Securing Critical Infrastructures: Deep Learning-based Threat Detection in the IIoT

Keping Yu, Liang Tan, Shahid Mumtaz, Saba Al-Rubaye, Anwer Al-Dulaimi, Ali Kashif Bashir, Farrukh Aslam Khan

***Abstract*—The Industrial Internet of Things (IIoT) is a physical information system developed based on traditional industrial control networks. As one of the most critical infrastructure systems, the IIoT is also a preferred target for adversaries engaged in advanced persistent threats (APTs). To address this issue, we explore a deep learning-based proactive APT detection scheme in the IIoT. In this scheme, considering the characteristics of long attack sequences and long-term continuous APT attacks, our solution adopts a well-known deep learning model, bidirectional encoder representations from transformers (BERT), to detect APT attack sequences. The APT attack sequence is also optimized to ensure the model's long-term sequence judgment effectiveness. The experimental results not only show that the proposed deep learning method has feasibility and effectiveness for APT detection but also certify that the BERT model has better accuracy and a lower false alarm rate when detecting APT attack sequences than other time series models.**

***Index Terms*—Cybersecurity, critical infrastructure systems, deep learning, APT, proactive detection, IIoT.**

I. Introduction

THE Industrial Internet of Things (IIoT) relies on large amounts of collected industrial data for troubleshooting, identifying performance bottlenecks, and detecting malicious behavior to achieve efficient control of the physical world [1] [2]. Over time, the IIoT has been gradually applied to national critical infrastructure systems such as those supporting the petrochemical industry, power grids, water conservancy, nuclear energy, and transportation [3]. The rapid development of the IIoT has been accompanied by the emergence of cyberattacks against critical infrastructure. In addition to traditional network attacks, advanced persistent threat (APT) attacks are also increasing. An APT is a prolonged and targeted cyberattack in which an intruder gains access to critical infrastructure systems and remains undetected until the target system is destroyed [4]. APTs pose a serious threat to critical infrastructure systems and have caused many serious accidents.

APT attacks are a major threat to critical infrastructure systems. Current APT detection methods are mainly based on distributed computing, big data, cloud computing, and data mining technologies [5], such as host malicious code anomaly detection, sandbox malicious code anomaly detection, correlation analysis, traffic anomaly detection, and comprehensive network capture. Traditional methods have not been directly applied to the IIoT. In [6], the authors propose an APT detection method that analyzes social network security events. By combining cloud computing with network traffic analysis, a flow reverse detection model based on traffic changes is developed [7]. The main idea is to establish an application container that runs a program to detect APT attacks by monitoring the behavior of network endpoints. In addition, a defense architecture, including APT gateway detection and an APT management console, is developed to monitor and analyze the host system environment, application environment, communication environment, data environment, traffic characteristics, and network protocol characteristics. The APT detection methods mentioned above have some validity, but they are mainly used for APT attacks with short attack durations and fixed attack patterns, while APT attacks in the IIoT are usually characterized by large scale and long duration. Thus, existing detection methods are not able to provide high detection accuracy when used with the IIoT [8].

The integration of artificial intelligence (AI) and IoT is currently a hot research topic [9][10]. The combination of AI and IoT leads to a very powerful technology, the AIoT, that can enable devices to collect data and analyze it to make human-like decisions. With the popularity of the AIoT, APTs and zero-day vulnerabilities have appeared in some important IIoTs. AI, especially deep learning, has advantages over traditional methods in detecting and defending against such attacks. To propose an APT detection method suitable for the IIoT in critical infrastructure systems, this paper investigates a deep learning-based proactive APT detection scheme in the IIoT based on the characteristics of APT attacks in the IIoT, such as long attack sequences and a long-term attack duration. The main contributions of this paper are as follows.

1) We apply the well-known bidirectional encoder represen-

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant JP18K18044 and JP21K17736, in part by the National Natural Science Foundation of China under Grant No. 61373162, in part by the Sichuan Provincial Science and Technology Department Project under Grant No. 2019YFG0183, and in part by the Sichuan Provincial Key Laboratory Project under Grant No. KJ201402. (Corresponding author: Liang Tan)

Keping Yu is with the College of Computer Science, Sichuan Normal University, Chengdu 610101, China and the Global Information and Telecommunication Institute, Waseda University, Tokyo, Japan. (e-mail: keping.yu@aoni.waseda.jp)

Liang Tan is with the College of Computer Science, Sichuan Normal University, Chengdu 610101, China, and the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China. (e-mail: jkxy_tl@sicnu.edu.cn, Corresponding author)

S. Mumtaz is with the Instituto de Telecomunicações, Aveiro, Portugal. (e-mail: Dr.shahid.mumtaz@ieee.org)

Saba Al-Rubaye is with the School of Aerospace, Transport and Manufacturing, Cranfield University, UK. (e-mail: s.al-rubaye@cranfield.ac.uk)

A. Al-Dulaimi is with EXFO Inc., Montreal, Canada. (e-mail: anwer.al-dulaimi@exfo.com)

A. K. Bashir is with Department of Computing and Mathematics, E-154, John Dolton, Chester Street, M15 6H, Manchester Metropolitan University, Manchester, United Kingdom. (e-mail: dr.alikashif.b@ieee.org).

Farrukh Aslam Khan is with Center of Excellence in Information Assurance (CoEIA), King Saud University, Saudi Arabia (e-mail: fakhan@ksu.edu.sa).

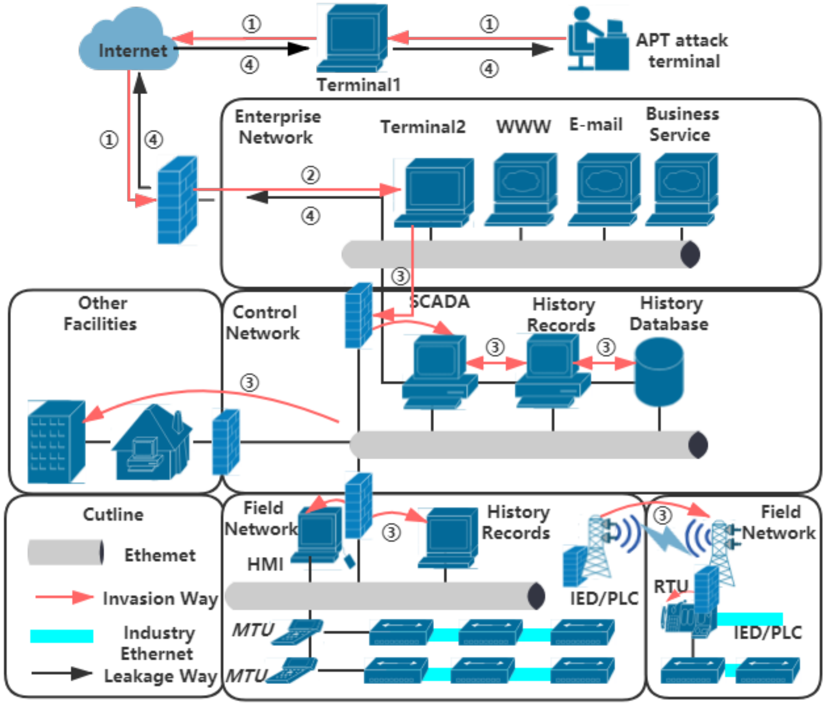

Fig. 1: IIoT-APT attack structure

tations from transformers (BERT) sequence processing model in the deep learning area to APT attack sequence detection.

2) We optimize the APT attack sequence data to normalize it and make it more obvious without destroying any content of the original data, ensuring the effectiveness of the trained model in long-term sequence detection.

The rest of this paper is structured as follows. Section II describes APT attacks in the IIoT. Section III describes the deep learning method for proactive APT detection in the IIoT. The experimental analysis is presented in Section IV. Finally, Section V summarizes the paper.

II. Description of APT attacks in the IIoT

In this section, we briefly describe APT attacks in the IIoT; an entire life cycle of an APT attack in the IIoT is shown in Fig. 1. The process of attack can be divided into the following five stages: data collection, plan formulation, privilege escalation and internal penetration, authority maintenance and anti-tracing, and goal realization and trace cleanup. We describe each stage as follows:

(1) The first stage is data collection, in which information is collected on the IIoT network by scanning and detecting all networks, including the enterprise network, control network and field network. Data collection runsthrough the entire attack. The purpose is to obtain information such as asset information, user information, supplier information, mailboxes and account numbers in the target network. All these actions serve as preparation for formulating an effective attack plan.

(2) The second stage is plan formulation. After the basic information and topological structure of the IIoT network are mastered through the collected information, the hacker obtains access to the IIoT network. An effective attack plan is formulated using e-mail, instant messaging, social networks or application weaknesses to attack devices or servers in the IIoT network as a preliminary entry to the attack.

(3) The third stage is privilege escalation and internal penetration, which is the core goal of IIoT attacks. Hackers first use software or management vulnerabilities such as www, e-mail or business services in the enterprise network of the IIoT to implant malicious code, such as recorders, Trojan horses, password cracking and file collection programs. As long as this operation is successful, hackers can lurk in the enterprise network to collect increasingly complete network information. It later becomes convenient to use horizontal penetration to find the configuration file and business interface of Terminal2 in the enterprise network and thereby enter the control network of the IIoT. Hackers will crack the domain control server as the focus of the attack. Once the domain control server's authority is obtained, the hacker can smoothly enter the IIoT field network and further control the production equipment.

(4) The fourth stage is authority maintenance and anti-tracing. The process of stealing enterprise, control, or field network data in the IIoT is complicated and cumbersome. To maintain the obtained permissions, hackers develop feasible techniques to prevent traceability and

remain deeply hidden, such as using a homemade anti-killing RAT cluster system or Cobalt Strike. When transmitting data, methods such as byte splitting and combining can also be used to transmit the collected information slowly to avoid detection by abnormal traffic analysis tools. As mentioned earlier, the use of forged identities and springboards [11] will also prevent the traceability of the entire APT attack even when the attack is discovered.

(5) The final stage is goal realization and trace cleanup, that is, the illegal transmission of sensitive data from the enterprise, control and field networks to an external system controlled by the hackers. The hacker leverages data mining or related algorithms to locate the root user's computer from the collected network traffic information and log files of the target system. After obtaining the root user's information and root permissions, the hackers can access the data center or issue instructions to other machines. The core data inside the data center are then transmitted back to the attacker through encrypted channels. Finally, the access traces, logs, and other related information are cleaned up.

It should be noted that APT attacks have unique features compared with traditional network attacks.

Feature 1: APT attacks on the IIoT utilize the power of a government or even an entire country. Regardless of how much manpower and resources are involved, such attacks do not stop until they reach their target. This creates high pressure to prevent attacks.

Feature 2: During APT attacks on the IIoT, the hackers are silent and latent for a long time at every stage. This means that existing intrusion detection and antivirus software are unable to identify APT attacks by means of data association analysis and anomaly detection.

III. Deep learning for proactive APT detection in the IIoT

A deep learning method for proactive APT detection in the IIoT is explored in this section. Anomaly detection based on unsupervised learning is very popular for the following reasons. First, in the IIoT, some normal and abnormal data have no clear boundaries. Second, the collected IIoT data also contain noise, which is difficult to distinguish from anomalies. Third, as time passes, normal behavior may change. Finally, labeled data are difficult to obtain. Popular unsupervised learning methods include statistical-based anomaly detection, density-based anomaly detection, cluster-based anomaly detection, anomaly detection using OneClassSVM, and anomaly detection using the isolation forest integrated learning method. In actual projects, if there is relatively little prelabeled data, unsupervised methods can be used. However, since the labeled data in this article come from a private power grid, this article uses a supervised learning method.

A. APT attack sequence analysis

As indicated by the above analysis, APT attacks in the IIoT are characterized by long-term sustainability and can range from a few minutes to several years. Therefore, the APT attack sequence is an unknown sequence with variable length. To complete the attack, the IIoT APT attacker must combine traditional network attacks with various advanced attack methods. The related activities underlying APT attacks are targeted and continuous, but they may be indirect, such as information collection, plan formulation, privilege escalation, etc. Based on an analysis of existing APT events for the IIoT, the behavioral data of the IIoT can be classified into five categories: normal data, monitoring and detection activities, privilege escalation, command operation, and attack and steal.

(1) Normal data: In the enterprise, control and field networks, the industrial control site and machine status, etc., are in the APT attack state or have entered the APT attack state. In the intermittent or incubation attack period, the network can carry out normal real-time network communication and obtain requests through the browser. For example, it can fetch HTTP request messages with safe and reliable data.

(2) Monitoring and detection activities: The attacker needs to collect a large amount of information about the target system (including the port communication status of the host in the enterprise, control, and field networks and the host's network status, data flow, network information transmission status, etc.). Port scanning, code analysis, SQL statement detection and other methods are used to obtain useful data. At this stage, we identify these activities through audit system log analysis and identify and record such data through network traffic analysis, defense system alarms, etc.

(3) Privilege escalation: Privilege escalation includes horizontal and vertical privilege escalation. Horizontal privilege escalation means that the attacker has obtained the access permission of a certain user, but the information obtained is not sufficient. At this time, the attacker will try to exploit various analyzed vulnerabilities and obtain other users' information through precise phishing. With vertical privilege escalation, the attacker expands the single user's privileges to administrator privileges to fully control the system.

(4) Command operation: Whether it is the field network, the control network or the enterprise network, the attacker performs related operations on files. Since the server system may be based on Linux, the directories and programs in the host, data, services, devices, drivers, and I/O files have the read, write, and execute permissions of the corresponding owner. Execution of these commands by the attacker involves illegal access and illegal operations on remote machines, and this kind of command operation data can be recorded.

(5) Attack and steal: Attackers use multiple attack methods, such as spear attacks, watering hole attacks, DoS attacks, flood attacks, WinNuke attacks, Land attacks, Script/ActiveX attacks, Smurf attacks, and routing protocol attacks. Attackers can perform corresponding attacks on the target host or machine in the IIoT, and those data can also be recorded.

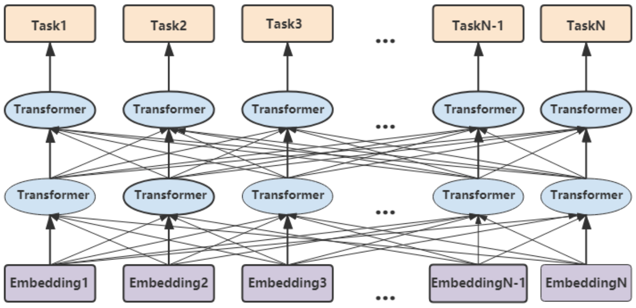

Fig. 2: BERT model structure

B. APT attack word vector generation

A "word vector" is a representation of words as vectors, usually in terms of multidimensional continuous floating-point numbers, where similar words are mapped to similar locations in geometric space. Representing words as vectors allows them to be applied to mathematical operations. An APT is an attack that remains undetected over a long period. The collected APT attack data in the IIoT need to be vectorized to be used as input for the classification model in order to distinguish the attack type. Across the various stages of the entire APT network attack process, all network data packets constitute a time series. Since there may be a certain correlation or logical relationship among the five attack intents, the five attack intent labels are converted into corresponding word vectors. A traditional discrete representation cannot show this relationship, so a distributed representation is used to transform words into a distributed word vector representation. This distributed representation can depict the interrelationships between APT attack intentions in the IIoT. In traditional text information, the most representative semantic feature words are usually selected for word vector representation. This is generally done using the word2vec model [12], which can transform each feature word into the same shape. However, when the word2vec model is used for word vector representation, the APT attack feature words cannot be distinguished through contextual semantic information, and different attack intentions may represent the same vector, which causes subsequent classifier misjudgments. For these reasons, this paper uses the BERT model [13] to represent the APT attack intention word vector. We choose BERT because of the long persistence of APT attacks. If the attack time span is very long, it is difficult to detect the complete attack chain based on real-time point-in-time detection technology. The BERT used in this article contains a transformer structure that relies on the attention mechanism to model the global dependency of inputs and outputs. It can capture the key features of APT attacks in an unknown time series. Moreover, this method can combine the context and semantic information from the APT attack sequence and can be more reasonably expressed as a word vector for judgment.

C. BERT pretraining language model

In recent years, researchers have achieved good results in pretraining language models [14], e.g., the embeddings from language model (ELMo) and generative pretraining (GPT). In 2018, Google Research Institute Devlin and others proposed the BERT pretraining language model.

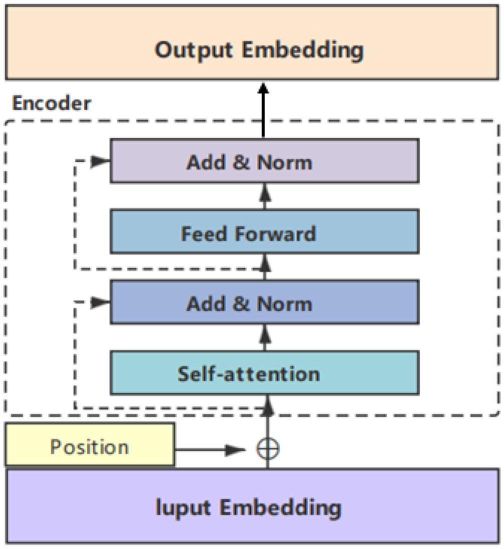

Fig. 3: Transformer encoder structure

As shown in Fig. 2, the BERT pretraining model uses a bidirectional transformer as the encoder; this transformer is based on the attention mechanism to model text, which has good parallel computing capabilities. Originally, it was proposed that the masked language model and sentence continuity prediction be used for joint training. The aforementioned BERT model can capture the key characteristics in an unknown time series of varying length for the APT attack on the IIoT. We use the embedding output of the processed APT attack sequences as the input word representation of the BERT internal transformer coding network, combined with a series of transformer encoders. In the BERT model, following application of the multilayer bidirectional transformer encoder, the vectorized representation tasks of the APT attack word are finally obtained. The transformer is a Seq2Seq model [15] based on self-attention that has an encoder-decoder structure. The encoder adds a variable length sequence to a fixed length sequence, and the decoder decodes this fixed length vector into an output sequence of variable length. The BERT model mainly uses the encoder part of the transformer, which can encode the variable length time series in the APT attack. The structure of the encoder is shown in Fig. 3.

The input of the encoder is a word embedding representation, where the position information is added. The input then passes through the self-attention layer, which helps the encoder view the information of the words before and after as it encodes each character. Its output passes through an add & norm layer. The add layer adds the input and output of the self-attention layer, followed by normalization by the norm layer. The normalized vector list will be passed into a feed forward layer, which is a fully connected neural layer. The feed forward layer follows a corresponding add & norm layer, which outputs a new list of normalized word vectors. The core idea behind determining the most important input using self-attention in an encoder is to calculate the relationship between a text message and other text messages in a sequence and then use this relationship to adjust the weight of each text to obtain a new expression. This new expression contains not only its own semantics but also the relationship between them. Therefore, compared with the traditional word vector, it provides a more global expression, and the APT attack sequence of the IIoT can better reasonably express the semantic information of the context in the temporal feature.

D. Attack sequence optimization in the IIoT

In the APT attack, each attack has steps, each step has key "word vectors", and the word vectors between the steps are correlated. We use the word vector association relationship to optimize the attack sequence data. Some operations of APT attacks that do not damage the host can be considered as normal data and are not counted in the APT attack sequence samples. Only the word vector association relationship of the before and after operations that satisfy the attack features is used as the input of the APT detection model. In other words, the hacker performs some operations in some steps, but as long as this action does not extract the word vector that can cause damage, we still regard this operation of the hacker as normal data that will not be used as a step of the attack. In this way, we avoid considering all operations as attacks as soon as they are regarded as hackers and thus optimize the attack sequence samples.

After optimization, the length of the APT attack sequence can be greatly reduced, which not only ensures the feature integrity of the APT attack sequence but also simplifies the APT attack sequence, significantly mitigating the cost of training the model.

E. APT attack detection algorithm based on BERT

This paper proposes an APT attack detection algorithm based on BERT in the IIoT. The detailed algorithm process is as follows:

Input: APT attack sequence training set in the IIoT where two variables are included. One is the characteristic information of the APT attack. The other is the APT category to which the attack belongs.

Output: APT attack intention classification model.

- **Step 1**: First, a clean data set is obtained by cleaning and preprocessing the collected data. For data cleaning, we mainly delete the null attribute values in the APT attack sample. For data preprocessing, we numericalize character attributes of APT attacks, perform one-hot encoding on multiple values of the attribute, and concatenate the result with the original attribute. After numerical and one-hot encoding processing, because the numerical range of each attribute is different, using the original value directly will affect the shift in the model's focus, so the features in the data set are standardized and normalized.
- **Step 2**: The preprocessed APT attack sequence data are sent to the BERT model, where the timing feature starts with [CLASS (CLS)], and the feature and tag category are separated by [SEPARATING (SEP)] notation.

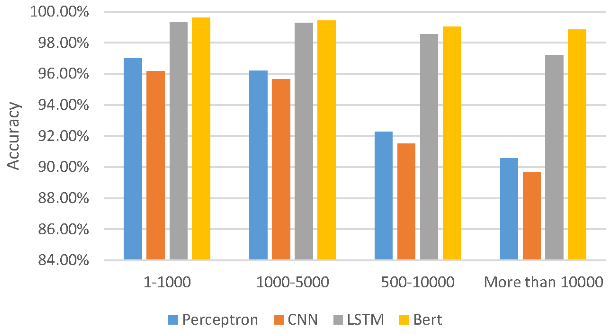

Fig. 4: Sequence model accuracy rate comparison histogram

- **Step 3**: In the BERT pretraining model, the data are encoded by the encoder of the bidirectional transformer. The characteristics corresponding to the APT attack sequence are denoted.
- **Step 4**: The characteristic representation obtained in step 3 is input into the Softmax regression classifier model for classification training, where the dimension of the word vector is set to 5 to obtain the probability that the APT attack intention belongs to each APT attack sequence.
- **Step 5**: The data in the training set are trained in batches and output to the classification model of APT attack intentions.
- **Step 6**: The obtained classification model is used on the test set to test the generalization ability of the model and obtain various performance indicators of the algorithm.

IV. Experimental Analysis

A. Experimental environment and settings

The experiment in this paper is performed using the TensorFlow deep learning framework. The experimental hardware environment is a PC with a Windows 10 operating system and an NVIDIA GTX 1080TI GPU. We take the data collected from a certain equipment manufacturer as the experimental training and test data sets and divide the simulated attacks into five categories: 'NORMAL' (i.e. normal network connections), 'PROBE' (i.e. ipsweep, nmap, portsweep, satan), 'DOS' (i.e. back, land, neptune, pod, smurf, teardrop), 'U2R' (i.e. buffer_overflow, loadmodule, perl, rootkit), 'R2L' (i.e. buffer_overflow, loadmodule, perl, rootkit).

B. Experimental scheme

To evaluate the effectiveness of the scheme proposed in this paper (BERT), when designing the comparison schemes, we choose a single-layer perceptron model (hereafter, Perceptron), a single-layer LSTM network (hereafter, LSTM) and a single-layer CNN network (hereafter, CNN). The reason for selecting this network model is that the training set consists of only 5-word vectors; because the length of the word vectors is small, fewer parameters are required. The detailed experimental plan is as follows:

(1) APT attack sequence detection accuracy: The data are divided into four levels according to the sequence length,

i.e., 1-1000, 1000-5000, 5000-10000 and more than 10000. The 10-fold cross-validation method is used to obtain the APT attack sequence detection accuracy rate.

(2) Receiver operating characteristic (ROC) curve: The steps are the same as those of (1). The four trained models are predicted on the test set, and the ROC curves are obtained for model performance comparison.

C. Experimental results and analysis

The APT detection accuracy rates for the four models with different test set sequence lengths are shown in Table I. From Table I, we can also obtain the histogram of the APT detection accuracy for the four models, as shown in Fig. 4.

TABLE I: Accuracy comparison of the sequence models

Sequence length	Perceptron	CNN	LSTM	BERT
1-1000	97.00%	96.17%	99.31%	99.62%
1000-5000	96.20%	95.67%	99.26%	99.44%
5000-10000	92.26%	91.52%	98.56%	99.04%
More than 10000	90.58%	89.66%	97.21%	98.85%

Combining TABLE I and Fig. 4 reveals that when the unknown attack sequence is short, such as within 1 to 5000, all four models can detect unknown APT attacks with a correct rate of over 95.67%; this success is due to the shorter attack. The features of the sequence can be well extracted by the four models for discrimination. Among them, the BERT model has the highest accuracy, reaching more than 99.42%, and LSTM is second, indicating that our model can effectively detect an APT attack when the latency is not long. As the length of the attack sequence increases, the CNN and Perceptron models lose their effectiveness in detecting long-term latency APT attacks. This is because such long unknown sequences cannot effectively extract the characteristics of the attack using a single-layer network or CNN network structure. The LSTM and BERT models can also obtain good results because they have memory for longer sequence data features; in particular, the self-attention of the transformer contained in BERT can properly note the APT attack, allowing it to achieve the best detection effects. In short, the BERT model has better performance than other models. We use the four trained models for the test set and obtain the ROC curves, which are shown in Fig. 5. The figure clearly shows that the degree to which the ROC curves deviate from the 45-degree diagonal is almost the same for all four models. The accuracy area under the ROC curve is slightly larger for the proposed method than for the other three models. Hence, we conclude that our approach delivers better performance.

Taken together, the comparative analysis results of (1) and (2) demonstrate that the method in this paper can effectively detect an APT attack sequence with a long attack duration in the IIoT and show the feasibility and effectiveness of this method.

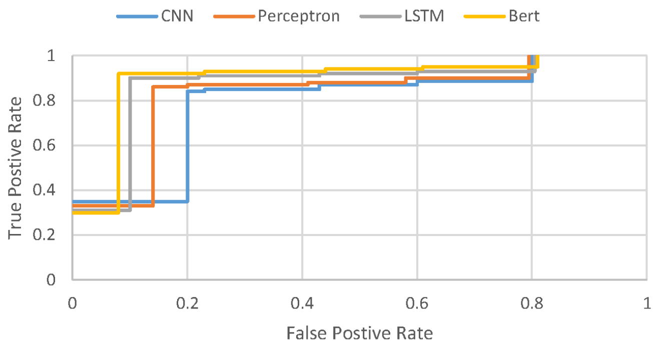

Fig. 5: ROC curve comparison chart

V. Conclusion

IIoT APT detection has gradually become a research hotspot in both industry and academia, and many new technologies, algorithms and systems related to APT detection have recently emerged. According to the analysis of the 2016 Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) Industrial Internet Security Situation Report, more than 80% of the country's critical infrastructure relies on the Industrial Internet to automate the production process. However, APT detection in the existing Industrial Internet has many drawbacks, such as a low accuracy rate and a high false alarm rate. This paper proposes a deep learning-based proactive APT detection scheme for the IIoT. The experimental results show that the proposed method not only has feasibility and effectiveness in detecting APTs but also has an accuracy rate as high as 99%. Our future work will optimize this model and promote this technology in the IIoT.

References

[1] K. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-enhanced data sharing with traceable and direct revocation in iiot," *IEEE Transactions on Industrial Informatics*, vol. 17, 2021.

[2] A. Al-Dulaimi, A. Anpalagan, S. Al-Rubaye, and Q. Ni, "Adaptive management of cognitive radio networks employing femtocells," *IEEE Systems Journal*, vol. 11, no. 4, pp. 2687–2698, 2017.

[3] C. Zhu, J. J. P. C. Rodrigues, V. C. M. Leung, L. Shu, and L. T. Yang, "Trust-based communication for the industrial internet of things," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 16–22, 2018.

[4] A. Alshamrani, S. Myneni, A. Chowdhary, and D. Huang, "A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities," *IEEE Communications Surveys Tutorials*, vol. 21, no. 2, pp. 1851–1877, 2019.

[5] Y. Li, W. Dai, J. Bai, X. Gan, J. Wang, and X. Wang, "An intelligence-driven security-aware defense mechanism for advanced persistent threats," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 646–661, 2019.

[6] P. Hu, H. Li, H. Fu, D. Cansever, and P. Mohapatra, "Dynamic defense strategy against advanced persistent threat with insiders," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 747–755.

[7] A. Canovas, J. M. Jimenez, O. Romero, and J. Lloret, "Multimedia data flow traffic classification using intelligent models based on traffic patterns," *IEEE Network*, vol. 32, no. 6, pp. 100–107, 2018.

[8] L. Xiao, D. Xu, N. B. Mandayam, and H. V. Poor, "Attacker-centric view of a detection game against advanced persistent threats," *IEEE Transactions on Mobile Computing*, vol. 17, no. 11, pp. 2512–2523, 2018.

[9] J. Zhang, K. Yu, Z. Wen, X. Qi, and A. K. Paul, "3d reconstruction for motion blurred images using deep learning-based intelligent systems," *Computers, Materials & Continua*, vol. 66, no. 2, pp. 2087–2104, 2021.

[10] K. Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5g-enabled intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, 2021.

[11] A. Lemay, J. Calvet, F. Menet, and J. M. Fernandez, "Survey of publicly available reports on advanced persistent threat actors," *Computers & Security*, vol. 72, pp. 26–59, 2018.

[12] S. Ji, N. Satish, S. Li, and P. K. Dubey, "Parallelizing word2vec in shared and distributed memory," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 9, pp. 2090–2100, 2019.

[13] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 165–178, 2020.

[14] G. K. W. Huang and J. C. Lee, "Hyperpartisan news and articles detection using bert and elmo," in *2019 International Conference on Computer and Drone Applications (IConDA)*, 2019, pp. 29–32.

[15] Y. Li, M. Zhu, Y. Ma, and J. Yang, "Work modes recognition and boundary identification of mfr pulse sequences with a hierarchical seq2seq lstm," *IET Radar, Sonar Navigation*, vol. 14, no. 9, pp. 1343–1353, 2020.

Keping Yu (S'11-M'17) received his Ph.D. from Waseda University, Japan, in 2016. He is currently an Assistant Professor at Waseda University, Japan and a Visiting Professor with the College of Computer Science, Sichuan Normal University, China. His research interests include smart grids, information-centric networking, the Internet of Things, artificial intelligence, blockchain, and information security.

Liang Tan received his Ph.D. from the University of Electronic Science and Technology of China, China, in 2007. He is currently a professor at Sichuan Normal University. His research interests include network security, trusted computing, big data and cloud computing.

Shahid Mumtaz received his Ph.D. from the University of Aveiro, Portugal. He is currently a senior research engineer with the Instituto de Telecomunicações, Aveiro, where he is involved in EU-funded projects. His research interests include co-operative techniques, MIMO techniques, cognitive radios, multihop relaying communication and game theory.

Saba Al-Rubaye received her Ph.D. in electrical and electronic engineering from Brunel University London, United Kingdom. She is currently a senior lecturer and leading connectivity and networking researcher at the School of Aerospace, Transport and Manufacturing at Cranfield University, United Kingdom. Her research interests lie primarily in the areas of UAV connectivity, networking, AI and autonomous vehicle systems.

Anwer Al-Dulaimi is a Technical Product Owner in the Center of Excellence at EXFO Inc., Montreal, Canada. He received his Ph.D. in electronic and computer engineering from Brunel University, London, United Kingdom, in 2012 after receiving a BSc and MSc honors degree in communication engineering. His research interests include 5G and beyond networks, cloud computing, V2X and the Internet of Things.

Ali Kashif Bashir is a senior lecturer at the Department of Computing and Mathematics, Manchester Metropolitan University, United Kingdom. He is a senior member of IEEE and a Distinguished Speaker of ACM.

Farrukh Aslam Khan is a Professor at the Center of Excellence in Information Assurance (CoEIA), King Saud University, Saudi Arabia. His research interests include cybersecurity, the Internet of Things, E-health, and computational intelligence. He received his Ph.D. in Computer Engineering from Jeju National University, South Korea. He is a Fellow of the British Computer Society and a Senior Member of IEEE.

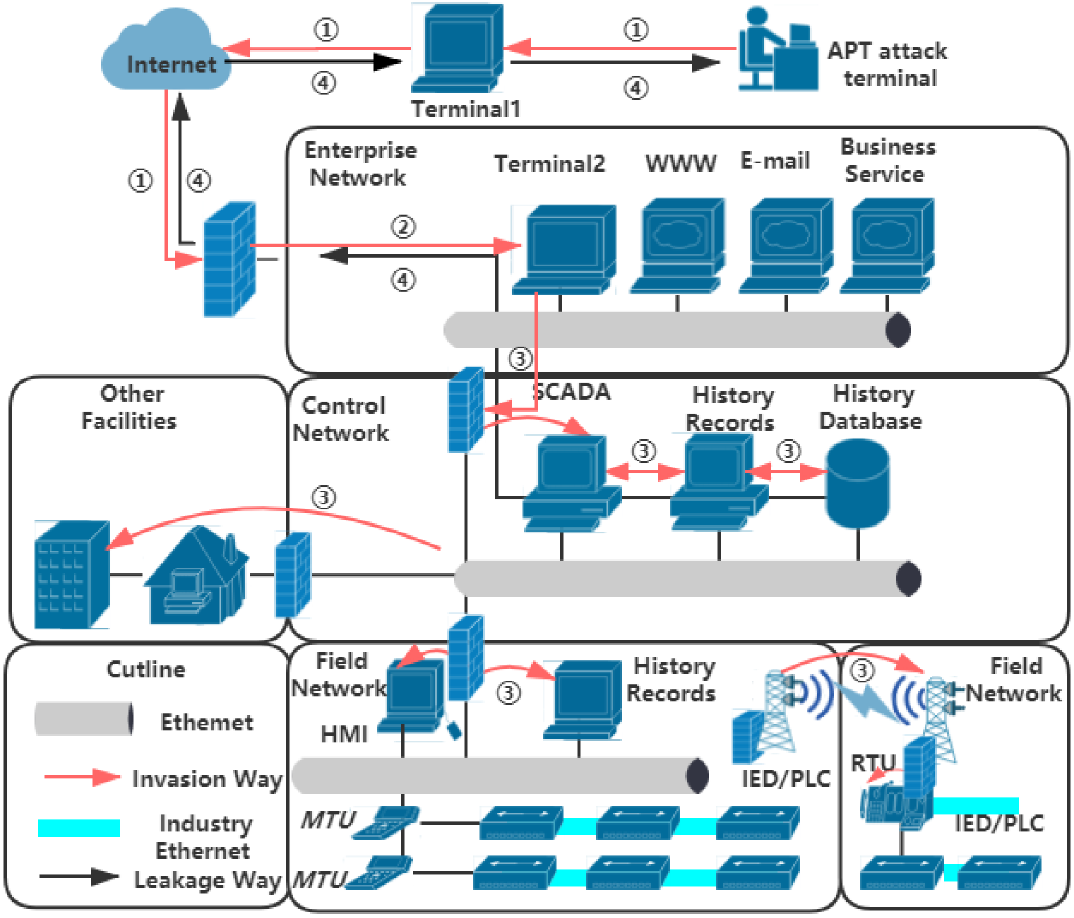

Fig. 1: IIoT-APT attack structure

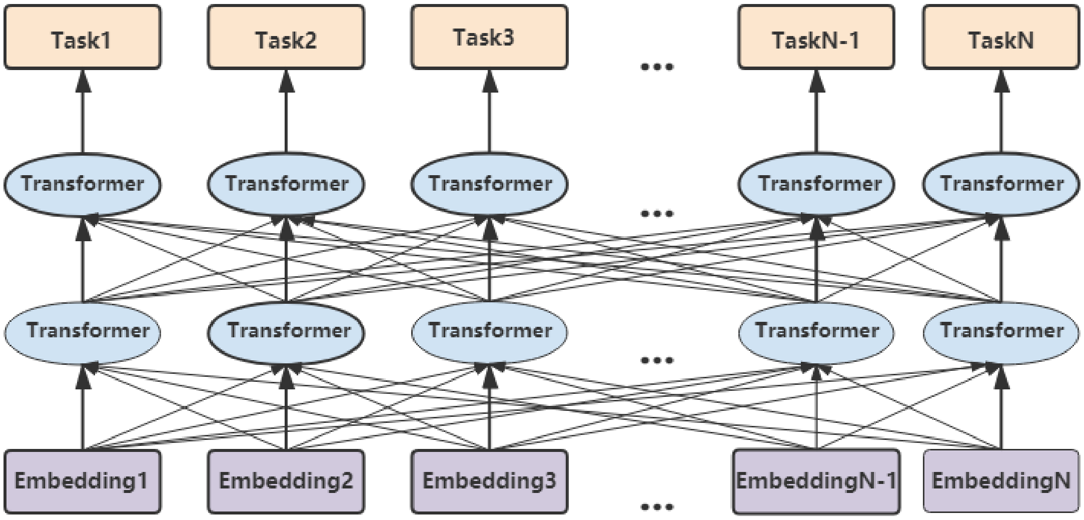

Fig. 2: BERT model structure

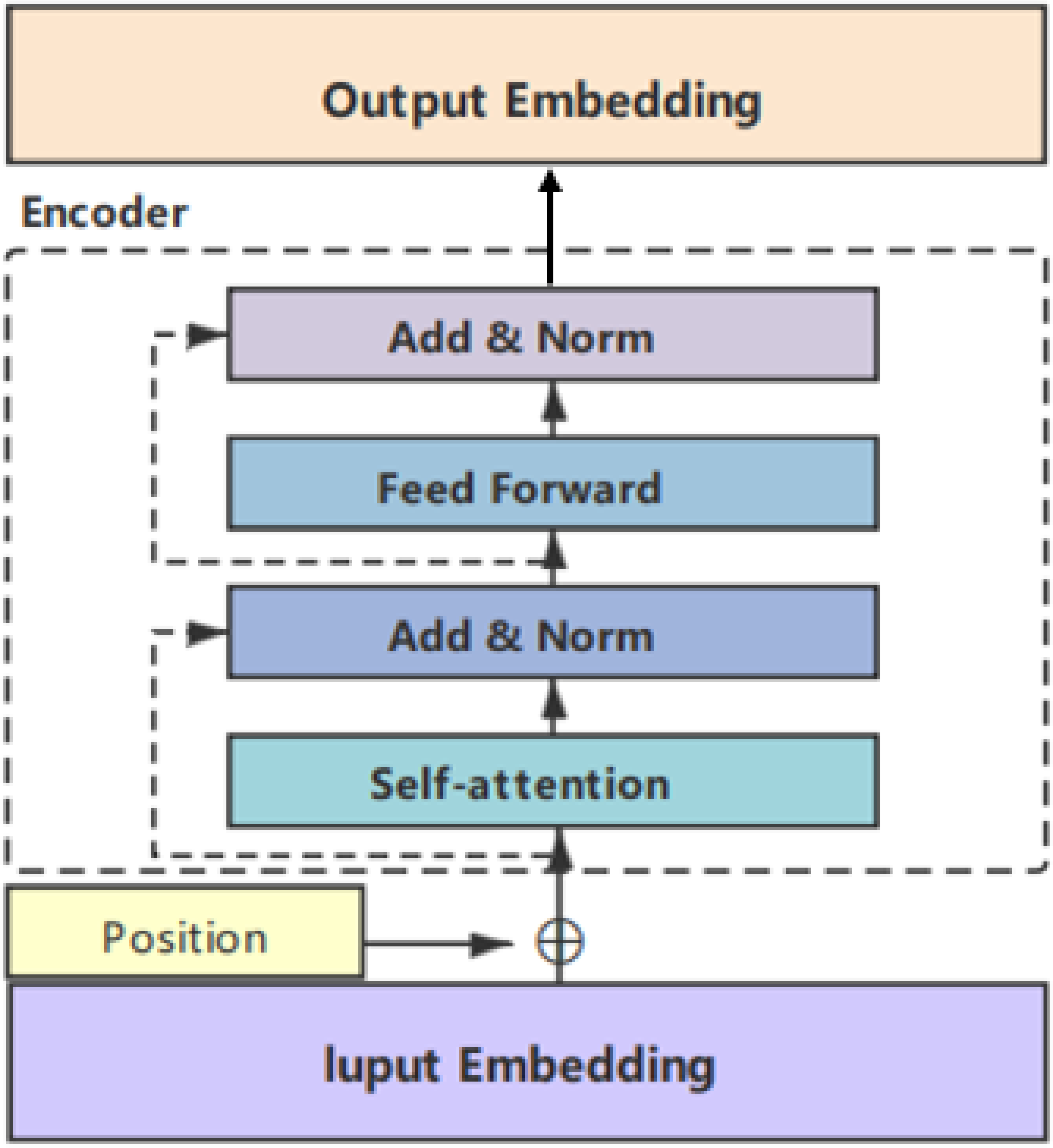

Fig. 3: Transformer encoder structure

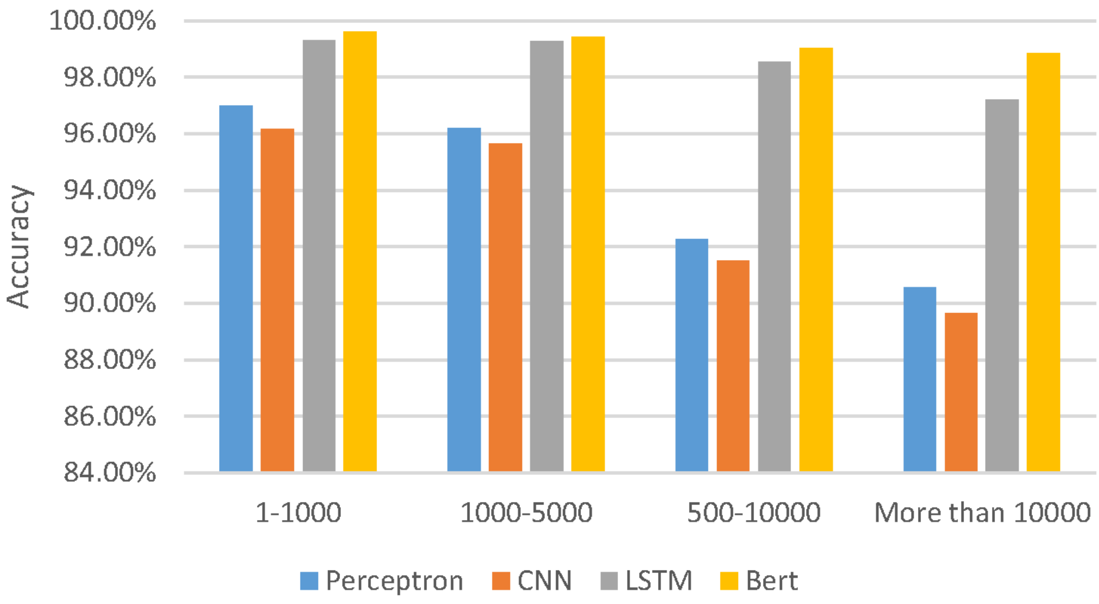

Fig. 4: Sequence model accuracy rate comparison histogram

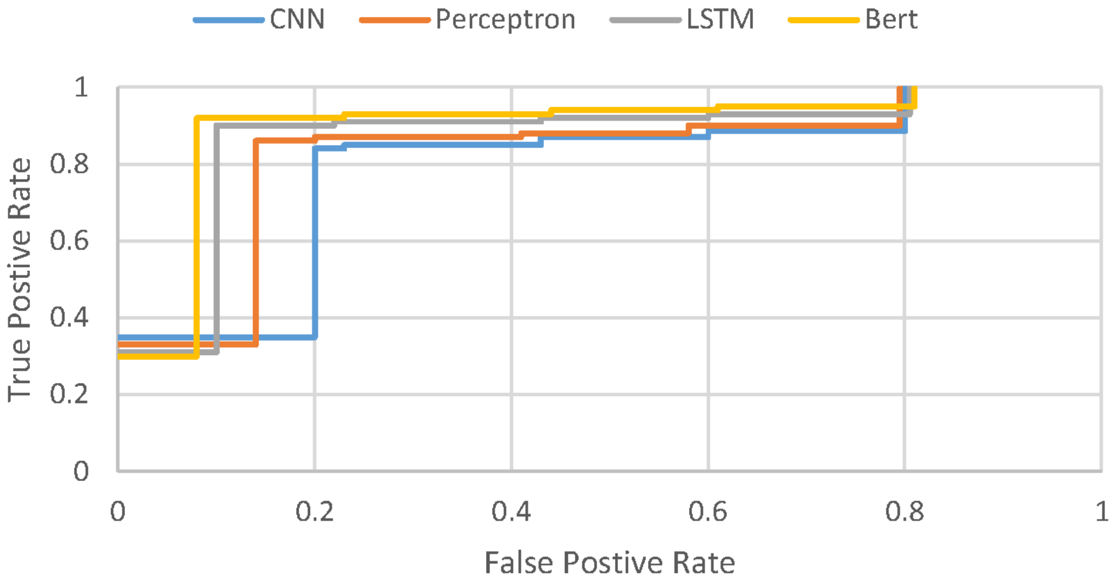

Fig. 5: ROC curve comparison chart

TABLE I: Accuracy comparison of the sequence models

Sequence length	Perceptron	CNN	LSTM	BERT
1-1000	97.00%	96.17%	99.31%	99.62%
1000-5000	96.20%	95.67%	99.26%	99.44%
5000-10000	92.26%	91.52%	98.56%	99.04%
More than 10000	90.58%	89.66%	97.21%	98.85%