvement, especially when dealing with highly complex and noisy datasets. Future research efforts will focus on enhancing the model's robustness and adaptability in diverse industrial control system settings, particularly in refining data preprocessing procedures and strengthening the model's ability to understand intricate data features.

4.6. Effectiveness Evaluation. Expanding on the insights from Section 4.5, this section further explores the CAE-T model's effectiveness through an in-depth evaluation. Initially, an ablation study (Section 4.6.1) is performed to analyze the impact of individual components within the CAE-T model. This assessment is crucial for understanding each element's contribution to the model's overall efficacy, as demonstrated by the results in Section 4.5. Following this, a training set sensitivity analysis (Section 4.6.2) and a parameter sensitivity analysis (Section 4.6.3) are conducted. These analyses further highlight the model's adaptability and robustness in various operational scenarios.

4.6.1. Ablation Study Design. In this study, several variants of the CAE-T model were systematically assessed, each incorporating specific technological enhancements. We evaluated three primary versions: CAE_$T_{(\text{Rec+Svdd})}$, which exclusively integrated the support vector data description (SVDD) as an enhancement; CAE_$T_{(\text{Rec+Pre})}$, solely amalgamating the transformer-based anomaly score; and the comprehensive CAE-T model, which combined both SVDD and the transformer-based anomaly score enhancements. Detailed experimental results are presented in Table 5 and Figures 3 and 4.

In our ablation study, different variants of the CAE-T model exhibited significant performance variations across three datasets, owing to complex technical reasons. The model's underlying architecture was meticulously optimized, particularly for high-dimensional and time-series data anomaly detection, enabling the CAE-T to sharply differentiate between normal and anomalous patterns. For instance, on the SWaT dataset, the CAE-T model achieved a precision of 0.9260 and an AUC of 0.9483. The incorporation of SVDD in CAE_$T_{(\text{Rec+Svdd})}$ led to marked improvements in $F1$ score and AUC, reaching 0.9805 and 0.9680, respectively, highlighting SVDD's critical role in

Table 2: The survey of the compared algorithms.

Research work	Key technique	Categories	Evaluation dataset	Baseline methods
TranAD [31]	Deep transformer networks for anomaly detection	Combinatorial model	SWaT, WADI, SMD, SMAP, MSL	(1) Transformer (2) AE (3) SPOT
DAGMM [20]	Deep autoencoding Gaussian mixture model	Reconstruction model	SMD, SMAP, MSL	(1) AE (2) GMM
OmniAnomaly [29]	Multitask learning and adaptive thresholding	Predictive model	SMD, SMAP, MSL	(1) GAN (2) AE (3) LSTM
USAD [28]	Variational autoencoder	Reconstruction model	SWaT, WADI, SMD, SMAP, MSL	(1) VAE (2) GAN
MSCRED [21]	Attention-based ConvLSTM	Combinatorial model	Synthetic data, power plant data	(1) CAE (2) LSTM (3) Attention
CAE-M [22]	Convolutional autoencoding memory network	Combinatorial model	PAMAP2, CAP, mental fatigue dataset	(1) CAE (2) LSTM (3) Attention
MAD_GAN [32]	Multivariate anomaly detection based on generative adversarial networks	Reconstruction model	SWaT, WADI	(1) GAN (2) LSTM (3) AE
LSTM-AD [30]	LSTM autoencoders	Predictive model	WADI, BATADAL	(1) AE (2) LSTM

Table 3: The parameters set during the training process.

Parameters	Values
λ_1	0.1
λ_2	$1e-5$
λ_3	$5e-5$
Window size	128
Batch size	256
Epochs	20
Learning rate	0.0001

Table 4: The precision, recall, and $F1$ score and AUC score of baselines and our proposed method.

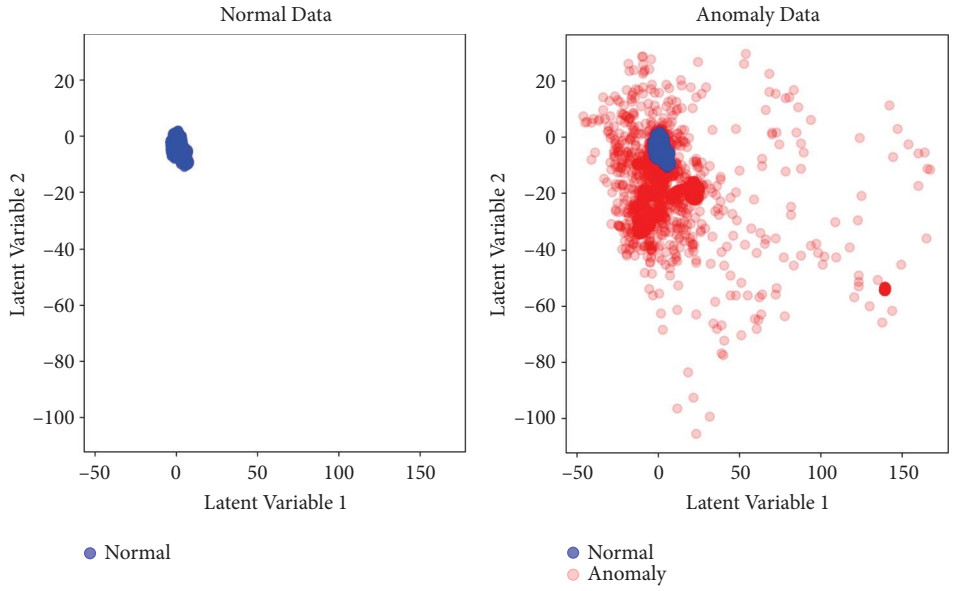
Model	SWaT				WADI				Power system			
	P	R	$F1$	AUC	P	R	$F1$	AUC	P	R	$F1$	AUC
LSTM-AD	0.9255	0.9952	0.9591	0.9480	0.0138	0.6623	0.0271	0.4986	0.2352	0.5186	0.6584	0.4698
USAD	0.9245	0.9952	0.9686	0.9472	0.1579	0.8295	0.2653	0.8625	0.9208	0.8154	0.8649	0.8941
MSCRED	0.9255	0.9952	0.9591	0.9479	0.1307	0.8943	0.1331	0.8461	0.9211	0.8154	0.8650	0.8942
MADGAN	0.9247	0.9952	0.9586	0.9473	0.0952	0.6541	0.1662	0.7536	0.9220	0.8154	0.8654	0.8943
DAGMM	0.9253	0.9952	0.9590	0.9478	0.0756	0.9971	0.1412	0.8563	0.9211	0.8154	0.8652	0.8943
OmniAnomaly	0.9235	0.9952	0.9580	0.9465	0.3158	0.6541	0.4260	0.8198	0.9225	0.8154	0.8646	0.8945
CAE-M	0.9260	0.9952	0.9594	0.9483	0.2621	0.7918	0.4119	0.8788	0.9208	0.9208	0.8649	0.8941
TranAD	0.9293	0.9952	0.9611	0.9510	0.2660	0.8295	0.4029	0.8877	0.9914	0.2931	0.4524	0.6460
Ours	**0.9764**	0.9923	**0.9878**	**0.9983**	**0.6477**	0.8295	**0.7274**	**0.9694**	0.9215	0.8154	**0.8659**	**0.9032**

Bold values indicate the model's highest score for the corresponding metric.

Table 5: The precision, recall, and $F1$ score and AUC score from variants.

Model	SWaT				WADI				Power system			
	P	R	$F1$	AUC	P	R	$F1$	AUC	P	R	$F1$	AUC
CAE_T_{Rec}	0.9260	0.9952	0.9594	0.9483	0.1653	0.5614	0.2554	0.7472	0.9208	0.9208	0.8649	0.8941
CAE_$T_{(\text{Rec+Svdd})}$	0.9271	0.9941	0.9805	0.9680	0.2648	0.9801	0.2756	0.8452	0.9234	0.8154	0.8661	0.8946
CAE_$T_{(\text{Rec+Pre})}$	0.9762	0.9997	0.9878	0.9671	0.2407	0.9999	0.3881	0.9627	0.9478	0.8154	0.8765	0.8890
CAE_T	0.9764	0.9923	0.9877	**0.9983**	**0.6477**	0.8295	**0.7274**	**0.9694**	0.9215	0.8154	0.8659	**0.9032**

Bold values indicate the model's highest score for the corresponding metric.



(a)

Figure 3: Continued.

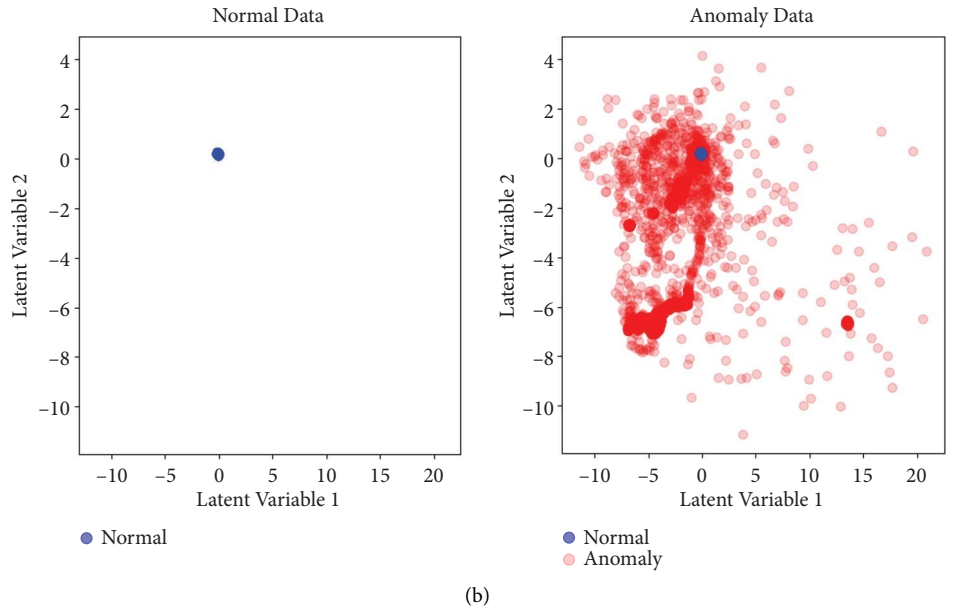


FIGURE 3: Visualization of representation: (a) visualization of representation before loss improvement and (b) visualization of representation after loss improvement.

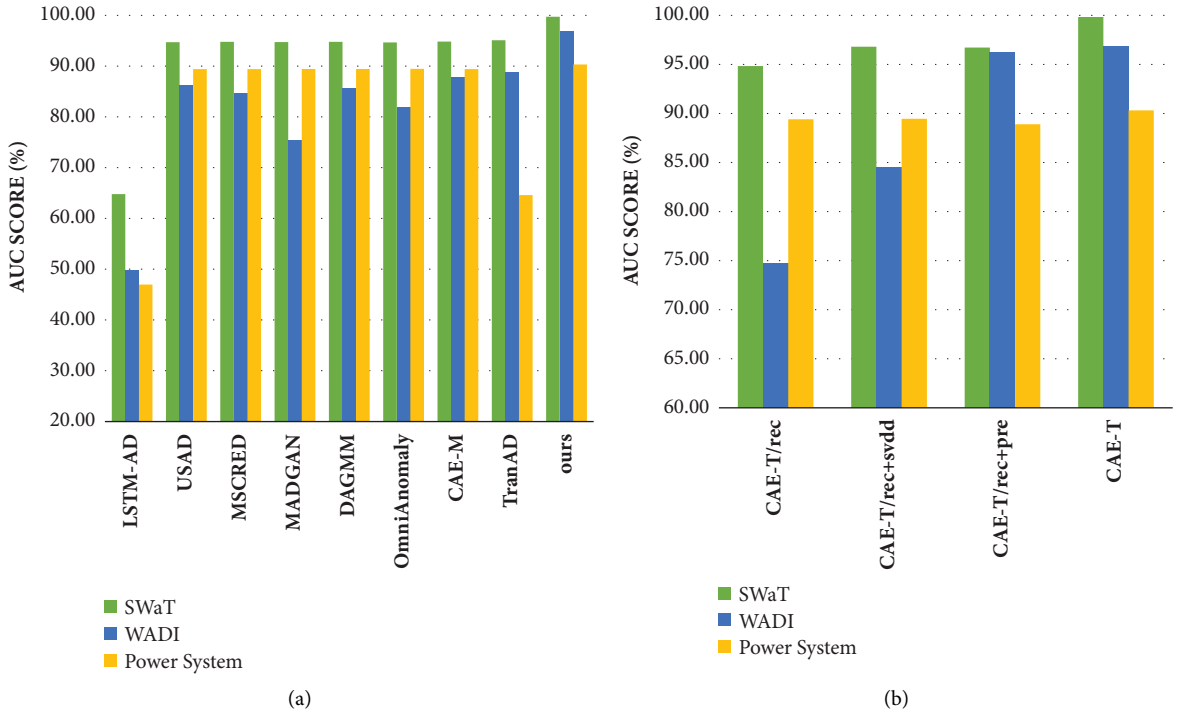


FIGURE 4: Ablation experiments. (a) Overall performance evaluation. (b) Effectiveness evaluation.

enhancing the model's ability to identify anomalies. As illustrated in Figure 3, the integration of SVDD significantly improved the representation of data features, defining a clear boundary for normal data in the high-dimensional feature space and effectively distinguishing anomalous points from normal ones.

On the other hand, the inclusion of the transformer-based anomaly score in CAE_$T_{(\text{Rec+Pre})}$ resulted in a substantial increase in recall to 0.9997 while maintaining a high AUC of 0.9671, underscoring the importance of time-series analysis in predicting and identifying anomalous behaviors. The full CAE-T model, combining both SVDD and the

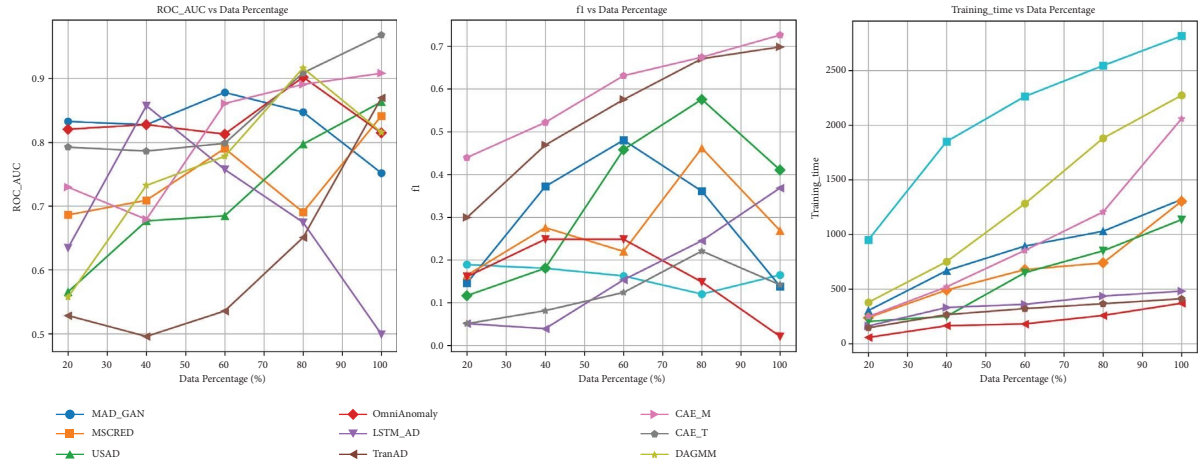


Figure 5: Training set size sensitivity experiments.

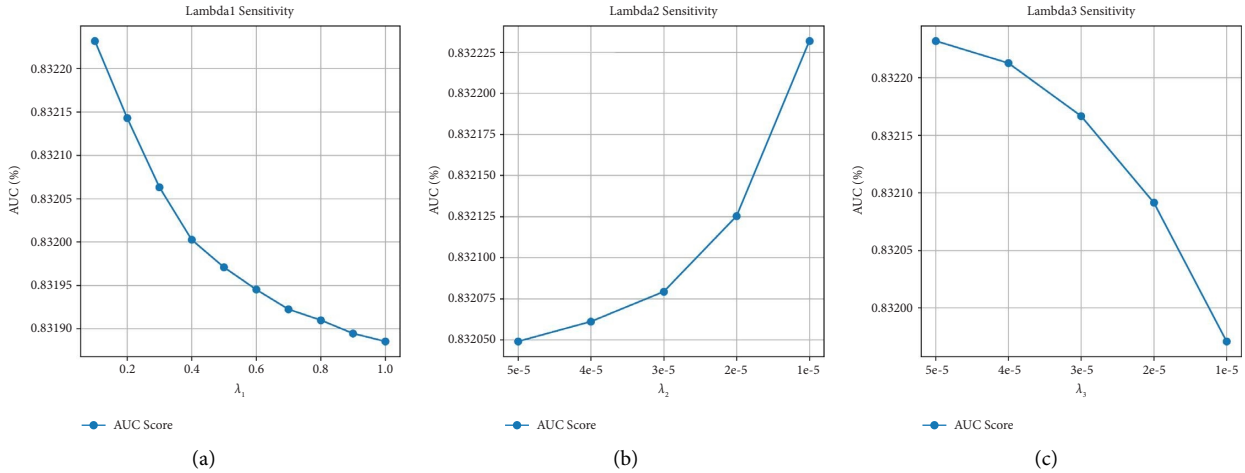


Figure 6: Parameters sensitivity experiments. (a) Lambda1 sensitivity. (b) Lambda2 sensitivity. (c) Lambda3 sensitivity.

transformer score, achieved peak performance in $F1$ score and AUC at 0.9877 and 0.9983, respectively, further validating the theory that the synergistic effect of these components significantly enhances anomaly detection capabilities.

On the WADI dataset, the complete CAE-T model excelled with its ability to handle complex data structures, achieving the highest $F1$ score of 0.7274 and an AUC of 0.9694, reaffirming its effectiveness in recognizing complex patterns. On the power system dataset, the CAE-T model demonstrated its excellent adaptability to different data distributions with an AUC of 0.9032, highlighting its strong capability in processing diverse data.

In conclusion, the technological enhancements of the CAE-T model not only exhibited outstanding performance on individual metrics but also demonstrated robustness and adaptability across multiple datasets, establishing its potential as an effective tool for anomaly detection in industrial control systems.

4.6.2. Training Set Size Sensitivity. This section evaluates the CAE-T model's sensitivity to training set size through an extensive analysis of Figure 5 data, focusing on $F1$ scores, AUC values, and training times for various models using the WADI dataset. The analysis aims to explore how different proportions of training data affect model performance and why the CAE-T model remains stable in varied training data environments.

Significant performance variations were noted across different training data proportions among the models, with the CAE-T model's performance standing out. The CAE-T model consistently showed superior AUC and a clear advantage in $F1$ scores across all data proportions. This finding highlights the efficacy of the CAE-T model in handling training data of varying sizes.

In contrast, models like LSTM-AD and MAD_GAN experienced declining $F1$ and AUC values with increasing training data proportions, indicating challenges in handling complex relationships in larger datasets. However, models

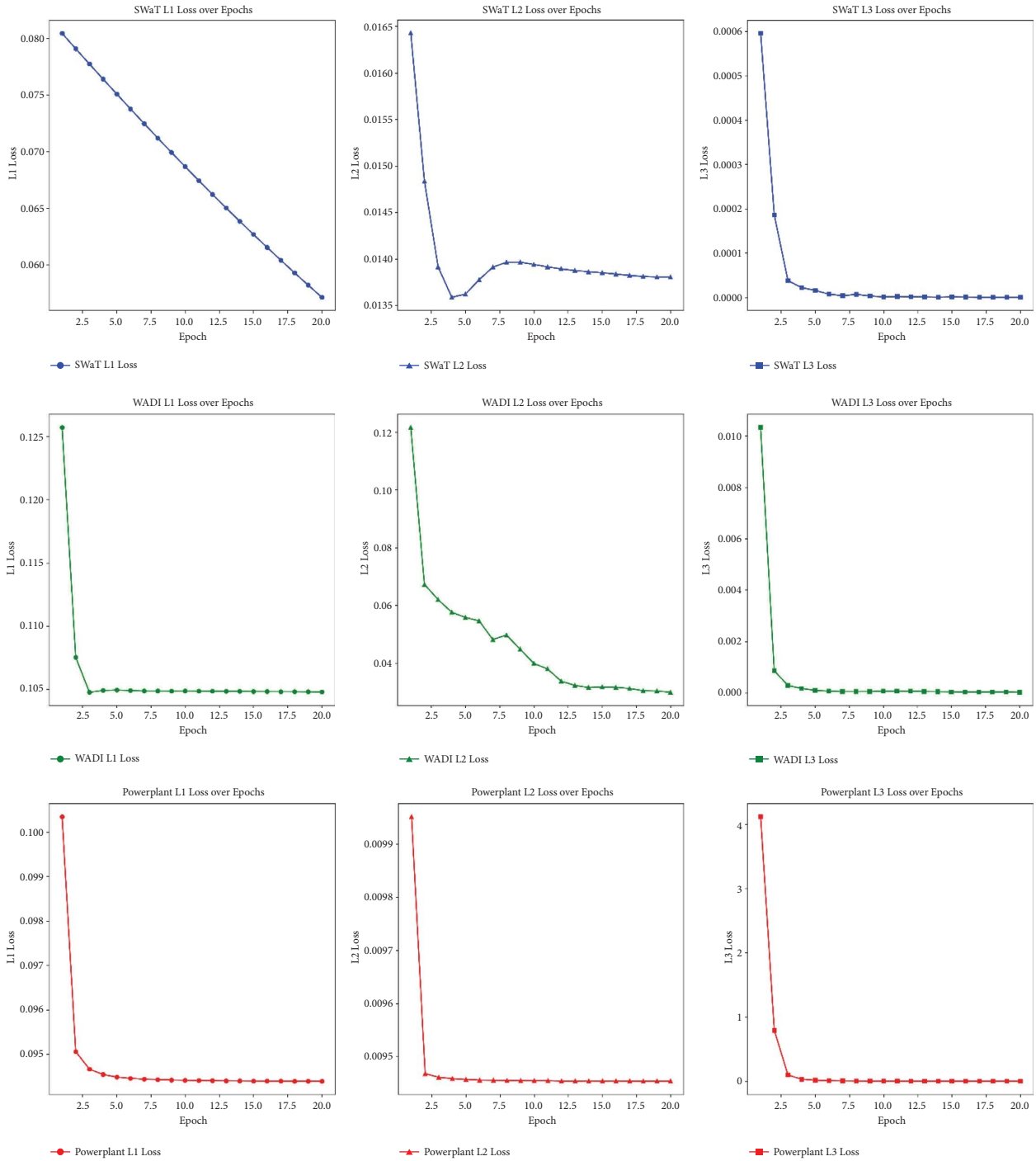


Figure 7: Parameters for loss experiments.

such as CAE-M and TranAD showed improved detection performance with larger training sets, demonstrating their ability to utilize more data to reveal complex patterns and enhance detection efficiency.

Regarding training time, the TranAD model significantly reduced its duration, benefiting from fully utilizing the transformer model. The transformer network in our temporal information extraction network achieved faster training speeds compared to the LSTM network in the CAE-M model. These results suggest that the CAE-T model excels not only in performance metrics but also in training efficiency, which is crucial for handling large datasets and complex data relationships.

We further explored the CAE-T model's ability to maintain stable performance across various training set sizes, likely due to its architectural strengths that enhance data generalization, regardless of volume. Moreover, our implementation potentially avoids overfitting, often a cause of performance degradation in other models with increased training set sizes. Investigating these aspects can provide

valuable insights into why the CAE-T model maintains consistent performance and what factors lead to the decline in performance of other models.

4.6.3. Parameters Sensitivity. In this research section, we conducted a sensitivity analysis of key parameters $\lambda 1$, $\lambda 2$, and $\lambda 3$ in our anomaly detection model. These parameters are crucial, regulating different model aspects: $\lambda 1$ influences the reconstruction error (Rec), $\lambda 2$ controls the weight of the support vector data description (SVDD), and $\lambda 3$ relates to the anomaly scoring of the transformer.

We conducted experiments using the SWaT dataset, following parameter configurations detailed in Table 3, including window size, batch size, training epochs, and learning rate. Our objective was to analyze each parameter's specific impact on model performance and identify the optimal settings. We adjusted $\lambda 1$ (0.1 to 1), $\lambda 2$, and $\lambda 3$ (5e-5 to 1e-5) values, ensuring the other two parameters remained at optimal AUC values for accurate and reliable results. As illustrated in Figure 6, the experiments revealed that the model exhibits optimal anomaly detection performance with $\lambda 1$ set to 0.1, $\lambda 2$ to 1e-5, and $\lambda 3$ to 5e-5.

Furthermore, we conducted a detailed convergence analysis of the model's reconstruction error, SVDD regularization term, and transformer-based prediction error term on three datasets. As shown in Figure 7, each type of loss function exhibited a stable convergence trend as the number of training iterations increased. This not only demonstrates the stability of the design of the objective function but also clearly shows that each parameter has a significant positive impact on model performance, which has been validated across different datasets.

These convergence results directly prove the optimizing role of our objective function in anomaly detection tasks with multidimensional time-series data. They provide strong evidence for the application of the model in complex scenarios and confirm the wide applicability of our chosen parameters. This further supports the effectiveness of our choices in model parameter optimization and offers practical grounds for adjusting and optimizing the model in similar industrial applications in the future.

In summary, through these supplementary experiments and analyses, we have gained a more comprehensive understanding of the impact of model parameters on performance and have obtained clear guidance for the further development and refined optimization of our model.

5. Conclusion and Future Work

This research introduces the CAE-T model, a novel anomaly detection framework based on transformer architecture, designed specifically for multivariate time series data. Featuring an encoder-decoder architecture, the CAE-T model efficiently trains and robustly identifies various anomaly patterns in complex datasets. While excelling in anomaly detection for multivariate time series data, the model shows potential for improvement in detection speed and data dependency. In particular, compared to the TranAD model, the CAE-T needs further structural optimization for more effective real-time or near-real-time anomaly detection applications [38].

Currently, the CAE-T model faces challenges in accurately localizing anomalies to specific sensors in complex industrial environments, particularly amid signal interference or obstruction. To overcome this challenge, we aim to integrate advanced collaborative detection and localization techniques, improving the model's ability to pinpoint anomalies accurately at the sensor level [39, 40]. Applying these techniques will lead to more precise device localization in challenging signal conditions and enhance real-time monitoring and anomaly detection in large sensor networks.

Future work involves integrating various transformer models, like Linformer [41], to bolster the CAE-T model's generalization across diverse temporal patterns. Additionally, we plan to explore advanced feature extraction technologies to enhance anomaly detection in limited data scenarios.

In summary, our research is dedicated to significantly enhancing the CAE-T model's performance, generalization capacity, and operational efficiency. This enhancement will improve the adaptability and robustness of anomaly detection systems in cybersecurity and industrial applications. Our goal is to broaden the CAE-T model's application across various practical scenarios, significantly contributing to anomaly detection [42–47].

Data Availability

We use public datasets, which are deposited in public repositories SWaT (https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/#swat), WADI (https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/#wadi), and Power System (https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets).

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The authors extend our sincere appreciation to the iTrust Centre for providing the industrial control systems dataset, which was instrumental in the completion of this research. This work was supported in part by National Natural Science Foundation of China under grant no. 62173101, the Basic and Applied Basic Research Funding of Guangdong Province under grant no. 2022A1515010865, the Guangzhou Science and Technology Funding under grant no. 202201020217, and the Open Fund Project of Information Security Assessment Center of Civil Aviation University of China under grant no. ISECCA-202201. Guangzhou University Graduate Student Basic Innovation Project 2022GDJC-M26.

References

[1] M. Serror, S. Hack, M. Henze, M. Schuba, and K. Wehrle, "Challenges and opportunities in securing the industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 2985–2996, 2020.

[2] D. Pliatsios, P. Sarigiannidis, T. Lagkas, and A. G. Sarigiannidis, "A survey on SCADA systems: secure protocols, incidents, threats and tactics," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1942–1976, 2020.

[3] C. Alcaraz and J. Lopez, "Digital twin: a comprehensive survey of security threats," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 3, pp. 1475–1503, 2022.

[4] M. Conti, D. Donadel, and F. Turrin, "A survey on industrial control system testbeds and datasets for security research," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2248–2294, 2021.

[5] G. Falco, C. Caldera, and H. Shrobe, "IIoT cybersecurity risk modeling for SCADA systems," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4486–4495, 2018.

[6] I. Makhdoom, M. Abolhasan, J. Lipman, R. P. Liu, and W. Ni, "Anatomy of threats to the internet of things," *IEEE communications surveys & tutorials*, vol. 21, no. 2, pp. 1636–1675, 2019.

[7] N. Neshenko, E. Bou-Harb, J. Crichigno, G. Kaddoum, and N. Ghani, "Demystifying IoT security: an exhaustive survey on IoT vulnerabilities and a first empirical look on Internet-scale IoT exploitations," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2702–2733, 2019.

[8] K. Tange, M. De Donno, X. Fafoutis, and N. Dragoni, "A systematic survey of industrial Internet of Things security: requirements and fog computing opportunities," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2489–2520, 2020.

[9] I. Stellios, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, "A survey of iot-enabled cyberattacks: assessing attack paths to critical infrastructures and services," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3453–3495, 2018.

[10] T. K. Das, S. Adepu, and J. Zhou, "Anomaly detection in industrial control systems using logical analysis of data," *Computers & Security*, vol. 96, 2020.

[11] S. E. Benkabou, K. Benabdeslem, V. Kraus, K. Bourhis, and B. Canitia, "Local anomaly detection for multivariate time series by temporal dependency based on Poisson model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6701–6711, 2022.

[12] Z. H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[13] X. Cui, S. Liu, Z. Lin et al., "Two-step electricity theft detection strategy considering economic return based on convolutional autoencoder and improved regression algorithm," *IEEE Transactions on Power Systems*, vol. 37, no. 3, pp. 2346–2359, 2022.

[14] J. Kuang, G. Xu, T. Tao, and Q. Wu, "Class-imbalance adversarial transfer learning network for cross-domain fault diagnosis with imbalanced data," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[15] M. Petković, S. Džeroski, and D. Kocev, "Feature ranking for semi-supervised learning," *Machine Learning*, vol. 112, no. 11, pp. 4379–4408, 2023.

[16] Z. Li, Y. Sun, L. Yang, Z. Zhao, and X. Chen, "Unsupervised machine anomaly detection using autoencoder and temporal convolutional network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[17] Z. Liao, Y. Li, E. Xia, Y. Liu, and R. Hu, "A twice denoising autoencoder framework for random seismic noise attenuation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[18] T. Ergen and S. S. Kozat, "Unsupervised anomaly detection with LSTM neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 8, pp. 3127–3141, 2020.

[19] M. Astekin, H. Zengin, and H. Sözer, "Evaluation of distributed machine learning algorithms for anomaly detection from large-scale system logs: a case study," in *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, pp. 2071–2077, IEEE, Venice, Italy, December 2018.

[20] B. Zong, Q. Song, M. R. Min et al., "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, February 2018.

[21] C. Zhang, D. Song, Y. Chen et al., "A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 1409–1416, 2019.

[22] Y. Zhang, Y. Chen, J. Wang, and Z. Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 2, pp. 1–2132, 2021.

[23] A. Dukkipati, D. Ghoshdastidar, and J. Krishnan, "Mixture modeling with compact support distributions for unsupervised learning," in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2706–2713, IEEE, Columbia, Canada, July 2016.

[24] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[25] D. Renaudie, M. A. Zuluaga, and R. Acuna-Agost, "Benchmarking anomaly detection algorithms in an industrial context: dealing with scarce labels and multiple positive types," in *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, pp. 1228–1237, IEEE, Venice, Italy, December 2018.

[26] E. Gyamfi and A. D. Jurcut, "Novel online network intrusion detection system for industrial IoT based on OI-SVDD and AS-ELM," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3827–3839, 2023.

[27] Z. Tian, C. Luo, J. Qiu, X. Du, and M. Guizani, "A distributed deep learning system for web attack detection on edge devices," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1963–1971, 2020.

[28] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Usad: unsupervised anomaly detection on multivariate time series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3395–3404, Anchorage, AK, USA, August 2020.

[29] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837, Anchorage, AK, USA, July 2019.

[30] A. Erba, R. Taormina, S. Galelli et al., "Constrained concealment attacks against reconstruction-based anomaly

detectors in industrial control systems," in *Proceedings of the 36th Annual Computer Security Applications Conference*, pp. 480–495, Austin, TX, USA, December 2020.

[31] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: deep transformer networks for anomaly detection in multivariate time series data," 2022, https://arxiv.org/abs/2201.07284.

[32] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S. K. Ng, "MAD-GAN: multivariate anomaly detection for time series data with generative adversarial networks," in *Proceedings of the International Conference on Artificial Neural Networks*, pp. 703–716, Berlin, Germany, September 2019.

[33] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979–16995, 2021.

[34] A. Siffer, P. A. Fouque, A. Termier, and C. Largouet, "Anomaly detection in streams with extreme value theory," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1067–1075, Halifax, NS, USA, August 2017.

[35] A. P. Mathur and N. O. Tippenhauer, "SWaT: a water treatment testbed for research and training on ICS security," in *Proceedings of the 2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater)*, pp. 31–36, IEEE, Vienna, Austria, April 2016.

[36] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: a water distribution testbed for research in the design of secure cyber physical systems," in *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, pp. 25–28, Pittsburgh, PA, USA, April 2017.

[37] S. Pan, T. Morris, and U. Adhikari, "Developing a hybrid intrusion detection system using data mining for power systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 3104–3113, 2015.

[38] K. Wang, A. Zhang, H. Sun, and B. Wang, "Analysis of recent deep-learning-based intrusion detection methods for in-vehicle network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1–12, 2022.

[39] Y. Xiong, N. Wu, H. Wang, and J. Kuang, "Cooperative detection-assisted localization in wireless networks in the presence of ranging outliers," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5165–5179, 2017.

[40] Y. Xiong, N. Wu, Y. Shen, and M. Z. Win, "Cooperative network synchronization: asymptotic analysis," *IEEE Transactions on Signal Processing*, vol. 66, no. 3, pp. 757–772, 2018.

[41] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: self-attention with linear complexity," 2020, https://arxiv.org/abs/2006.04768.

[42] S. Bagchi, T. F. Abdelzaher, R. Govindan et al., "New frontiers in IoT: networking, systems, reliability, and security challenges," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11330–11346, 2020.

[43] D. Upadhyay, J. Manero, M. Zaman, and S. Sampalli, "Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 1104–1116, 2021.

[44] M. Deng, X. Wu, P. Chen, and W. Zeng, "A hybrid column and constraint generation method for network behavior anomaly detection," in *Proceedings of the 2020 IEEE 20th International Conference on Communication Technology (ICCT)*, pp. 1107–1111, Nanning, China, October 2020.

[45] B. Hussain, Q. Du, B. Sun, and Z. Han, "Deep learning-based DDoS-attack detection for cyber–physical system over 5G network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 860–870, 2021.

[46] M. O. Mustafa, G. Georgoulas, and G. Nikolakopoulos, "Principal component analysis anomaly detector for rotor broken bars," in *Proceedings of theIECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society*, pp. 3462–3467, IEEE, Dallas, TX, USA, October 2014.

[47] J. Hu, K. Kaur, H. Lin et al., "Intelligent anomaly detection of trajectories for IoT empowered maritime transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 2382–2391, 2022.