A Standardized ICS Network Data Processing Flow With Generative Model in Anomaly Detection

TAO YANG, YIBO HU, YANG LI, WEI HU, AND QUAN PAN

1School of Automation, Northwestern Polytechnical University, Xi’an 710072, China
2School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798
3School of Cybersecurity, Northwestern Polytechnical University, Xi’an 710072, China

Corresponding author: Yibo Hu (hyb@mail.nwpu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 3102017OQD094 and in part by the Shanghai Aerospace Science and Technology Innovation Foundation under Grant 2018-008.

ABSTRACT
Industrial control systems (ICS) now usually connect to Wireless Sensor Networks and the Internet, exposing them to security threats resulting from cyber-attacks. However, detecting such attacks is a non-trivial task. The high-dimensional network data pose significant challenges on security anomaly detection. In this work, we propose a network flow data processing method, which can make the complex network data more standardized and unified to assist security anomaly detection. Then, data generation method is applied to collect enough training data. We also propose an evaluation method for generated data. Finally, the bidirectional recurrent neural networks with attention mechanism is proposed to extract the latent feature, and give an explainable result in identifying the dominant attributes. Empirical results show our method outperforms the state-of-the-art models.

INDEX TERMS
Industrial control systems network security, imbalanced data, data generation and evaluation, attention.

I. INTRODUCTION
The core of Industry 4.0 strategy is the deep integration of information systems and physical systems, which aims at promoting the intelligence, informatization and digitization of industrial production. However, while enjoying the convenience of the integration, the ICS also faces the threats caused by the complex network environment. ICS networks are more vulnerable than traditional Internet. Once the ICS network is attacked, the harm is much greater than that of the traditional Internet.

Traditional software security has relatively developed protection solutions [1], [2], but attacks and defenses on ICS have always been a concern of the academic. Gelone achieved the attack by consuming the battery energy of the sensor node, and evaluated the effect of the attack through a mathematical model [3], [4]. Domanska described the SerIoT which can optimize the information security in IoT platforms and networks in a holistic, cross-layered manner [5].

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

With the interconnection of ICS networks and traditional Internet, network security have become the core issue of ICS security. In view of the reliability and stability of anomaly detection systems (ADS) in traditional IT network [6], [7], many researches on ADS for ICS network has gradually emerged. Morris designed the Snort-based analysis and detection system, which could detect illegal data packets in Modbus traffic effectively. But the proposed system only could detect the known attacks, and the missing report rate (MRR) was high once an unknown attack occurs [8]. Ponomarev proposed a method to check session duration to verify whether the ICS network was attacked, as noticed that the periods of communication were longer than a given threshold, when injection attacks happened [9], [10]. Kebrilis designed a process-awareness defense and mitigation strategy using the support vector machine (SVM) model. The SVM model trained with anomaly data was used to monitor malicious behavior in real time [11]. Brun used dense random neural networks for online detection of IoT network attacks [12]. However, these studies were not really targeted at data features, and can’t capture the potential characteristics.
of network flows. With the advent of the KDD 99 dataset, researchers paid more attention to high-dimensional network flow data features in ADS [13]. Some researchers gradually used artificial intelligence techniques to analyze network flow data, and studied the general rules existing in network flow data [14]–[17].

Unlike KDD 99 dataset, the real ICS network data is not only small in size, but also have class distribution skews. Since the size of anomaly samples in the data domain may be much smaller than the normal samples, there is a situation where the dataset is imbalanced. If such imbalanced data is used for IDS model training directly, the detections results will be unreliable. In this situation, the lack of representative data will make learning difficult [18]. There are many researches on imbalanced learning [19]–[22], but there is barely on ADS, especially on ICS networks anomaly detection.

For its superiority in feature learning and Mega-data processing, CNN is widely used in computer vision, speech synthesis, natural language processing, and information filtering. Using deep learning for ADS is seldom tested before. The main reason perhaps is lack of ICS data and limited in low dimensional data. Therefore, proposing a standardized processing and data generation method for ICS Network Data with high dimension. This paper is organized as follows: Section II describes the proposed network data feature process methods. Section III discusses the data generation methods and proposes a evaluation method. Section IV describes the proposed data selection methods and attention mechanism. Section V discusses the experiments and results. Section VI presents conclusions.

II. NETWORK DATA FEATURE MINING
A. NETWORK FLOW DATA ACQUISITION
ICS network and traditional Internet have much in common in terms of the means of network attacks. Attacks such as port-scan, ping-sweep, and denial of service, which are common in traditional Internet, can also be used for attacks on ICS networks. Of course, they also have some differences, ICS is more special. Many behaviors may be normal in general terms, but are the attacks for the ICS systems, such as “Sending a fake command”, “Uploading exe files”, and so on. ADS is a defense means that is very practical and easy to implement. In order to have accurate ADS, sophisticated network data is necessary. At the network physical layer, an Ethernet frame is composed of the Ethernet header, and up to 1500 bytes of payload. This payload contains many attributes of the network connection. Analyzing the network data can help to defend against attacks. Data from ICS networks can be acquired through a API called PCAP. Libpcap and WinPcap are the front-end packet capture software libraries for many network tools. Tcpdump, Wireshark, Snort, and Nmap are popular sniffing programs when using pcap data [23]. The data acquisition of ICS networks is shown in Fig. 1.

Due to the lack of data in the field of ICS network security research, Lemay et al. used the SCADA sandbox to build an ICS environment similar to that in Fig. 1. The MTU polls the controllers to generate data, which is collected as the datasets that can be used for ICS network intrusion detection [24]. The datasets are certain universality and can be used as a typical representative of ICS network data. The ICS network datasets involved in this paper as shown in Table 1.

![Diagram](image.png)
TABLE 1. Four ICS network datasets generated under four different abnormal events respectively. Each dataset contains a varying number of abnormal traffic packages.

| Name | Content | Total Number | Anomaly Number |
|------|---------|--------------|----------------|
| Sending a fake command modbus 6RTU with operate | 11 minutes of regular Modbus traffic including polling and manual operation - 1 MTU, 6 RTU and 10 seconds polling interval. Also includes sending a Modbus write operation from a compromised RTU using Metasplot proxy functionality and the proxychains tool | 11166 | 10 |
| Moving two files modbus 6RTU with operate | 3 minutes of regular Modbus traffic including polling only - 1 MTU, 6 RTU and 10 seconds polling interval | 3319 | 75 |
| ChC uploading exe modbus 6RTU with operate | 1 minute of regular Modbus traffic including polling and manual operation - 1 MTU, 6 RTU and 10 seconds polling interval. Also includes sending an EXE file from a compromised RTU to another compromised RTU through a Metasplot meterpreter channel | 1426 | 121 |
| Exploiting ms08_netapi modbus 6RTU with operate | 5 minutes of regular Modbus traffic including polling and manual operation - 1 MTU, 6 RTU and 10 seconds polling interval. Also includes using an exploit ms08_netapi from a compromised RTU to compromise another RTU using Metasplot | 1856 | 1199 |

**B. NETWORK DATA PROCESSING**

Whether it is a traditional Internet or an ICS network, there are often multiple attributes in their network flows, such as destination address, packet length, protocol type, and so on. The original data are PCAP packets, which are parsed so that the data in PCAP format becomes structured data. After data cleaning, the dataset having 54-dimensional features is obtained. The following is a piece of the data flow:

1424799320, 658571, 62, 62, 0, C, 29, 3C, 11, 3F, 0, C, 29, EE, B7, 84, 8, 0, IP4, 192, 168, 1, 100, 192, 168, 1, 104, 20, 060, 48, 58160, 1, 0, 128, 6, 28, TCP, 3206, 502, 3482588417, 0, 28, 0, 0, 0, 0, 0, 1, 0, 64240, 0, 8, 0.

Among these features, some feature subsets define a certain attribute of the data stream. Therefore, these feature subsets are processed such that retaining the basic information while reducing redundancy.

For example, the feature subset \{0, C, 29, 3C, 11, 3F\} represents the source MAC address of the network packet which is “0:C:29:3C:11:3F”. The MAC address is a 48-bit address information and is the unique identifier of each hardware device. Every 4-bit are usually expressed as a hexadecimal number, so there is the above MAC address form. In other words, radix conversion of MAC address information does not change its physical meaning. Therefore, in order to reduce the dimension of the data, and retain the physical meaning of the MAC address at the same time, we convert it to decimal which is “0, 12, 41, 60, 17, 63”. Then, treating it as a 6-dimensional vector \(x_{MACs} = (0, 12, 41, 60, 17, 63)\), and calculating the 2-norm by

\[
||x_{MACs}|| = \left(\sum_{i=1}^{6} |x_{MACs}(i)|^2\right)^{\frac{1}{2}}
\]

Thus, the result is 98.4, and use it to replace the original 6-dimensional feature subset. The same method can be applied to the destination MAC address.

For the feature subset \{192, 168, 1, 100\}, which represents the source IP address of the network packet. Defining a 4-dimensional vector \(x_{IPs} = (192, 168, 1, 100)\) to quantify the IP address, and a 4-dimensional vector \(z = (192, 168, 1, 1)\) to represent the vector of the gateway address. Then, calculating the Euclidean distance between two vectors

\[
D_{IPs} = \left(\sum_{i=1}^{4} |x_{IPs}(i) - z(i)|^2\right)^{\frac{1}{2}}
\]

This result is 99, and use it to replace the original 4-dimensional feature subset. The same method can be applied to the destination IP address.

In this way, the initial 54-dimensional data is reduced into 36-dimensional data, which is shown in Fig. 2.

The acquired network packages have high-dimensional attributes that contain a large amount of characteristics. These characteristics mainly include the basic characteristics, traffic characteristics, and content characteristics of the network connection. The basic characteristics include all the attributes that can be extracted from the TCP/IP connection. The traffic characteristics are information such as the IP address and the
services of the network connections. The content characteristics are the attributes that describe the states of the network connection. Each row represents a network flow data, each column represents an attribute, and the last column is the label corresponding to each piece of data.

Faced with such high-dimensional network data, it is difficult to directly measure the impact of each attribute on network attack behavior analysis. The data types between the various attributes of network data are quite different. According to the data types, we divide all the attributes of network data into three sub-attribute sets: Descriptive fields, Proportional fields and Statistical fields. The specific divisions of ICS network data are shown in Table 2.

III. THE BALANCED METHODS BASED ON DATA GENERATION AND EVALUATION

In fact, the ICS network will be in a normal state in most cases. Therefore, anomaly data is only a small part, while the normal data is the overwhelming majority in the collected network data, which results in samples imbalanced problem. As shown in Table 1, the number of anomaly samples in "Sending a fake command modbus 6RTU with operate" is less than one thousandth of the total sample size. Training with such data will seriously affect the performance of the anomaly detection model. In order to represent the distribution of samples more intuitively, using the T-SNE dimensionality reduction method to visualize these 4 datasets. T-SNE is a manifold learning method that attempts to find a low-dimensional representation of the data [25]–[27]. The visualization results are shown in Fig. 3.

Common methods for solving sample imbalance problems include oversampling, undersampling, and cost sensitivity. Due to the small scale of ICS network data, this paper uses data generation methods to generate new samples of the minority (anomaly) class, which based on the idea of oversampling, but not simply repeating a few class samples. It’s not only make the number of normal and anomaly samples more balanced, but can increase the total amount of training samples.

A. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is an improved scheme based on random oversampling algorithm [28]. The generation of new samples is achieved by interpolating between the samples of the minority class. Assuming that there are a total of $n$ samples in the fewer classes, choosing a sample $x_i$ randomly and finding the $k$ nearest neighbors among these samples. Then, selecting one of the $k$ nearest neighbors sample $x_{ij}$ arbitrarily, the interpolation between it and $x_i$ will be a new sample

$$x_{new} = x_i + rand(0, 1) \cdot (x_{ij} - x_i)$$

s.t. $i \in \{1, 2, \cdots, n\}, j \in \{1, 2, \cdots, k\}$

where $rand(0, 1)$ represents a random number between 0 and 1. The anomaly samples are expanded to lower the proportion of positive and negative samples. The Fig. 4 are visualization results with a positive to negative sample ratio of 1:1.

B. MAHAKIL

MAHAKIL is a novel synthetic oversampling approach based on the theory of inheritance and the Mahalanobis distance [29]. It takes inspiration from the field of biology that offspring inherit traits from their parents by obtaining chromosomes from each parent in equal quantity. Splitting
TABLE 2. Dividing all attributes into three parts depending on the data type. The symbolic data is divided into Descriptive fields, the float-point data is divided into Proportional fields, and the integer data is divided into Statistical fields.

| ICS network data | Descriptive field | Proportional field | Statistical field |
|------------------|-------------------|---------------------|-------------------|
|                  | Type, Protocol    | Connection proportion (can be calculated from IP destination, IP source, MAC destination and MAC source, e.g., the proportion of connections with the same target host during a certain period of time.) | Epoch time, Frame length, Capture length, IP header length, DSF, IP total length, ID, Flags DF, Flags MF, Time to live, IP Payload, Source port, Destination port, Sequence number, Acknowledgment number, TCP header length, CWR, ECE, URG, ACK, PSH, RST, SYN, FIN, Window size value, Urgent pointer, Option, TCP payload |

FIGURE 5. The new samples generated by MAHAKIL are scattered, but many of them overlap the heterogeneous samples.

FIGURE 6. The processing of ICS network data generate by GAN.

FIGURE 7. The minority class data of 4 datasets are generated so that its ratio to majority class reaches 1:1 by GAN respectively, and the results are visualized by T-SNE and shown in the Fig. 7.

D. THE EVALUATION METHOD BASED ON MAHALANOBIS DISTANCE OF INTRA-CLASS AND INTER-CLASS SAMPLES

The above three data generation methods can perform sample expansion and balancing on ICS network data. But it’s unknown whether the generated data is valid and available. Therefore, an appropriate strategy must be established to evaluate the effectiveness of the generated data. According to the previous analysis of the visualization results, when the generated samples are too similar with the original same class samples, there will be clustering. It’s almost a multiple minimax game with value function

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_{z}(z)}[\log(1 - D(G(z)))]$$

where $p_{data}$ is the data distribution, $p_{z}$ is the prior distribution of the generative network. The discriminator aims to maximize the probability of distinguishing real from generated data. Whereas the generator is keeping the differentiation between real and generated data to a minimum so as to trick the discriminator into believing that generated examples are real.

Based on the originally given noise, GAN can generate data that matches the real situation and has diversity by self-game. Applying it to the ICS network data, and the process is shown in the Fig. 6.

The minority class data of 4 datasets are generated so that its ratio to majority class reaches 1:1 by GAN respectively, and the results are visualized by T-SNE and shown in the Fig. 7.
repetition of the original data. If the generated samples are
too different from the original same class samples, it will
blur the boundary between the anomaly data and the normal
data. This paper proposes that evaluating the quality of the
generated data needs to carried out in three claim: (i) Whether
the generated data has a good inheritance of the original
data characteristics. (ii) The diversity of the generated data.
(iii) The speed at which the data is generated. The visual-
ization results from the data generated by the previous three
methods can be roughly analyzed for the three methods. The
new samples generated by SMOTE preserve the characteris-
tics of the original data, but it gives rise to clustering. The
new samples generated by the MAHAKIL are diverse but
ignore the boundaries between samples of different classes.
GAN has the advantages of the former two methods, the data
generated by which is not only diverse, but also inherits the
characteristics of the original data well.

In order to further evaluate the performance of the three
methods, we propose a evaluation method based on Mahala-
nobis distance of intra-class and inter-class samples. The
Mahalanobis distance has been explained by equation 4,
which can measure the similarity between two sample sets.
Inspired by this, the similarity between the intra-class sam-
ple sets, which can measure the similarity between two sample sets.

\[ I_{x \sim y} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(x_i - \mu_x)^T \Sigma_y^{-1} (x_i - \mu_y)} \] (7)

where the \( I_{x \sim y} \) is similarity measure between the intra-class samples, the \( I_{y \sim x} \) is inheritance measure between the inter-class samples, \( \mu_x \) and \( \mu_y \) are the mean of the two class samples, \( \Sigma_x \) and \( \Sigma_y \) are the covariance matrix of the two class samples. These two measures can evaluate the claim
(i) and (ii) for data generated by three methods respectively.

IV. NETWORK FLOW DATA SELECTION AND ATTENTION

Descriptive fields are all symbolic data. These data are dis-
crete, and it is difficult to characterize and measure them
directly. We quantify these symbolic data with numbers. For
example, for the ICS network data, the protocol types are
[TCP, UDP, Modbus], then they can be quantified as [1]–[3].
In this way, all discrete data are quantized to facilitate feature
analysis. The quantized features are represented as \( F_D \).

Proportional fields are all float data. According to the
information such as the IP address and the MAC address
of the network connection, it is possible to obtain the number
of connections with the same target host or the same service
for a period of time, so that multiple proportional data can be
obtained. Some of these data is between [0, 1], and the value
of other data are greater than 1. We use Sigmoid function to
deal with the data, and map all of them to (0, 1]. Sigmoid
function is defined as

\[ S(x_{ij}) = \frac{1}{1 + e^{-x_{ij}}} \] (8)

where \( x_{ij} \) is the \( j \)-th sub-attribute value of the \( i \)-th piece
of data in the Proportional fields, \( S(x_{ij}) \) is the value after being
mapped by the Sigmoid function. The processed features are
represented as \( F_P \).

Statistical fields are all integer data, each attribute has a
different interval and cannot be directly analyzed. We use
the normalization method to process the statistical fields.
Let there be a total of \( n \) pieces of data, for the \( i \)-th piece
of data, the \( j \)-th attribute value in the statistical field is \( a_{ij} \),
the normalization process is

\[ Z_{ij} = \frac{a_{ij}}{\sum_{i=1}^{m} a_{ij}}, j \in [1, m] \] (9)

where \( Z_{ij} \) represents the normalized results, \( m \) is The total
number of attributes in the statistics field. The processed
features are represented as \( F_S \). With this method, the data
range of each attribute is as close as possible for horizontal
comparison and analysis. Meanwhile, it can keep the original
data distribution as much as possible.
### TABLE 3. The augmented datasets are the data with new generated samples, and the ratio of normal to anomaly is 1:1.

| Name | Original datasets | Augmented datasets |
|------|------------------|-------------------|
|      | total number | anomaly number | total number | anomaly number |
| Sending a fake command modbus 6RTU with operate | 11166 | 10 | 22291 | 11135 |
| Moving two files modbus 6RTU | 3319 | 75 | 6479 | 3235 |
| CnC uploading exe modbus 6RTU with operate | 1426 | 121 | 2605 | 1300 |
| Exploiting mo98 netapi modbus 6RTU with operate | 1856 | 1199 | 2348 | 1199 |

### TABLE 4. CNN model parameters.

| Parameter       | Value       |
|-----------------|-------------|
| Filter size     | 2 x 2, 3 x 3 |
| Filter Number   | 128, 128    |
| Dropout rate    | 0.5         |
| Batch size      | 64          |
| Epoch           | 100         |
| Learning rate   | 4e-4        |

![Diagram of CNN model parameters](image)

**FIGURE 8. The attention for ICS network data.**

### B. THE ATTENTION FOR ICS NETWORK DATA

In order to find out which attributes of the ICS network data have greater impact on the anomaly classification, this paper uses the attention mechanism [31], [32]. For a ICS network data \( \{F_1, F_2, \ldots, F_n\} \), where \( n \) is the number of features, and some of these features have connection with the neighbors, such as IP source & IP destination, MAC source & MAC destination. Therefore, using a Bi-RNN (bidirectional RNN) to encode the them. As shown in Fig. 8, \( h_i \) and \( \bar{h}_i \) are the hidden state that computed by Bi-RNN.

Concatenate \( h_i \) and \( \bar{h}_i \) to get an annotation of \( F_i \), i.e., \( \tilde{h}_i = \left[ h_i, \bar{h}_i \right] \). \( h_i \) summarizes the neighbor features around \( F_i \) but still focus on \( F_i \). Then, multiplying the each annotation \( h_i \) by each attention value \( \alpha_i \) and add them all together

\[
S = \sum_{i=1}^{n} \alpha_i h_i
\]  

where \( S \) is the attention output that that summarizes all the information of features in the ICS network data. The attention output \( S \) is a high level representation of the ICS network data and can be used as features for anomaly classification by Softmax. The attention value \( \alpha_i \) of each annotation \( h_i \) is computed by

\[
\alpha_i = \frac{\exp(u_i^T u_k)}{\sum_{k=1}^{n} \exp(u_i^T u_k)}
\]  

where \( u_i \) is the a hidden representation of \( h_i \), \( u_k \) is the nearby attributes vector. The \( \alpha_i \) reflects the importance of the annotation \( h_i \) with respect to the previous hidden state \( h_{i-1} \) in predicting the network data categories. This paper uses attention to analyze the importance of each feature in ICS network data for anomaly classification.

### V. RESULTS

#### A. DATASET

This paper uses the data generation methods to expand the minority samples in the ICS network dataset, so that the originally imbalanced data tends to be balanced. Then, we proposes an evaluation strategy for the effects of the three data generation methods. The datasets in Table 1 are collected in a complete and realistic scenario, which provided valuable data for ICS network security research [24]. For the 4 original datasets in the Table 1, using the three data generation methods to expand the original data, so that the normal and anomaly samples reach a certain proportion. The total number and the anomaly number are shown in the Table 3, and then performs the classification tests and evaluations.

#### B. THE CLASSIFICATION RESULTS OF NEW ICS NETWORK DATA

Convolution Neural Network (CNN) has become an industry standard technology [33], [34]. In academic direction, CNN is widely used in computer vision, speech recognition, speech synthesis, image synthesis, and natural language processing [35]–[37]. In industrial applications, CNN is widely used in autonomous driving, medical image recognition, and information filtering [38]–[40].

CNN also has good performance in extracting and classifying network traffic [41]. We uses CNN to classify ICS network anomaly behaviors. In fact, CNN has better talent for matrix operations, and it can extract object features in depth. Therefore, it has great performance in image recognition. Inspired by this, transforming the network data, fold the sequence data into matrix. For each piece of data in ICS network datasets, includes 2 Descriptive fields, 4 Proportional fields, and 30 Statistical fields. Transforming the \( \{F_D, F_P, \ldots\} \),
TABLE 5. The detection result of original data.

| Name | Pre  | Rec  | $F_1$ - Score |
|------|------|------|---------------|
| Sending a fake command modbus 6RTU with operate | -    | -    | -             |
| Moving two files modbus 6RTU  | -    | -    | -             |
| CnC uploading exe modbus 6RTU  | -    | -    | -             |
| Exploiting ms08 netapi modbus 6RTU with operate | 95.53 % | 71.10 % | 81.52 % |

TABLE 6. The detection result of the data after SMOTE.

| Name | Pre  | Rec  | $F_1$ - Score |
|------|------|------|---------------|
| Sending a fake command modbus 6RTU with operate | 96.79 % | 85.71 % | 81.01 % |
| Moving two files modbus 6RTU  | 86.72 % | 88.44 % | 87.57 % |
| CnC uploading exe modbus 6RTU  | 87.10 % | 90.63 % | 88.83 % |
| Exploiting ms08 netapi modbus 6RTU with operate | 96.36 % | 87.50 % | 91.72 % |

TABLE 7. The detection result of the data after MAHAKIL.

| Name | Pre  | Rec  | $F_1$ - Score |
|------|------|------|---------------|
| Sending a fake command modbus 6RTU with operate | 73.57 % | 83.33 % | 78.15 % |
| Moving two files modbus 6RTU  | 89.48 % | 66.67 % | 76.41 % |
| CnC uploading exe modbus 6RTU  | 90.44 % | 83.53 % | 86.85 % |
| Exploiting ms08 netapi modbus 6RTU with operate | 85.16 % | 88.24 % | 86.67 % |

TABLE 8. The detection result of the data after GAN.

| Name | Pre  | Rec  | $F_1$ - Score |
|------|------|------|---------------|
| Sending a fake command modbus 6RTU with operate | 76.31 % | 92.62 % | 82.25 % |
| Moving two files modbus 6RTU  | 90.67 % | 86.11 % | 83.33 % |
| CnC uploading exe modbus 6RTU  | 96.04 % | 85.71 % | 90.58 % |
| Exploiting ms08 netapi modbus 6RTU with operate | 93.45 % | 92.31 % | 92.88 % |

$F_S \}$ to a $6 \times 6$ matrix. The vacant part is filled with 0. The main parameters of the CNN model designed in this paper are shown in Table 4.

Finally, using CNN model to the ICS network data, and $Pre$ (Precision Rate), $Rec$ (Recall Rate), and $F_1$ – $Score$ are used to evaluate the classification effect. The results on original datasets are shown in Table 5. Due to the serious imbalance between the normal and anomaly ratios in the datasets, the classification results of the three datasets are seriously over-fitting. Then, we verify the augmented datasets with new data generated by three methods respectively, the results are shown in Table 6, 7, 8.

Comparing the experimental results of the above three methods. In general, the data generated by GAN has the best classification results. In other words, the data generated by GAN can assist the original data more effectively, thereby the training effect of the classification model is better. Compared with GAN, the SMOTE method is slightly inferior in terms of classification results, and followed by the MAHAKIL. For the first three ICS network datasets, the CNN model has experienced a serious over-fitting before data generation, and the classification results cannot be measured accurately. After data expansion and augment, these three datasets can be classified effectively, and there are good results on $Pre$, $Rec$ and $F_1$ – $Score$. In addition, GAN is a process of self-game, and the process of training the generator takes a long time. However, when the training is completed, it takes less time to generate the data than the other two methods.

C. THE EVALUATION OF GENERATED DATA

For the above three data generation methods, after analyzing the anomaly classification of the data, the experiments are further carried out by using the evaluation method proposed in this paper. For each data generation method, we measure the generated data by $I_{x \sim x}$ and $I_{x \sim y}$, and the results are shown in Table 9.

From the Table 9, the data generated by the MAHAKIL has the maximum $I_{x \sim x}$, which indicates that the generated data has the largest difference and is much more diverse. Followed by GAN, and then the SMOTE. For the $I_{x \sim y}$, the data generated by the GAN has the maximum value, the SMOTE gets a smaller value, and the MAHAKIL far from the other two methods. This result illustrates that the data generated by GAN has the best inheritance of the original data features, SMOTE is slightly inferior, and MAHAKIL is also not very good at it. In general, MAHAKIL achieves the highest diversity at the expense of the inheritance, while GAN is the most balanced method that get good diversity as well as retaining original data characteristics.

D. ATTENTION RESULT ON NETWORK DATA

For an anomaly data, using the attention mechanism to explore the features that has the greatest effect on the anomaly classification in the 36-dimensional data. Fig. 9 is an attention result visualized with a heat map.

Through the analysis by attention, it can be seen from Fig. 9 that the most important attributes in the ICS network
data: ‘‘325.1753373181921’’, ‘‘1621’’, ‘‘54’’, ‘‘104.0’’, ‘‘3223194217’’, ‘‘277.0523416251882’’, ‘‘100.0’’, ‘‘128’’, ‘‘54’’. These attributes represent: IP destination, Destination port, Frame length, MAC destination, Acknowledgment number, IP source, MAC source, Time to live, Capture length. These 9 attributes play a important role in the ICS network anomaly classification, although other attributes also play a more or less role. From this, we can known that there are only 1/4 of the attributes of 36-dimensional ICS network data play a key role.

VI. CONCLUSION

The paper aims to conduct intelligent anomaly detection on the ICS network, but finds that there is a serious problem of imbalanced sample in the ICS network data. Therefore, this paper uses three data generation methods to expand the ICS network data, and proposes an innovative evaluation method for generated data. Experiments show that this method can effectively evaluate the availability of data from the aspects of diversity and inheritance, and further prove that GAN is the state-of-the-art methods on the generation of ICS network data.

REFERENCES

[1] E. Gelenbe, P. Campiegiani, T. Czachórski, S. K. Katsikas, L. Komnios, L. Romano, and D. Tzovaras, “Security in computer and information sciences,” in Proc. 1st Int. ISICS Secur. Workshop, 2018, pp. 26–27.
[2] M. Siavvas, E. Gelenbe, D. Khegas, and D. Tzovaras, “Static analysis-based approaches for secure software development,” in Proc. Int. ISICS Secur. Workshop, 2018, pp. 142–157.
[3] E. Gelenbe and Y. M. Kadioglu, “Battery attacks on sensors,” in Proc. Int. Symp. Comput. Inf. Secur., Secure Workshop, 2018.
[4] E. Gelenbe and Y. M. Kadioglu, “Energy life-time of wireless nodes with network attacks and mitigation,” in Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops), May 2018, pp. 1–6.
[5] J. Domanska, E. Gelenbe, T. Czachorski, A. Drosou, and D. Tzovaras, “Research and innovation action for the security of the Internet of Things: The SerIoT project,” in Security in Computer and Information Sciences, vol. 821. Cham, Switzerland: Springer, 2018, pp. 101–118.
[6] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, “Detecting large-scale system problems by mining console logs,” in Proc. ACM SIGOPS 22nd Symp. Oper. Syst. Princ. (SOSP), 2009, pp. 57–46.
[7] D. Hadžiosmanović, D. Bolzoni, and P. H. Hartel, “A log mining approach for process monitoring in SCADA,” Int. J. Inf. Secur., vol. 11, no. 4, pp. 231–251, Aug. 2012.
[8] T. H. Morris, B. A. Jones, R. B. Vaughn, and Y. S. Dandass, “Deterministic intrusion detection rules for MODBUS protocols,” in Proc. 46th Hawaii Int. Conf. Syst. Sci., Jan. 2013.
[9] S. Ponomarev and T. Atkinson, “Industrial control system network intrusion detection by telemetry analysis,” IEEE Trans. Depend. Secur. Comput., vol. 13, no. 2, pp. 252–260, Apr. 2016.
[10] S. Ponomarev and T. Atkinson, “Session duration based feature extraction for network intrusion detection in control system networks,” in Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), Dec. 2016.
[11] A. Keliris, H. Salehghaffari, B. Cairl, P. Krishnamurthy, M. Maniakas, and F. Horrari, “Machine learning-based defense against process-aware attacks on industrial control systems,” in Proc. IEEE Int. Test Conf. (ITC), Nov. 2016.
[12] O. Brun, Y. Yin, and E. Gelenbe, “Deep learning with dense random neural network for detecting attacks against IoT-connected home environments,” Procedia Comput. Sci., vol. 134, pp. 458–463, 2018.
[13] N. V. Chawla, “Data mining for imbalanced datasets: An overview,” in Data Mining and Knowledge Discovery Handbook. Boston, MA, USA: Springer, May 2006, pp. 853–867.
[14] M. S. Pervez and D. M. Farid, “Feature selection and intrusion classification in NSL–KDD cup 99 dataset employing SVMs,” in Proc. 8th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA), Dec. 2014.
[15] D. H. Deshmukh, T. Ghorpade, and P. Padiya, “Intrusion detection system by improved preprocessing methods and Naïve Bayes classifier using NSL–KDD 99 Dataset,” in Proc. Int. Conf. Electron. Commun. Syst., Feb. 2014.
[16] H. H. Pajouh, G. Dastghaibyfard, and S. Hashemi, “Two-tier network anomaly detection model: A machine learning approach,” J. Intell. Inf. Syst., vol. 48, no. 1, pp. 61–74, Feb. 2017.
[17] M. S. Pervez and D. M. Farid, “Feature selection and intrusion classification in NSL–KDD cup 99 dataset employing SVMs,” in Proc. 8th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA), Dec. 2014.
[18] G. M. Weiss, “Mining with rarity: A unifying framework,” SIGKDD Explor. Newslett., vol. 6, no. 1, p. 7, Jun. 2004.
[19] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[20] N. V. Chawla, “Data mining for imbalanced datasets: An overview,” in Data Mining and Knowledge Discovery Handbook. Boston, MA, USA: Springer, May 2006, pp. 853–867.
[21] B. Krawczyk, “Learning from imbalanced data: Open challenges and future directions,” Prog. Artif. Intell., vol. 5, no. 4, pp. 221–232, Nov. 2016.
[22] G. E. A. P. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” SIGKDD Explor. Newslett., vol. 6, no. 1, p. 20, Jun. 2004.
[23] A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[24] A. Lemay and J. M. Fernandez, “Providing SCADA network data sets for intrusion detection research,” in *Proc. Workshop Cyber Secur. Exp. Test.* Berkeley, CA, USA: USENIX, 2016.

[25] L. van der Maaten and G. E. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[26] Y. Ma and Y. Fu, *Manifold Learning Theory and Applications.* Boca Raton, FL, USA: CRC Press, 2011.

[27] L. Cayton, “Algorithms for manifold learning,” Univ. California Berkeley, CA, USA: Tech. Rep. CS2008-0923, 2005, vol. 44.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jul. 2018.

[29] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah, “MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction,” *IEEE Trans. Softw. Eng.*, vol. 44, no. 6, pp. 534–550, Jun. 2018.

[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2014.

[31] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, arXiv:1409.0473. [Online]. Available: https://arxiv.org/abs/1409.0473

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016.

[33] J. Koushik, “Understanding convolutional neural networks,” 2016, arXiv:1605.09081. [Online]. Available: https://arxiv.org/abs/1605.09081

[34] A. Jacovi, O. Sar Shalom, and Y. Goldberg, “Understanding convolutional neural networks for text classification,” in *Proc. EMNLP Workshop Black-boxNLP, Analyzing Interpreting Neural Netw.* NLP, 2018.

[35] R. Girshick, “Fast R–CNN,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[36] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2015.

[37] Y. Hu, Y. Yi, T. Yang, and Q. Pan, “Short text classification with a convolutional neural networks based method,” in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018.

[38] B. Wu, A. Wan, F. Iandola, P. H. Jin, and K. Keutzer, “SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017.

[39] F. Milletari, N. Navab, and S.-A. Ahmadi, “V–Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016.

[40] O. Wu and W. Hu, “Web sensitive text filtering by combing semantics and statistics,” in *Proc. Int. Conf. Natural Lang. Process. Knowl. Eng.*, Mar. 2006.

[41] R. Blanco, J. J. Cilla, P. Malagón, I. Penas, and J. M. Moya, “Tuning CNN input layout for IDS with genetic algorithms,” in *Hybrid Artificial Intelligent Systems*. Cham, Switzerland: Springer, 2018, pp. 197–209.

**TAO YANG** received the B.S. and M.S. degrees from the School of Automation, Northwestern Polytechnical University, in 2006 and 2009, respectively, and the Ph.D. degree from the School of Computer, Northwestern Polytechnical University, in 2012. From 2009 to 2010, he was a Graduate Student with The Ohio State University, USA. From 2012 to 2015, he was an Assistant Professor with Northwestern Polytechnical University, where he is currently an Associate Professor and a Master’s Tutor. His research interests include data mining methodologies, machine learning algorithms, and information security.

**YIBO HU** received the B.S. degree in information security from Northwestern Polytechnical University, in 2017, where he is currently pursuing the master’s degree under the supervision of Q. Pan. His research interests include the industrial control system information security, deep learning, and natural language processing.

**YANG LI** received the B.S., M.S., and Ph.D. degrees from Northwestern Polytechnical University, in 2005, 2008, and 2012, respectively. From 2009 to 2011, he was a Graduate Student with the Laboratory of Professor Ryan Kastner, Department of Computer Science and Engineering, University of California at San Diego, San Diego. From 2014 to 2017, he worked as a Postdoctoral Researcher with the Laboratory of Professor Kastner and Department of Computer Science and Engineering, University of California at San Diego, San Diego. His current research interests include machine learning, deep learning, and natural language processing.

**WEI HU** received the B.S., M.S., and Ph.D. degrees from Northwestern Polytechnical University, in 2005, 2008, and 2012, respectively. From 2009 to 2011, he was a Graduate Student with the Laboratory of Professor Ryan Kastner, Department of Computer Science and Engineering, University of California at San Diego, San Diego. From 2014 to 2017, he worked as a Postdoctoral Researcher with the Laboratory of Professor Kastner and Department of Computer Science and Engineering, University of California at San Diego, San Diego. His current research interests include machine learning, deep learning, and natural language processing.

**QUAN PAN** was born in China, in 1961. He received the B.S. degree in automatic control from the Huazhong University of Science and Technology, in 1982, and the M.S. and Ph.D. degrees in control theory and application from Northwestern Polytechnical University. From 1991 to 1993, he was an Associate Professor with Northwestern Polytechnical University, where he has been a Professor with the Automatic Control Department, since 1997. He has authored 11 books, more than 400 articles. His research interests include information fusion, target tracking and recognition, deep network and machine learning, UAV detection navigation and security control, polarization spectral imaging and image processing, industrial control system information security, commercial password applications and modern security technologies. He is an Associate Editor of the journal *Information Fusion and Modern Weapons Testing Technology.*