Network Security Situational Awareness of Enterprise Control Systems under Machine Learning

Hui You

(Corresponding author: Hui You)

Network Security Defense Department, Beijing Police College
Nanjian Road, Changping District, Beijing 102202, China
Email: yhui1984@outlook.com

(Received Dec. 29, 2022; Revised and Accepted Dec. 26, 2023; First Online Feb. 23, 2024)

Abstract

In the process of operation, enterprise control system networks face complex network attacks and require enhanced protection. This study investigated the method of network security situational awareness (NSSA) for enterprise control systems. XGBoost was used to implement situational assessment, and an improved bat algorithm (IBA) was designed to optimize the parameters of XGBoost to obtain the IBA-XGBoost situational assessment method. Bidirectional long short-term memory (BiLSTM) was applied for situational prediction, and an IBA was also used to optimize parameters to achieve the IBA-BiLSTM situational prediction method. Tests were conducted using the NSL-KDD dataset. It was observed that the IBA-XGBoost method outperformed other machine learning methods, such as the KNN algorithm, in situational assessment. The obtained situation values closely aligned with actual values, demonstrating root-mean-square error (RMSE) and mean absolute error (MAE) values as low as 0.051 and 0.016, respectively. Additionally, IBA-BiLSTM outperformed the other algorithms in situation prediction, achieving an RMSE of 0.028 and an MAE of 0.021. These results validate the effectiveness of the proposed situation evaluation and prediction methods, showcasing their applicability in real-world enterprise control systems.

Keywords: Enterprise Control Network; Machine Learning; Network Security Situational Awareness; XGBoost

1 Introduction

The enterprise control system enables the automated control of various equipment and software involved in the production processes of enterprises, widely utilized in industries such as aviation and electric energy. With the continuous advancement of intelligent technology, an increasing number of sensors and devices are integrated into enterprise control systems, making the system network more susceptible to threats such as viruses and hackers. Traditional protection methods for enterprise control system networks include firewalls [11], intrusion detection [1], etc. However, faced with the growing complexity and frequency of external attacks, these conventional approaches struggle to comprehensively control the system's security status.

In contrast to traditional methods, network security situational awareness (NSSA) [9] technology can extract effective features from vast amounts of data, providing a timely and effective reflection of the system's security status. This enhances the network defense capability of enterprise control systems. Machine learning methods find extensive applications in NSSA [21]. Yao *et al.* [20] designed a framework that combines multivariate heterogeneous data based on the attack behavior model and proved its feasibility through experiments. He *et al.* [8] designed a situation prediction method using the dual-feedback Elman model and found through experiments that only four samples did not match the actual outcomes. Pavol *et al.* [12] compared statistical and neural network models in NSSA, concluding that the neural network method was more accurate than the traditional statistical model.

Tao *et al.* [16] reduced data dimensions through a stacked auto-coding network and used the output low-dimensional data as the input for a back-propagation neural network (BPNN) to assess situation. To further enhance NSSA effectiveness, this paper introduces a situation assessment and prediction method based on machine learning. Experiments were conducted on the designed method to evaluate its performance in addressing NSSA challenges. This work offers a novel security method for enterprise control systems, contributing to the safe operation of network systems.

2 Network Security Situation Awareness Methods

2.1 XGBoost-Based Situational Assessment Method

As industrialization and informationization continue to integrate, the interconnection between the enterprise control system and the Internet deepens. This integration aims to simultaneously enhance production efficiency and elevate management standards within enterprises. However, a surge in security threats emanating from the network also appears. Attacks on the control system network of enterprises often have a direct impact on their production and operational efficiency. In 2012, Saudi Arabia's National Petroleum Company experienced a cyberattack that paralyzed its internal network. In 2014, a Norwegian offshore oil platform fell victim to a network attack, resulting in production interruptions. Additionally, in 2015, the Ukrainian power system faced a large-scale cyberattack, leading to a prolonged power outage [18]. As technology continues to advance, the network security threats confronting enterprise control systems are becoming increasingly complex and diverse. This evolution necessitates heightened standards for network security protection.

NSSA can realize effective monitoring and analysis of the enterprise control system network, enabling early detection of potential attacks and reducing the extent of damage. NSSA includes situation assessment and situation prediction. The first one pertains to evaluating the present state of security, while the second one relates to forecasting the forthcoming status of network security. First of all, in terms of situation assessment, it is necessary to quantify the security situation according to certain indicators. This paper is based on the characteristics of the enterprise control system network and quantifies the security situation based on the attacks on the network. The details are shown below.

1) Attack probability P: The percentage of attack data out of the total amount of network data over a period of time.

2) Impact degree of attack Y: According to the common vulnerability scoring system (CVSS) [6], the impact on the network is categorized into three categories: confidentiality, integrity, and availability, and the weights are taken as 0.3, 0.1, and 0.6 respectively. The impact degree of the i-th kind of attack can be written as:

$$Y_i = round_2[\log_2(\frac{0.3 \times 2^{C_i} + 0.1 \times 2^{I_i} + 0.6 \times 2^{A_i}}{3})],$$

where $round_2$ means reserving two decimal places, C_i, I_i, and A_i corresponds to the impact value of confidentiality, integrity, and availability of the i-th kind of attack. The impact value is set as 0/0.2/0.6 corresponding to no (N), low (L), and high (H) impact.

Ultimately, the quantization yields a situation value of:

$$V = \frac{P \times \sum_{i=1}^{n} Y_i \times N_i}{N_A}$$

where N_i stands for the number of the i-th kind of attack and N_A stands for the count of attacks on the network.

Referring to the National Internet Emergency Center, the situation values are divided into four levels (Table 1).

Table 1: Classification of situation levels

Situation value	Security level
0-0.2	Excellent
0.2-0.4	Good
0.4-0.75	Medium
0.75-0.9	Poor
0.9-0.1	Dangerous

For the classification of attacks on the network, this paper chooses the XGBoost algorithm [10], a machine learning method, whose objective function can be written as:

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k),$$

where $\sum_(i = 1)^n l(y_i, \hat{y_i})$ is the error between the actual and predicted values, and $\sum_{k=1}^{K} \Omega(f_k)$ is the regularity term, which is employed for managing the intricacy of the model. Performing a Taylor second-order expansion on the above equation, at the t-th iteration, the objective function can be rewritten as:

$$obj^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t),$$

where g_i is the first-order derivative and h_i is the second-order derivative. After sorting, there is:

$$obj^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} (\frac{G_j^2}{H_j + \lambda}) + \gamma T,$$

where T is the count of leaf nodes, γ and λ are penalty factors.

The formula for the iterative decision tree can be written as: *** please add a formula. ***

where η is the learning rate of the iterative decision tree, which is used to control the iteration speed (0-1, 0.1 by default). In addition, the values of maximum depth of the tree and the number of weak classifiers, i.e., m and n, will also affect the accuracy of the algorithm for the classification of cyber-attacks. In order to obtain better performance, appropriate parameter adjustment is needed [15]. Therefore, this paper uses an improved bat

algorithm (IBA) to realize the optimization of η, m, and n in the XGBoost algorithm.

The bat algorithm (BA) is a swarm intelligence algorithm [7] and finds extensive usage in various optimization problems [19]. Assume that a bat flies at position h_i with a velocity of v_i, the frequency range of sound wave is $[f_{\min}, f_{\max}]$, the loudness range is $[A_{\min}, A_0]$, and the wavelength is λ, then the equation for updating the position and velocity of the bat can be written as:

$$\begin{aligned} v_i^t &= v_i^{t-1} + (x_i^{t-1} - x_g) f_i \\ x_i^t &= x_i^{t-1} + v_i^t, \end{aligned}$$

where x_g is the current global optimal position of the bat and f_i is the frequency of adjusting the bat's velocity, whose calculation formula is:

$$f_i = f_{\min} + \beta \times (f_{\max} - f_{\min}),$$

where β is a random number obeying a normal distribution in [0,1]. The local search process for the bat can be written as:

$$x_{new} = x_{old} + \alpha A_{avg}^t$$

where x_{new} is the new solution, x_{old} is the selected optimal old solution, $\alpha \in [-1, 1]$, and A_{avg}^t is the average loudness of the bat's sound waves at moment t. When a bat finds a prey, it raises the frequency of the sound wave and decreases the loudness of the sound wave to move towards the prey. The process can be written as:

$$\begin{aligned} A_i^{t+1} &= \alpha A_i^t, \\ r_i^{t+1} &= r_i^0 [1 - \exp(-\delta t)], \end{aligned}$$

where r_i^0 is the initial pulse emissivity, α and δ are constants.

In order to further improve the BA's optimization searching effect, the initialization of bat populations is implemented based on Tent chaotic mapping [4], and the process is as follows:

1) Initial value x_0 is randomly generated within the range of (0,1).

2) A sequence of Tent chaotic mappings is generated based on the following equation:

$$x_{k+1} = \begin{cases} 2x_k + rand(u\frac{n-k}{n}), & 0 \le x_k \le 0.5 \\ 2(1 - x_k) + rand(u\frac{n-k}{n}), & 0.5 < x_k \le 1 \end{cases}$$

where x_k denotes the value after k times of Tent chaotic mapping, n denotes the total number of calculations to be performed, u stands for the disturbance coefficient, and rand denotes a random number between 0 and 1.

3) The sequence is intercepted to obtain a number of numerical sequences, which are the initialized bat population.

The specific procedure of the IBA-XGBoost algorithm-based situation assessment method is as follows.

1) The parameters of the XGBoost algorithm are initialized, and bat individuals are encoded according to the parameters that need to be optimized.

2) The bat population is initialized using Tent chaotic mapping, and the optimal bat individual, i.e., the optimal parameter of the XGBoost algorithm, is calculated by IBA using the mean square error (MSE) as the objective function.

3) The optimal parameters obtained are used to build a situation assessment model, and a test set is input for model evaluation.

2.2 Bidirectional Long Short-Term Memory-Based Situation Prediction

There is a certain temporal pattern in network attacks on enterprise control systems. To address situation prediction, this paper selects bidirectional long short-term memory (BiLSTM), renowned for its effectiveness in temporal prediction, as the model. BiLSTM [14] addresses the limitations of traditional unidirectional long short-term memory (LSTM), which can only capture information from one direction. The BiLSTM architecture is depicted in Figure 1.

LSTM obtains the output through the calculation of three gates. Suppose the state of the hidden layer at the previous moment is h_{t-1}, the input at the current moment is x_t, then the output of the forgetting gate is f_t:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f).$$

The input gate is used to update important information, and its output is i_t:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i).$$

Cell state C_t at the current moment can be written as:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t,$$

where $\tilde{C}_t$ is the interim cell state: $\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$. Finally, the calculation formulas of output gate o_t and hidden layer state h_t at the current moment are:

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \\ h_t &= o_t \times \tanh(C_t). \end{aligned}$$

There are also parameters in BiLSTM that can affect the effectiveness of the situation prediction, and IBA is also used for optimization. The specific procedure of the IBA-BiLSTM-based situation prediction approach based on is presented below.

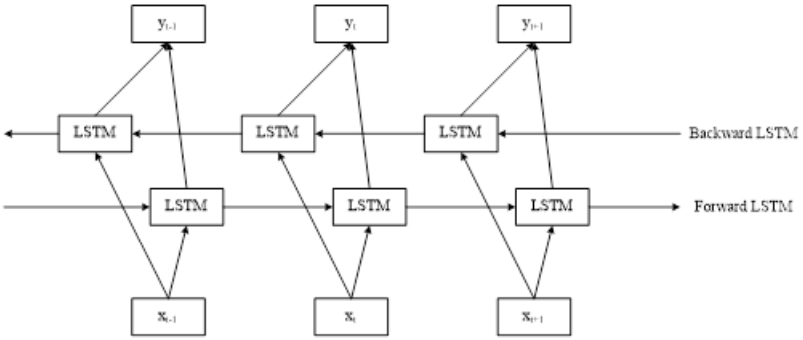

Figure 1: BiLSTM structure

1) The structure of BiLSTM is initialized. The optimization targets include the quantity of iterations, the Dropout ratio, and the count of units in hidden layers. Individual bats are encoded.

2) The parameters of BiLSTM are optimized using IBA.

3) The optimal parameters of BiLSTM are obtained to establish a situation prediction model. The model is utilized to generate prediction outcomes by inputting the test set.

3 Experiments and Analysis

3.1 Experimental Setup

Experiments were performed on a computer system that operated on Windows 10. This sytem was equipped with an Intel(R)Core(TM) i7-5500U processor and had a memory capacity of 8 GB. Python 3.9 programming language was used. The dataset used for the experiments was from NSL-KDD [17]. Each sample contained 41-dimensional features and one-dimensional labels, and the attacks are distributed as presented in Table 2.

Table 2: Distribution of cyber-attacks in the NSL-KDD dataset

	KDD Train+	KDD Test+
Normal	67343	9711
DoS	45927	7458
Probe	11656	2421
U2R	52	200
R2L	995	2654
Total	125973	22544

The situation values obtained using the IBA-XGBoost model were used as the data for the situation prediction, and the inputs and outputs of the IBA-BiLSTM model were determined using a sliding window with a value of 6. The situation values of the first five moments were taken as the inputs, which were used to predict the situation values of the latter moments. The performance of both the situation assessment and prediction methods was evaluated using the following two indicators.

Assuming that the actual value is y_i and the output value of the model is $\hat{y}_i$.

1) Root-mean-square error (RMSE): A quantification of the disparity between the observed value and the estimated value:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2};$$

2) Mean absolute error (MAE): The actual situation of the error of the estimated value:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

3.2 Results Analysis

Ten samples were randomly selected to compare the IBA-XGBoost model with other machine learning methods:

1) K-nearest neighbor (KNN) [5];

2) Support vector machine (SVM) [3];

3) Decision tree (DT) [13].

The findings are presented in Table 3.

Table 3 reveals that the KNN method exhibited two sample evaluation errors. Specifically, the evaluation result for sample 4 was medium, while the actual level was poor. Similarly, the evaluation result for sample 7 was medium, while the actual level was poor. The SVM method demonstrated two sample evaluation errors. Sample 4 received a medium evaluation, but it was poor actually; sample 8 was rated as poor, but its actual level was dangerous. The DT method obtained a sample evaluation error. Its assessment for sample 4 was medium, but it was actually poor. Both the XGBoost and IBA-XGBoost methods accurately evaluated all the ten samples, highlighting the superior performance of the XGBoost-based

Table 3: Results of security situation assessment

	KNN	SVM	DT	XGBoost	IBA-XGBoost	Actual value
1	0.155/excellent	0.107/excellent	0.145/excellent	0.133/excellent	0.125/excellent	0.12/excellent
2	0.268/good	0.213/good	0.262/good	0.251/good	0.244/good	0.24/good
3	0.584/medium	0.589/medium	0.579/medium	0.546/medium	0.551/medium	0.55/medium
4	0.721/medium	0.732/medium	0.748/medium	0.761/poor	0.782/poor	0.77/poor
5	0.264/good	0.256/good	0.212/good	0.225/good	0.232/good	0.23/good
6	0.397/good	0.391/good	0.384/good	0.357/good	0.367/good	0.36/good
7	0.734/medium	0.752/poor	0.801/poor	0.771/poor	0.785/poor	0.78/poor
8	0.951/dangerous	0.889/poor	0.935/dangerous	0.927/dangerous	0.912/dangerous	0.91/dangerous
9	0.289/good	0.221/good	0.231/good	0.247/good	0.255/good	0.25/good
10	0.961/dangerous	0.959/dangerous	0.907/dangerous	0.945/dangerous	0.934/dangerous	0.93/dangerous

Note: Bolding indicates that the output situation level does not match the reality.

assessment method. To further understand the performance of different assessment methods, RMSE and MAE were compared, as depicted in Figure 2.

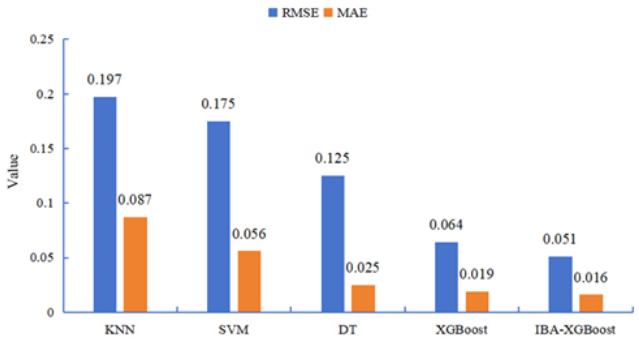

Figure 2: Comparative results of RMSE and MAE on the situational assessment

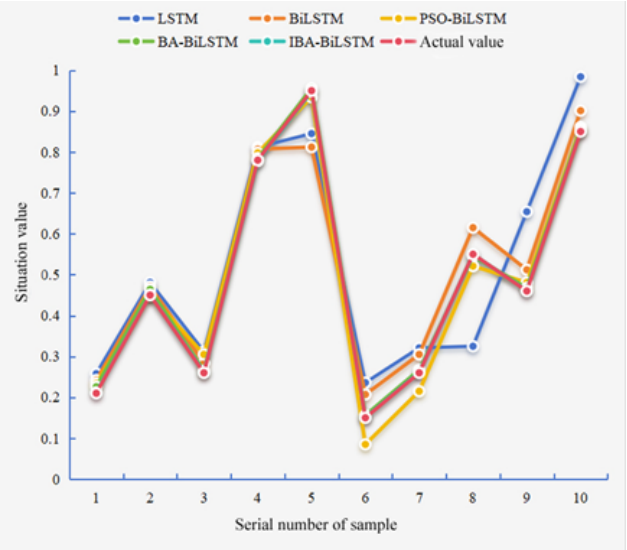

Figure 3: Security situation prediction results

Observing Figure 2, it becomes evident that the KNN method exhibited the poorest performance in assessing the cybersecurity situation of enterprise control systems, with an RMSE and MAE of 0.197 and 0.087 respectively. The SVM and DT methods had an RMSE value greater than 0.1. In comparison, the RMSE and MAE of the XGBoost method was 0.064 and 0.019, respectively, both markedly lower than the values of the KNN, SVM, and DT methods. Subsequently, after parameter optimization by IBA, the RMSE of the IBA-XGBoost approach was 0.051, reflecting a 20.31% reduction compared to the XGBoost method, and the MAE was 0.016, demonstrating a 15.79% reduction compared to the XGBoost method. This result demonstrated the efficacy of IBA in enhancing the performance of the XGBoost method. Again with ten samples, the IBA-XGBoost method was compared with the following methods: LSTM, BiLSTM, BiLSTM optimized by particle swarm algorithm (PSO) [2]: PSO-BiLSTM, BA-BiLSTM, and IBA-BiLSTM. The comparative results are presented in Figure 3.

In Figure 3, it is evident that the predicted values obtained by the LSTM method exhibited a large gap from the real values, accompanied by considerable fluctuations, notably in the prediction for sample 8 where the disparity was pronounced. Subsequently, the prediction results of the BiLSTM and PSO-BiLSTM methods slightly outperformed the LSTM method, yet there remained a discernible gap from the real values. In contrast, the predicted values of the BA-BiLSTM and IBA-BiLSTM methods aligned closely with the real values, indicating their superior prediction capabilities. The calculated RMSE and MAE results were compared in Figure 4.

The findings illustrated in Figure 4 suggest that, among the compared methods, the LSTM method exhibited a poor performance in situation prediction. In contrast, the BiLSTM method outperformed the LSTM method with a lower RMSE of 0.107 and MAE of 0.061, indicating that the BiLSTM approach was more adept at capturing temporal information in situation values over time, leading to superior results compared to the LSTM

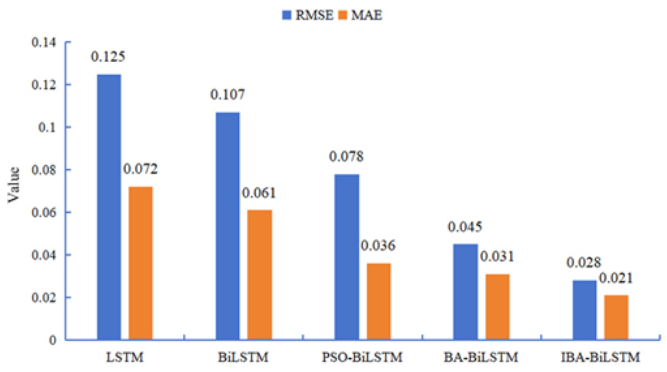

Figure 4: Comparative results of RMSE and MAE

method. The RMSE of the BA-BiLSTM method was 0.045, representing a 42.31% reduction compared to the PSO-BiLSTM method, while the MAE was 0.031, signifying a 13.89% reduction compared to the PSO-BiLSTM method. This result demonstrated that, in comparison to PSO, BA yielded superior parameter optimization effects for BiLSTM. Finally, the RMSE of the IBA-BiLSTM approach was 0.028, demonstrating a 37.78% reduction compared to the BA-BiLSTM approach, and its MAE was 0.021, indicating a 32.26% reduction compared to the BA-BiLSTM approach. This result confirmed that the IBA was more effective in enhancing prediction performance.

4 Conclusion

This paper studied the NSSA challenges faced by enterprise control systems based on machine learning. The IBA-XGBoost method and the IBA-BiLSTM method were designed for situation evaluation and prediction respectively. Through experimentation on the NSL-KDD dataset, it was observed that the two approaches exhibited superior performance in both situation assessment and prediction. They are effective in obtaining more accurate situation values for determining security levels and making reliable predictions for future situation values. The two methods hold promising potential for practical applications in real-world enterprise control systems.

References

[1] A. Alamleh, O. S. Albahri, A. A. Zaidan, A. H. Alamoodi, A. S. Albahri, B. B. Zaidan, S. Qahtan, "Multi-attribute decision-making for intrusion detection systems: A systematic review," *International Journal of Information Technology & Decision Making*, vol. 22, no. 01, pp. 589-636, 2023.

[2] A. Amiri, A. Salmasnia, M. Zarifi, M. R. Maleki, "Adaptive Shewhart control charts under fuzzy parameters with tuned particle swarm optimization algorithm," *Journal of Industrial Integration and Management*, vol. 8, no. 2, pp. 241-276, 2023.

[3] F. Camastra, V. Capone, A. Ciaramella, A. Riccio, A. Staiano, "Prediction of environmental missing data time series by support vector machine regression and correlation dimension estimation," *Environmental Modelling & Software*, vol. 150, pp. 1-7, 2022.

[4] S. Chen, S. Wang, "An optimization method for an integrated energy system scheduling process based on NSGA-II improved by tent mapping chaotic algorithms," *Processes*, vol. 8, no. 4, pp. 1-11, 2020.

[5] P. Chumnanpuen, "K-nearest neighbor and random forest-based prediction of putative tyrosinase inhibitory peptides of abalone haliotis diversicolor," *Molecules*, vol. 26, no. 12, pp. 1-10, 2021.

[6] K. Gencer, F. Baifti, "The fuzzy common vulnerability scoring system (F-CVSS) based on a least squares approach with fuzzy logistic regression," *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 145-153, 2020.

[7] Y. Guo, J. Chen, "An anomaly feature mining method for software test data based on bat algorithm," *International Journal of Data Mining and Bioinformatics*, vol. 27, pp. 58-72, 2022.

[8] J. He, J. Yang, "Network security situational level prediction based on a double-feedback elman model," *Informatica: An International Journal of Computing and Informatics*, vol. 46, no. 1, pp. 87-93, 2022.

[9] M. Husák, L. Sadlek, S. Špaček, M. Laštovička, M. Javorník, J. Komárková, "CRUSOE: A toolset for cyber situational awareness and decision support in incident handling," *Computers & Security*, vol. 115, pp. 1-5, 2022.

[10] S. T. Ikram, A. K. Cherukuri, B. Poorva, P. S. Ushasree, Y. S. Zhang, X. Liu, G. Li, "Anomaly detection using xgboost ensemble of deep neural network models," *Cybernetics and Information Technologies*, vol. 21, no. 3, pp. 175-188, 2021.

[11] S. H. Mohammed, A. D. Jasim, "Evaluation of firewall and load balance in fat-tree topology based on floodlight controller," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, pp. 1157-1164, 2020.

[12] S. Pavol, Staňa Richard, G. Andrej, et al., "Network security situation awareness forecasting based on statistical approach and neural networks," *Logic Journal of the IGPL*, vol. 31, no. 2, pp. 352–374, 2023.

[13] A. Pradeepika, R. Sabitha, "Examination of diabetes mellitus for early forecast using decision tree classifier and an innovative dependent feature vector based naive bayes classifier," *ECS Transactions*, vol. 107, no. 1, 2022.

[14] A. Shaikh, M. Bhargavi, C. P. Kumar, "An optimised Darknet traffic detection system using modified locally connected CNN - BiLSTM network," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 43, pp. 87-96, 2023.

[15] P. Srinivas, R. Katarya, "hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost,"

Biomedical Signal Processing and Control, vol. 73, pp. 1-10, 2022.

[16] X. Tao, K. Kong, F. Zhao, S. Cheng, S. Wang, "An efficient method for network security situation assessment," *International Journal of Distributed Sensor Networks*, vol. 16, no. 11, pp. 1-13, 2020.

[17] M. Tavallaee, E. Bagheri, W. Lu, A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA'09)*, pp. 53-58, 2009.

[18] J. Weiss, "Industrial control system cyber security and the critical infrastructures," *OR Insight*, vol. 19, no. 4, pp. 33-36, 2016.

[19] C. Yang, L. L. Sun, H. Guo, Y. S. Wang, Y. Shao, "A fast 3D-MUSIC method for near-field sound source localization based on the bat algorithm," *International Journal of Aeroacoustics*, vol. 21, no. 3/4, pp. 98-114, 2022.

[20] Y. Yao, Y. Sun, Z. Liu, X. Meng, Z. Liu, "A data fusion framework of multi-source heterogeneous network security situational awareness based on attack pattern," *Journal of Physics: Conference Series*, vol. 1550, no. 6, pp. 1-12, 2020.

[21] B. Zhu, Y. Chen, Y. Cai, "Three kinds of network security situation awareness model based on big data," *International Journal of Network Security*, vol. 21, no. 1, pp. 115-121, 2019.

Biography

Hui You, born in June 1984, has received the doctor's degree from Beijing Normal University in 2020. She is an associate professor and is working in Beijing Police College. She is interested in mathematical modeling, network security, and big data.