Universal Adversarial Perturbations Against Machine-Learning-Based Intrusion Detection Systems in Industrial Internet of Things

Sicong Zhang, Yang Xu, and Xiaoyao Xie, *Member, IEEE*

***Abstract*—The security of the Industrial Internet of Things (IIoT) has emerged as a prominent concern in cyber-security due to the potential impact of attacks against IIoT on physical infrastructure. Machine learning (ML)-based intrusion detection systems (IDSs) recently have been demonstrated to be an effective tool for protecting the systems in IIoT. However, the vulnerability of ML-based IDSs to adversarial attacks hinders their further application in IIoT. This article aims to further explore the adversarial attacks in IIoT to better evaluate the security of ML-based IDSs in this area. Our research primarily focuses on the generation of universal adversarial perturbations in IIoT, a topic that received limited attention in previous literature. Two novel attack methods based on a unified framework are proposed to utilize the original input-dependent adversarial perturbations of gradient-based or optimization-based adversarial attack methods to craft universal adversarial perturbations with better performance and transferability. The proposed attack methods conceal the underlying implementation details of the target attack methods, exploiting the original adversarial perturbations in a closed-box manner. This enhances their flexibility, making them applicable in a wider range of scenarios and enabling them to be combined with most gradient-based or optimization-based attack methods. Comprehensive experiments are conducted on three mainstream intrusion detection data sets, i.e., NSL-KDD, Gas Pipeline, and edge-IIoTset, to validate the effectiveness of the proposed methods. The preliminary experimental results demonstrate the feasibility of universal adversarial perturbations in IIoT and the superiority of the proposed methods to state-of-the-art attack methods.**

***Index Terms*—Adversarial examples (AEs), industrial control systems (ICSs), Industrial Internet of Things (IIoT), intrusion detection, universal adversarial perturbations.**

I. Introduction

THE INDUSTRIAL Internet of Things is a very important infrastructure, nowadays. The combination of Internet technology and industrial control systems (ICSs) improves the efficiency and intelligence of industrial production [1]. Besides, the integration of the industrial control process with the enterprise's internal network builds up the management efficiency of industrial production, whereas the fusion of ICS and Internet technology also enables attacks against the physical world. The Industrial Internet of Things (IIoT) technology is utilized in the field of smart grids, smart manufacturing, smart healthcare, etc. [2]. The systems in these areas are generally critical infrastructures for the national welfare and the people's livelihood. Therefore, the systems in these domains are normally security-critical, against which malicious behaviors could cause damage to the national economy or even threaten national security. Consequently, the security of IIoT recently attracted more and more attention in academia and industry.

Intrusion detection is regarded as a very effective way to protect networks and computer systems. Lots of researchers have employed intrusion detection to recognize malicious behaviors in IIoT [3], [4], [5], [6], [7], [8], in which machine learning (ML) is proven to be effective enough to identify unknown attacks. Consequently, ML is widely and deeply utilized to improve the performance of intrusion detection systems (IDSs) in IIoT. However, recent works show that ML is vulnerable to adversarial examples (AEs) produced by adding deliberately crafted perturbations to original inputs [9]. The existence of AEs unmasks the weakness of ML algorithms, which threatens the IDS based on ML technologies. The skills of generating AEs can be used to disguise malicious network traffic as normal ones to evade the detection of the target ML-based IDSs, incapacitating the target IDSs. This factor significantly impacts the performance and reliability of ML-based IDSs. In the real world, IDSs are required to recognize as much potentially malicious traffic as possible. The evasion of malicious traffic could cause unpredictable losses to the protected systems. Consequently, the AEs hinder the further application of ML in intrusion detection for IIoT. As a security-critical domain, further research on AEs in intrusion detection for IIoT is urgent and necessary to promote the security application of ML in this area and better secure the security of IIoT. Although some existing works have already researched the AEs in IIoT [10], [11], [12], [13], [14], there is still a lack of further research on AEs in this area. Currently, the origin of AEs is still not clear and the influences of AEs on ML-based security systems in IIoT need further exploration. Hence, to better evaluate the robustness of ML-based IDSs in IIoT against AEs, it is imperative to design more efficacious and targeted methodologies.

Received 6 April 2024; accepted 16 September 2024. Date of publication 23 September 2024; date of current version 9 January 2025. This work was supported by the Science and Technology Planned Project of Guizhou Province, China, under Grant [2023]YB449. *(Corresponding author: Yang Xu.)*

The authors are with the School of Cyber Science and Technology, the Key Laboratory of Information and Computing Science Guizhou Province, Guizhou Normal University, Guiyang 550001, China (e-mail: 202103008@gznu.edu.cn; xy@gznu.edu.cn; xyx@gznu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2024.3465549

Currently, the research on AEs in intrusion detection mainly concentrates on the traditional Internet [15], [16], [17], [18], [19]. The feature of frequent interaction with the physical world makes IIoT different from the traditional Internet. The AEs against ML-based IDSs in IIoT are less discussed in the existing works. On the other hand, the existing research mostly studies the generation of input-dependent adversarial perturbations. This kind of adversarial perturbation is specific to the original inputs in a one-to-one way, which means that one original input corresponds to one AE. On the contrary, universal adversarial perturbations are input-independent. The universal adversarial perturbations are generated on a small set of samples and can be added to other inputs to mislead the target classifiers. Compared with input-dependent adversarial perturbations, the generation of universal ones can be regarded as the one-to-many scheme, which means that one production of the universal adversarial perturbations can be applied to many different original inputs to mislead the target classifiers. The universal adversarial perturbations are proved to be generalized well across different classifiers [20]. Accordingly, universal adversarial perturbations enable adversaries to attack the target classifiers without knowing their internal details. This poses more threats to ML-based systems in various domains. Consequently, to further promote the exploration of the cause of AEs and better evaluate the ML-based IDS in IIoT, research on generating universal adversarial perturbations against ML-based IDSs in IIoT is meaningful.

This article explores the generation of adversarial perturbations against ML-based IDSs in IIoT. Specifically, we mainly focus on the production of universal adversarial perturbations [20] that are less discussed in the existing works. In this article, we propose a new attack framework to generate universal adversarial perturbations against ML-based IDSs in IIoT through a unified interface. The proposed framework utilizes the adversarial perturbations produced by the mainstream glass-box adversarial attack methods, i.e., fast gradient sign method (FGSM) [21], projected gradient descent (PGD) [22], momentum iterative method (MIM) [23], DeepFool (DP) [24], and Carlini and Wagner (CW) [25], to generate universal adversarial perturbations. The proposed framework neglects the underlying details of the mainstream attack methods and leverages the target attack methods to produce universal adversarial perturbations in a closed-box way, which makes the proposed attack framework more flexible. Based on the proposed framework, two different adversarial attack methods, i.e., the universal adversarial perturbation generator for IIoT (UAPG-IIoT) and the ensemble UAPG-IIoT (EUAPG-IIoT), are designed to craft universal adversarial perturbations with better attack performance and transferability. The proposed attack methods are more effective because they build upon the existing results of the target attack methods. The two attack methods based on the same framework share a similar generation procedure but have different generation details. Specifically, the EUAPG-IIoT produces universal adversarial perturbations with better performance based on UAPG-IIoT. The ML algorithms that are widely utilized in the existing work [10], [15], [17], [18], [19], i.e., neural networks (NNs), support vector machines (SVMs), logistic regression (LR), decision trees (DTs), naive Bayes (NB), and random forest (RF), are adopted to build the target IDSs to validate the attack performance and transferability of the proposed methods.

In general, we make the following contributions in this article.

1) To better evaluate the robustness of ML-based IDSs in IIoT, a new attack framework is proposed to craft universal adversarial perturbations with better attack performance and transferability. Two adversarial attack methods are designed based on the proposed attack framework, i.e., UAPG-IIoT and EUAPG-IIoT. The two attack methods utilize the AEs produced by the classic glass-box adversarial attack methods to synthesize universal adversarial perturbations that are input-independent. The produced universal adversarial perturbations can be added to original inputs that are not involved in the generation of the perturbations to evade the detection of the target IDS in IIoT. The difference between the two attack methods is that the UAPG-IIoT crafts the universal adversarial perturbations on a specific target classifier using a small data set but the EUAPG-IIoT generates the universal adversarial perturbations on a set of target classifiers. The preliminary experimental results show that the universal adversarial perturbations produced by our attack methods show better attack performance and better transferability across heterogeneous ML-based IDSs compared with the classic attack methods.
2) The research on AEs against the IDS in IIoT is in its infancy. In this article, the widely used ML algorithms in the domain of intrusion detection for IIoT are adopted to build the target IDS, through which we further evaluate the robustness of the widely utilized ML classifiers in IIoT against adversarial attacks. Specifically, we assess the robustness of the ML-based IDS in IIoT against the universal adversarial perturbations. The transferability is very important for the adversarial attack methods because good transferability of AEs means that the corresponding attack methods can attack the target classifiers in a closed-box way more effectively when the internal information of the target classifiers is unknown [26], which makes the attack methods applicable in more scenarios. Therefore, the transferability of the universal adversarial perturbations across different ML classifiers is also evaluated. The preliminary results show that the universal adversarial perturbations produced by our attack methods show better transferability across different ML-based IDSs compared with the classic adversarial attack methods.
3) Unlike adversarial attacks in computer vision, adversarial attacks in cyber-security need to guarantee the validity of original inputs, which means that the number of modifiable features of the original input is normally constrained [15]. Consequently, the influence of the number of modifiable features in the original network traffic on the attack performance is also assessed within the scope of this study. The preliminary experimental results show that the attack performance of the proposed

attack methods is superior to the baseline attack methods even when modifiable features are restricted.

The remainder of this article is organized as follows: in Section II, the related works on intrusion detection for IIoT and adversarial attacks in IIoT are reviewed. The research background of this article is discussed in Section III, including the background knowledge of IIoT, adversarial attacks, and the details of data sets, employed in the experiment section. The details of the proposed attack methods are presented in Section IV, followed by a comprehensive experimental evaluation and analysis in Section V. The conclusion of our work and the future research directions are presented in Section VI.

II. Related Work

With the advent of Industry 4.0, the IIoT is assuming an increasingly pivotal role. The corresponding technology promotes the intelligence and efficiency of industrial production. The IIoT has penetrated into national critical infrastructures, such as the power grid, manufacturing, water treatment, etc. Unlike the traditional Internet, the security of IIoT can influence the physical world. Accordingly, the attacks against IIoT in these security-critical areas can cause serious security problems, such as property damage or personal injury. The security research of IIoT has become the current hotspot in academia and industry.

A. Intrusion Detection for IIoT

In the existing literature, ML-based IDSs have been proven to be effective in recognizing the complex and diverse threats in IIoT. Abdel-Basse et al. [3] proposed a new method called Deep-IFS to improve the detection performance of IDSs in IIoT under the environment of fog. The Deep-IFS combines the local gated recurrent unit and the multihead attention to overcome the key challenge of the long traffic sequence. The residual connection between different layers is leveraged to decrease the information loss. Besides, to better face the big IIoT traffic data, they adopt distributed architecture to accomplish the deployment and training of the proposed model. Telikan et al. [4] utilized a hybrid model of stacked autoencoders and convolutional NNs based on a cost-dependent loss function to mitigate the imbalanced distribution of intrusion data. A fog computing schema is adopted to solve the scalability of IDSs on the big IIoT traffic data. Mansour [5] fuses clustering technology, blockchain, and deep learning to detect malicious traffic in IIoT. The combination of Harris Hawks optimization-based clustering and the gated recurrent unit is adopted to build up the IDSs. The blockchain is used to ensure the security transmission of the data in IIoT. The proposed method is claimed to show better performance than the state-of-the-art IDSs. Li et al. [6] proposed an intrusion detection method based on the fusion of multiple convolutional NNs. The proposed intrusion detection method is claimed to be more robust in the complex environment of IIoT. The 1-D traffic feature data are transformed into grayscale graphs, through which the fusion of multiple convolutional NNs is introduced into intrusion detection to cope with the complex traffic data and diverse attacks in IIoT. Gu et al. [7] proposed a data expansion algorithm to improve the data quality and solve the imbalance of training data, in which attack data is normally insufficient. Based on the data expansion, a reconstructed convolutional NN with the classification activation map structure is designed to learn effective features from the raw traffic and improve the detection performance of IDSs in ICS. Ahmad et al. [8] utilized the deep random NN to construct their IDSs. To improve the performance of the proposed detection method, they introduce hybrid particle swarm optimization with sequential quadratic programming to optimize the parameters of the NN. They conducted extensive experiments on three mainstream IIoT data sets to validate the effectiveness of their method. Their experimental results demonstrate the superiority of their methods. As discussed above, the IDSs have become one of the critical solutions to the increasingly severe situation of cyber-security in IIoT.

B. Adversarial Attacks in IIoT

While the ML-based IDS can effectively recognize the threats in IIoT, the occurrence of AEs hinders the further application of ML technologies in this area. The AEs disclose the underlying weakness of ML algorithms, which makes the ML-based IDS unreliable. Consequently, to better evaluate the robustness of ML-based IDSs in IIoT, researchers have researched the generation of AEs in this area. Chen et al. [10] explored the generation of AEs against ML-based IDSs in ICS. They propose two strategies to overcome the problems confronted during the production of AEs in ICS, such as the legal range of data fields, the discreteness of data fields, the immutability of data fields, etc. One of the proposed strategies is to produce the AEs against ML-based IDSs in ICS using the optimal solution and the other utilizes the generative adversarial networks to craft the AE. They conduct comprehensive experiments on a semi-physical testbed to validate the effectiveness of the proposed attack methods against the detection systems in the physical world. Qiu et al. [11] proposed a closed-box attack method to compromise the deep learning-based network IDSs. They leverage a small amount of training data to extract the target model. Then, a saliency map is adopted to find out the critical traffic attributes that influence the detection results mostly. It is claimed in their experiments that the proposed attack methods can compromise the Kitsune, a state-of-the-art deep learning-based network IDS, by only modifying a small number of bytes in network packets. Esmaeili et al. [12] explored the defense of deep learning-based malware detectors against closed-box adversarial attacks in IIoT. They designed a new defense method named stateful query analysis. Unlike the existing methods that aim to detect an individual AE, the proposed method leverages the history information of queries to identify adversarial scenarios, through which adversarial attacks can be aborted before their completion. Zeng et al. [13] discussed the adversarial attacks against the multiagent reinforcement learning-based demand response management systems in power grids. They propose a robust adversarial multiagent

reinforcement learning framework to improve the resilience of the demand response of the power systems against adversarial attacks. The proposed framework mainly utilizes periodic robust adversarial training to help the target demand response management systems deal with adversarial attacks and enhance their resilience. Gungor et al. [14] focused on the adversarial attacks against predictive maintenance in IIoT. They propose a stacking ensemble learning-based framework to improve the resilience of predictive maintenance against classic glass-box adversarial attacks, i.e., FGSM, basic iterative method, MIM, and robust optimization method. Ten different deep learning models are ensembled by stacking with four different meta-learners, respectively, to achieve the most resilient model against adversarial attacks. They claim that their framework is more resilient than any other individual ML method. Mohammadian et al. [16] used the Jacobian saliency map to craft AEs against deep learning-based network IDSs. Because of the utilization of the Jacobian saliency map, the produced AEs are purported to achieve better performance while fewer features are modified. They evaluate success rates of the best feature sets, average confidence of the adversarial class, and adversarial samples transferability on three wildly used intrusion detection data sets to verify the effectiveness of the proposed method. Alhajjar et al. [17] introduced heuristic optimization into the generation of AEs. They adopt the particle swarm optimization and genetic algorithm to craft AEs against network IDSs. Besides, they also explore the generation of AEs with generative adversarial networks. They compare the performance of their methods with that of the Monte Carlo simulation method on NSL-KDD and UNSW-NB15 data sets. Their initial experimental results demonstrate that the proposed methods exhibit higher success rates against eleven different ML algorithms. Wang et al. [18] analyzed the underlying details of adversarial attacks in network intrusion detection: AEs tend to be close to their original clusters and the decision boundary. Based on the analysis, they propose a manifold and decision boundary-based defense method against adversarial attacks. The proposed method mainly leverages inconsistency between manifold evaluation and model inference to detect AEs, while evaluating the uncertainty of the target model against small perturbations. Although the security of ML-based IDSs has already attracted the attention of researchers with the increasing importance of the security of IIoT, current AE research primarily focuses on intrusion detection of the traditional Internet. The AE research in IIoT is still in its infancy. Besides, the current research on AEs against ML-based IDSs mainly explores the generation of input-dependent adversarial perturbations. The universal adversarial perturbations against ML-based IDSs received restricted attention. The universal adversarial perturbations expose the geometric correlation of the high-dimensional decision boundaries of distinct ML classifiers. Once the universal adversarial perturbations are produced, they can be applied to different original inputs and target classifiers, thereby exhibiting better transferability and efficiency. On the other hand, the research on universal adversarial perturbations can further contribute to understanding the origin of AEs. Therefore, research on universal adversarial perturbations in IIoT is imperative to better evaluate the robustness of the ML-based IDS in this security-critical area.

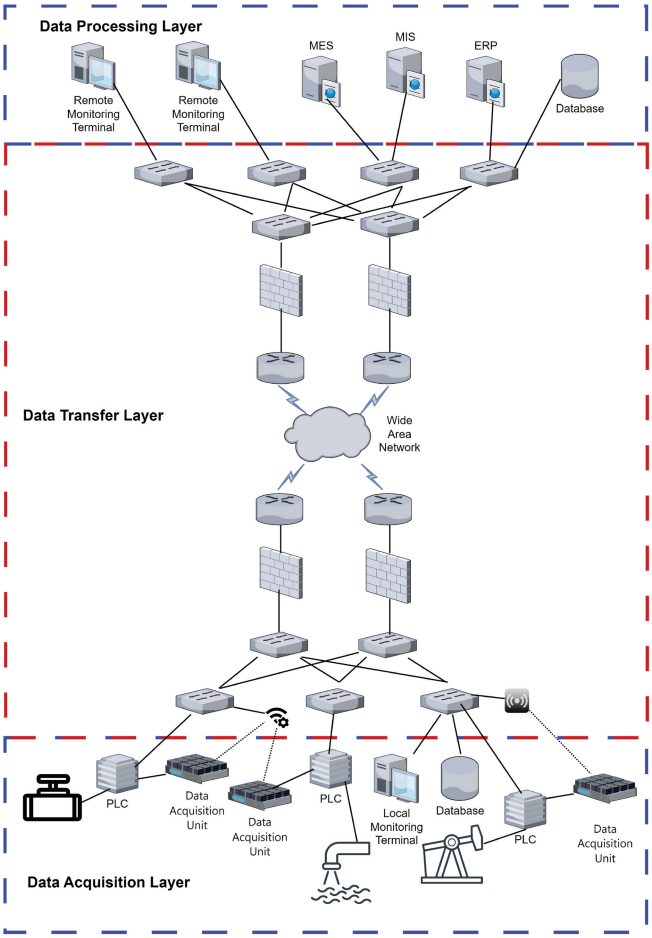


Fig. 1. Simplified structure of IIoT.

III. Background Knowledge

This section presents the background knowledge involved in the subsequent sections.

A. Industrial Internet of Things

A simplified structure of IIoT is shown in Fig. 1. The typical IIoT normally consists of three layers: 1) the data processing layer (DPL); 2) the data transfer layer (DTL); and 3) the data acquisition layer (DAL). The DAL consists of various sensors, monitoring systems, acquisition devices, programmable logic controllers, mechanical facilities, etc. The DTL is responsible for exchanging data between heterogeneous devices in IIoT. This layer is generally an integration of Internet, mobile networks, wireless networks, software-defined networks, etc. The DPL is responsible for monitoring the operating status of IIoT and intelligent decision-making based on the industrial process data. Also, this layer normally involves the process of enterprise management data, such as management information systems (MISs), enterprise resource planning (ERP) systems, manufacturing execution systems (MESs), etc. The IIoT combines traditional information technologies and industrial

control technologies through different networks to obtain real-time data on industrial processes, through which the efficiency of the management and production of industrial processes are improved and better enterprise-level intelligent decisions can be made. On the other hand, the frequent interaction between DPL and DAL gives adversaries the chance to interfere with the industrial devices, which finally leads to attacks that will influence the physical world.

The industrial control devices in the DAL communicate with each other through specific industrial communication protocols, such as Modbus, Profinet, intercontrol center communications protocol, etc. Instead, the communication between the DPL and the DAL normally utilizes TCP/IP protocols. The earlier ICSs tend to have a closed architecture, which results in inadequate security design for widely adopted industrial protocols. Therefore, mainstream industrial communication protocols are susceptible to various cyber-attacks. The security of IIoT has become the hotspot of the current research in cyber-security.

B. Adversarial Attacks

The emergence of AEs is initially observed in computer vision [9]. Intentionally crafted noises added to original inputs can mislead ML classifiers. With further exploration of AEs, researchers unveil the existence of AEs in diverse domains, such as intrusion detection [15], malware detection [12], reinforcement learning [13], natural language processing [27], and speech recognition [28]. These existing works show that the presence of AEs reveals the vulnerability of ML algorithms themselves. AEs hinder the further application of ML in security-critical fields, such as cyber-security and industrial control. ML-based IDSs have been proven to be an effective tool to mitigate cyber-attacks against IIoT. Accordingly, the robustness of ML-based IDSs against AEs has currently attracted attention in academia and industry. The current research predominantly employs classic adversarial attack methods to evaluate the robustness of ML-based IDSs, i.e., FGSM, PGD, MIM, DP, CW, etc. We mainly introduce the details of these classic attack methods in the subsequent parts of this section. All the attack methods introduced below except DP can be implemented in a targeted or nontargeted version. We only show their targeted versions here.

1) FGSM: The FGSM is a one-step method for constructing AEs [21]. The process of generating AEs by FGSM can be formalized as (1), where $\boldsymbol{x}$ denotes original inputs, $\boldsymbol{x}^*$ denotes the corresponding AE, ϵ is the applied perturbation strength, y^* is the target label that the adversary wants the target classifier to output for the AE, J denotes the loss function, and sign denotes the sign function. The FGSM utilizes the gradient information of the target classifier to craft AEs. The generation of the adversarial perturbation only needs a single computation of the gradient of the loss function with respect to the input. Consequently, the FGSM demonstrates high efficiency but the optimality of the generated adversarial perturbation can not be guaranteed

$$\boldsymbol{x}^* = \boldsymbol{x} - \epsilon \times \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}, y^*\right)\right). \tag{1}$$

2) PGD: The PGD produces AEs by iteratively applying (1). Therefore, PGD is regarded as the iterative version of FGSM. The formalization of PGD is shown in (2) [22], where $\boldsymbol{x}_i^*$ denotes the AE constructed in the ith iteration, $\boldsymbol{\delta}$ is a random initial perturbation, and α denotes the applied perturbation strength in each iteration. The other symbols in (2) share the same meaning as the corresponding ones in (1). The PGD constructs AEs from a random point in the ϵ-neighborhood of the original input. ϵ represents the acceptable maximum distortion imposed on the original input. The iterative adversarial attack methods are claimed to be stronger than the one-step ones [23]

$$\boldsymbol{x}_0^* = \boldsymbol{x} + \boldsymbol{\delta}, \quad \boldsymbol{x}_{i+1}^* = \boldsymbol{x}_i^* - \alpha \cdot \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_i^*, y^*\right)\right). \tag{2}$$

3) MIM: The MIM introduces the momentum into the iterative FGSM-based AE generation process [23]. The formalization of MIM is shown in (3) and (4), where $\boldsymbol{g}_i$ denotes the velocity vector accumulated in the ith iteration, $\|\cdot\|_1$ represents the L_1 norm, and μ is the decay factor. The other symbols in (3) and (4) share the same meaning as the corresponding ones in (2). The incorporation of the momentum is claimed to prevent converging to poor local extrema and construct better AEs

$$\boldsymbol{g}_{i+1} = \mu \cdot \boldsymbol{g}_i + \frac{\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_i^*, y^*\right)}{\left\|\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_i^*, y^*\right)\right\|_1} \tag{3}$$

$$\boldsymbol{x}_{i+1}^* = \boldsymbol{x}_i^* - \alpha \cdot \operatorname{sign}\left(\boldsymbol{g}_{i+1}\right). \tag{4}$$

4) DP: The DP is a nontargeted attack method [24]. The DP generates AEs by iteratively searching the closest distances between the original inputs to the decision boundaries of the target classifier. The approximate minimal adversarial perturbations are retrieved through linear approximation in each iteration. The approximate minimal adversarial perturbations produced in each iteration are accumulated to construct the final adversarial perturbations. Because the DP always looks for the closest distances to the decision boundaries to craft the perturbations, the crafted adversarial perturbations are claimed to be a good approximation of optimal ones. The AE generation of DP can be formalized as (5), where $\boldsymbol{x}$ denotes the original input, $\boldsymbol{x}_i$ denotes the AE produced at the ith iteration and the $\boldsymbol{r}_i$ is the approximate minimal perturbations generated in the ith iteration, which can be formalized as (6). f_l' and $\boldsymbol{\omega}_l'$ are the intermediate variables introduced during the linearization process of the target classifier. $\|\cdot\|_2$ represents the L_2 norm

$$\boldsymbol{x}_0 = \boldsymbol{x}, \boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \boldsymbol{r}_i \tag{5}$$

$$\boldsymbol{r}_i = \frac{\left|f_l'\right|}{\left\|\boldsymbol{\omega}_l'\right\|_2^2} \cdot \boldsymbol{\omega}_l'. \tag{6}$$

5) CW: CW design novel objective functions to effectively utilize traditional optimizers like Adam for generating AEs [25]. The L_2 attack of CW can be formalized as (7), where $\boldsymbol{x}$ still denotes the original input, $\boldsymbol{\omega}$ is an intermediate variable leveraged to eliminate the box constraint, and the constant c is used to balance the attack performance and the perturbation strength, typically achieved by an iterative binary searching process. CW have designed multiple candidates for

the function g and (8) is claimed to be the best choice in their experiments. The $\boldsymbol{x}'$ corresponds to the AE. The $Z(\boldsymbol{x}')$ denotes the output of the target classifier for $\boldsymbol{x}'$ before the final Softmax layer. $Z(\boldsymbol{x}')_i$ is the ith element of $Z(\boldsymbol{x}')$ and $Z(\boldsymbol{x}')_t$ represents the output for the target class specified by the adversary. The k is used to control the confidence score for outputting the target label. The CW claims to break down most state-of-the-art adversarial defense methods and is more effective than most existing adversarial attack methods

$$\min_{\boldsymbol{\omega}} \left\| \frac{1}{2}(\tanh(\boldsymbol{\omega})+1) - \boldsymbol{x} \right\|_2^2 + c \cdot g\left(\frac{1}{2}(\tanh(\boldsymbol{\omega})+1)\right) \tag{7}$$

$$g(\boldsymbol{x}') = \max\left(\max_{i\neq t}(Z(\boldsymbol{x}')_i) - Z(\boldsymbol{x}')_t, -k\right). \tag{8}$$

6) Universal Adversarial Perturbation Attack Method: The adversarial attack methods introduced above produce specific adversarial perturbations for every original input. Consequently, this kind of adversarial perturbation is specific to original inputs, i.e., input-dependent adversarial perturbations. The universal adversarial perturbations are another sort of perturbation and were first revealed by Moosavi-Dezfooli [20] in computer vision. Moosavi-Dezfooli et al. design a groundbreaking schema to iteratively produce universal adversarial perturbations using DP. This article labels their universal adversarial perturbation generation method as the universal adversarial perturbation attack method (UAPAM). A universal adversarial perturbation is constructed on a small sample set. Then it can be added to different original inputs, even the ones that are not involved in the generation of the universal adversarial perturbations, to mislead different classifiers. Therefore, the universal adversarial perturbations are normally thought to be input-independent. The generation process of universal adversarial perturbations using UAPAM can be formalized as (9), where $\boldsymbol{v}_i$ denotes the universal adversarial perturbations obtained in the ith iteration. P is the projected function to limit the perturbation strength according to norm p. The parameter ξ controls the strength of the universal adversarial perturbation. $\Delta\boldsymbol{v}_i$ is the small perturbation produced using DP and a random sample in the ith iteration. The existence of universal adversarial perturbations reveals the geometric correlation of high-dimensional decision boundaries of different classifiers [20] and further proves the transferability of AEs across heterogeneous classifiers. The integration of the universal adversarial perturbations with the substitute model technique can make adversaries attack the ML-based IDS in IIoT effectively without knowing the internal details of the systems, posing more threats to ML-based security systems in this area

$$\boldsymbol{v}_{i+1} = P_{p,\xi}(\boldsymbol{v}_i + \Delta\boldsymbol{v}_i). \tag{9}$$

The adversarial attack methods introduced above are initially proposed in computer vision. Generally, every image pixel is normally allowed to be perturbed. Instead, the modifiable features of network traffic are normally constrained. Consequently, while they are widely adopted in the existing literature to assess ML-based IDSs, these adversarial attack methods are not natively suitable for evaluating ML-based IDSs in IIoT. The attack performance of these attack methods against ML-based IDSs in IIoT can be enhanced. On the other hand, crafting universal adversarial perturbations in IIoT is normally not as straightforward as it is in computer vision due to the restriction on the modifiable features. This is why the works on universal adversarial perturbations mainly focused on computer vision and the discourse regarding universal adversarial perturbations in intrusion detection for IIoT is relatively limited. This article primarily explores the construction of universal adversarial perturbations in intrusion detection for IIoT. The proposed attack methods in this article utilize the outputs of the attack methods introduced above (namely, FGSM, PGD, MIM, DP, and CW) as the basis to craft enhanced universal adversarial perturbations.

TABLE I
NSL-KDD DATA SET

NSL-KDD		Training	Testing
Normal		67343	9711
Attack	DoS	45927	7458
	Probe	11656	2421
	R2L	995	2887
	U2R	52	67
Total		125973	22544

C. Data Sets

To comprehensively validate the attack performance of the proposed methods against ML-based IDSs in IIoT, this article adopts three extensively employed intrusion detection data sets, i.e., NSL-KDD [29], Gas Pipeline [30], and Edge-IIoTset [31].

The NSL-KDD data set is widely used to assess the performance and the robustness of ML-based IDSs in TCP/IP networks [6], [15], [17], [18], [19]. As discussed in Section III-A, the IIoT utilizes TCP/IP protocols to exchange data between DPL and DAL. Consequently, the NSL-KDD data set is leveraged to simulate the malicious traffic across DPL and DAL. The NSL-KDD data set encompasses four kinds of attack traffic, i.e., Denial of Service (DoS), probe, remote to local (R2L), and user to root (U2R). The detailed statistics of NSL-KDD are shown in Table I.

The gas pipeline data set is collected on a real laboratory-scale SCADA system. The gas pipeline data set is widely used to evaluate the performance of ML-based IDSs in ICS [32], [33], [34], [35]. Accordingly, the gas pipeline data set is adopted to simulate the traffic of industrial protocol among different industrial devices in DAL. The gas pipeline data set comprises four common types of attacks in ICS, which can be further divided into seven categories. The detailed information about the gas pipeline data set is shown in Table II.

The edge-IIoTset is a recently developed intrusion detection data set focusing on IoT and IIoT, which is designed to avoid some existing problems of the mainstream intrusion detection data sets in IoT or IIoT, such as insufficient industrial protocol and attack types. All the traffic data in edge-IIoTset are captured on an authentic testbed that comprises

TABLE II
GAS PIPELINE DATA SET

Attacks	Description	Counts
NMRI	Naive malicious response injection attack	2763
CMRI	Complex malicious response injection attack	15466
MSCI	Malicious state command injection attack	782
MPCI	Malicious parameter command injection attack	7637
MFCI	Malicious function command injection attack	573
DoS	Denial-of-service attack	1837
Reconnaissance	Reconnaissance attack	6805

cloud/edge configuration, diverse IoT/IIoT devices, and mainstream IoT/IIoT connection protocols. The edge-IIoTset data set is widely used to evaluate the ML-based IDSs in the context of centralized or federated learning [36], [37]. The edge-IIoTset data set covers the latest attack types in IoT and IIoT, which can be categorized into five common threats. The detailed statistics of edge-IIoTset are illustrated in Table III.

IV. METHODOLOGY

In this section, the details of the proposed attack methods (i.e., UAPG-IIoT and EUAPG-IIoT) are illustrated. The formal description of the problem concerning AE generation is initially presented, followed by a comprehensive theoretical derivation of the proposed methods.

Generating adversarial perturbations in a targeted way can be formalized as an optimization problem in (10) [9], [21], [22], [23], [24], [25], where $\boldsymbol{\delta}$ denotes the produced adversarial perturbations, C denotes the target ML classifier, and $\|\cdot\|_p$ represents an arbitrary norm. The other symbols in (10) have the same meaning as the corresponding ones in (1)

$$\begin{aligned} \min \quad & \|\boldsymbol{\delta}\|_p \\ \text{s.t.} \quad & C(\boldsymbol{x}+\boldsymbol{\delta}) = y^* \\ & \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^m. \end{aligned} \tag{10}$$

Generally, (10) can not be directly solved with the common solution in ML due to the nonconvexity and nonlinearity of the target classifier. Hence, most existing works obtain the adversarial perturbations by transforming the AE generation problem from (10) to (11), where f is a sort of objective function. For the gradient-based attack methods, the loss function is normally chosen as f. These attack methods leverage the gradient information of the loss function concerning original inputs to produce adversarial perturbations. For the gradient-based or optimization-based attack methods, the problem in (11) can be simplified as (12) to obtain the approximate adversarial perturbations. The function p takes the target classifier parameterized as $\boldsymbol{\theta}$, the original input $\boldsymbol{x}$, the label l, and loss function J as inputs to generate adversarial perturbations $\boldsymbol{\delta}$. The symbol $\pm$ corresponds to different types of adversarial targets. The $+$ denotes the generation of nontargeted AEs, while the $-$ indicates the generation of targeted AEs

$$\begin{aligned} \min_{\boldsymbol{x}^*} \quad & c \times \left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|_p + f\left(\boldsymbol{x}^*, y^*\right) \\ \text{s.t.} \quad & \boldsymbol{x}^* \in [0,1]^m. \end{aligned} \tag{11}$$

In the field of computer vision, the crafted perturbation $\boldsymbol{\delta}$ needs to be small enough to be imperceptible to human eyes. However, this restriction is not applicable to intrusion detection because the network traffic lacks vision semantics. In intrusion detection, this restriction is normally replaced by ensuring the validity of the traffic, which means that functional features must be preserved during the generation of adversarial traffic. On the other hand, the optimal solution in (10) and (11) is not necessary for the construction of AEs in intrusion detection for IIoT. The main objective of adversarial perturbations is to shift the original network traffic to the other side of the target classifier's decision boundary, making the original inputs misclassified. Therefore, the computation of (11) can be simplified during the practical construction of adversarial perturbations in intrusion detection for IIoT. First, the first term in (11) is replaced by modifying the limited feature subset of the original network traffic. Second, the constraint condition in (11) can be replaced by a clip function. Accordingly, the adversarial perturbations in intrusion detection for IIoT can be further represented as (13), where $\boldsymbol{\delta}_*$ denotes the ultimately produced adversarial perturbations. The function T performs a series of transformations, including the clipping, the discretization, etc. The $\boldsymbol{\psi}$ represents the parameter set for the transforming procedures, e.g., minima and maxima of original features, the threshold of discretization, etc. The function V utilizes the perturbations clipped by the function T to obtain the final perturbations that only modify the designated features of original inputs based on $\boldsymbol{\omega}$. The $\boldsymbol{\omega}$ is the index set of modifiable features of original network traffic. The final adversarial network traffic can be achieved through (14)

$$\begin{aligned} \boldsymbol{x}^* &= \boldsymbol{x} \pm \boldsymbol{\delta} \\ &= \boldsymbol{x} \pm p(\boldsymbol{\theta}, \boldsymbol{x}, l, J) \\ &= \boldsymbol{x} \pm p(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, l)) \end{aligned} \tag{12}$$

$$\begin{aligned} \boldsymbol{\delta}^* &= V(T(\boldsymbol{\delta}, \boldsymbol{\psi}), \boldsymbol{\omega}) \\ &= V(T(p(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, l), \boldsymbol{\psi}), \boldsymbol{\omega}) \end{aligned} \tag{13}$$

$$\begin{aligned} \boldsymbol{x}^* &= \boldsymbol{x} \pm \boldsymbol{\delta}^* \\ &= \boldsymbol{x} \pm V(T(\boldsymbol{\delta}, \boldsymbol{\psi}), \boldsymbol{\omega}) \\ &= \boldsymbol{x} \pm V(T(p(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, l), \boldsymbol{\psi}), \boldsymbol{\omega}). \end{aligned} \tag{14}$$

The adversarial perturbations $\boldsymbol{\delta}_*$ produced by the gradient-based or optimization-based attack methods comprise gradient information, which inspires us that the gradient information can be extracted from AEs to construct adversarial perturbations with better performance. Due to the shared paradigm among most gradient-based or optimization-based attack methods, the original perturbations with gradient information can be extracted through a unified interface. Consequently, we propose a new framework to utilize AEs produced by gradient-based or optimization-based attack methods to generate universal adversarial perturbations against ML-based IDSs in IIoT. The workflow of the proposed framework is shown in Fig. 2. The proposed framework shields the underlying details of the target adversarial attack methods, utilizing the adversarial perturbations of the target attack methods in a closed-box way and making it suitable for most adversarial scenarios in IIoT.

Based on the proposed framework UAPGF-IIoT, two adversarial attack methods are proposed to construct high-quality universal adversarial perturbations, i.e., UAPG-IIoT

TABLE III
EDGE-IIOTSET DATA SET

Threat Types	Details of the Threats	Counts
DoS/DDoS	Denial of service attacks including TCP SYN flood DDoS attacks, UDP flood DDoS attacks, HTTP flood DDoS attacks, and ICMP flood DDoS attacks	337977
Information Gathering	Collecting the information of the target system including port scanning, OS fingerprinting, and vulnerability scanning attack	73675
Man in the Middle	Attacks launched by the man in the middle including DNS spoofing attacks and ARP spoofing attacks	1214
Injection	Mainstream attacks against the B/S structure including XSS attacks, SQL injection, and uploading attacks	104752
Malware	Common malware threats in IoT/IIoT including backdoor attacks, password cracking attacks, and ransomware attacks	85940

and EUAPG-IIoT. The detailed workflow for UAPG-IIoT is shown in Fig. 3. The UAPG-IIoT first produces the original AEs using the target attack methods, such as FGSM, PGD, MIM, DP, CW, etc. Then, the original network traffic and corresponding original adversarial network traffic are utilized to extract the original adversarial perturbations. The extracting process can be formulated as (15), where $\boldsymbol{\delta}_i^*$ represents the original adversarial perturbation extracted from the seed sample $\boldsymbol{x}_i$, $\boldsymbol{x}_i^*$ denotes the corresponding AE of the seed sample $\boldsymbol{x}_i$ produced by the target attack methods, $\boldsymbol{\delta}_i'$ denotes the final universal adversarial perturbations accumulated on the first i seed sample, and $\boldsymbol{X}$ denotes a seed data set collected by the adversary to produce the final universal adversarial perturbations. The UAPG-IIoT extracts the gradient information from multiple seed traffic and integrates them into the ultimate universal adversarial perturbations in a straightforward way. Then, the produced universal adversarial perturbations can be added to arbitrary network traffic to construct adversarial network traffic that can mislead the ML-based IDSs in IIoT. The detailed process of generating universal adversarial perturbations by UAPG-IIoT is shown in Algorithm 1.

The test set $\boldsymbol{D}$ comprises two parts: 1) the original inputs $\boldsymbol{D}_{\text{data}}$ and 2) the corresponding labels $\boldsymbol{D}_{\text{label}}$. The evaluation function fool takes the target model, test set, and labels of the test set to calculate the fooling rate of current universal adversarial perturbations. The fooling rate r is a preset threshold controlling the attack strength. The UAPG-IIoT will terminate when the current fooling rate r^* achieved by $\boldsymbol{\delta}'$ reaches or exceeds r. The other parts in Algorithm 1 are self-explanatory. The 16th and 26th lines in Algorithm 1 guarantee the produced $\boldsymbol{\delta}'$ to be as small as possible, which means that the modification to the original network traffic is minimized as much as possible, ensuring the validity of the adversarial network traffic.

$$\begin{aligned} \boldsymbol{\delta}_i^* &= \mp\left(\boldsymbol{x}_i^* - \boldsymbol{x}_i\right) \\ \boldsymbol{\delta}_{i+1}' &= \boldsymbol{\delta}_i' + \boldsymbol{\delta}_{i+1}^* \\ \boldsymbol{x}_i &\in \boldsymbol{X}. \end{aligned} \tag{15}$$

The UAPG-IIoT constructs universal adversarial perturbations by extracting useful information from multiple seed AEs. The generated universal adversarial perturbations possess more adversarial information that can mislead the target ML-based IDS than input-dependent adversarial perturbations. Consequently, the AEs crafted with the universal adversarial perturbations normally show better transferability. However, the UAPG-IIoT produces universal adversarial perturbations on a specified classifier. The adversarial information is specific to the target classifiers. To further improve the transferability of the universal adversarial perturbations across different ML classifiers, we incorporate ensemble learning methodology into UAPG-IIoT to propose a new attack method, i.e., EUAPG-IIoT. Assuming that there is a target IDS set $\boldsymbol{\Phi} = \phi_0, \phi_1, \phi_2, \ldots, \phi_d$, the IDSs in this set are based on different ML algorithms. The computation of the gradient of ML classifiers except NN is nontrivial. Consequently, a substitute NN is trained for each ML classifier, i.e., SVM,

Algorithm 1 UAPG-IIoT

Input: Seed set $\boldsymbol{X}$; a gradient-based or optimization-based attack method φ; adversarial target o; test set $\boldsymbol{D}$; transformation parameters $\boldsymbol{\psi}$; the index set of modifiable features $\boldsymbol{\omega}$; the evaluation function fool; the target model C; the fooling rate threshold r.

Output: Universal adversarial perturbations $\boldsymbol{\delta}'$.

1: $\boldsymbol{\delta}' = 0$
2: $r^* = 0$
3: **for** $\boldsymbol{x}$ in $\boldsymbol{X}$ **do**
4: $\quad \boldsymbol{x}^* = \varphi(\boldsymbol{x})$
5: $\quad$ add $\boldsymbol{x}^*$ to $\boldsymbol{X}^*$
6: **end for**
7: **for** $\boldsymbol{x}$ in $\boldsymbol{X}$ and the corresponding $\boldsymbol{x}^*$ in $\boldsymbol{X}^*$ **do**
8: $\quad$ **if** o **then**
9: $\quad\quad \boldsymbol{\delta}^* = \boldsymbol{x} - \boldsymbol{x}^*$
10: $\quad\quad \boldsymbol{\delta}' = \boldsymbol{\delta}' + \boldsymbol{\delta}^*$
11: $\quad\quad \boldsymbol{D}'_{\text{data}} = \boldsymbol{D}_{data} - \boldsymbol{\delta}'$
12: $\quad\quad \boldsymbol{D}_{adv} = V(T(\boldsymbol{D}'_{\text{data}}, \boldsymbol{\psi}), \boldsymbol{\omega})$
13: $\quad\quad$ **if** fool$(\text{C}, \boldsymbol{D}_{adv}, \boldsymbol{D}_{\text{label}}) > r^*$ **then**
14: $\quad\quad\quad r^* = \text{fool}(\text{C}, \boldsymbol{D}_{adv}, \boldsymbol{D}_{\text{label}})$
15: $\quad\quad$ **else**
16: $\quad\quad\quad \boldsymbol{\delta}' = \boldsymbol{\delta}' - \boldsymbol{\delta}^*$
17: $\quad\quad$ **end if**
18: $\quad$ **else**
19: $\quad\quad \boldsymbol{\delta}^* = \boldsymbol{x}^* - \boldsymbol{x}$
20: $\quad\quad \boldsymbol{\delta}' = \boldsymbol{\delta}' + \boldsymbol{\delta}^*$
21: $\quad\quad \boldsymbol{D}'_{\text{data}} = \boldsymbol{D}_{data} + \boldsymbol{\delta}'$
22: $\quad\quad \boldsymbol{D}_{adv} = V(T(\boldsymbol{D}'_{\text{data}}, \boldsymbol{\psi}), \boldsymbol{\omega})$
23: $\quad\quad$ **if** fool$(\text{C}, \boldsymbol{D}_{adv}, \boldsymbol{D}_{\text{label}}) > r^*$ **then**
24: $\quad\quad\quad r^* = \text{fool}(\text{C}, \boldsymbol{D}_{adv}, \boldsymbol{D}_{\text{label}})$
25: $\quad\quad$ **else**
26: $\quad\quad\quad \boldsymbol{\delta}' = \boldsymbol{\delta}' - \boldsymbol{\delta}^*$
27: $\quad\quad$ **end if**
28: $\quad$ **end if**
29: $\quad$ **if** $r^* \geq r$ **then**
30: $\quad\quad$ break
31: $\quad$ **end if**
32: **end for**
33: **return** $\boldsymbol{\delta}'$

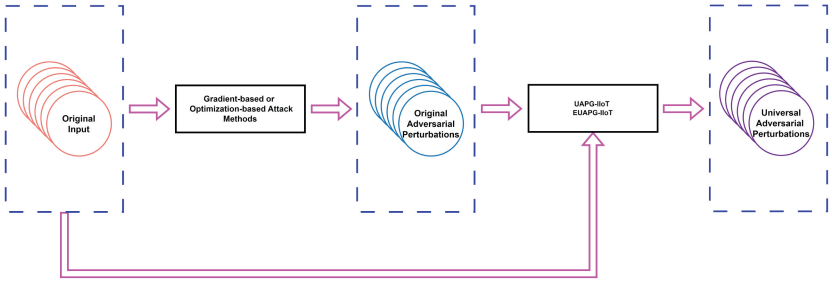


Fig. 2. Workflow of universal adversarial perturbation generation framework for IIoT. The UAPGF-IIoT leverages the original perturbations produced by the target attack methods to construct universal adversarial perturbations against ML-based IDSs in IIoT.

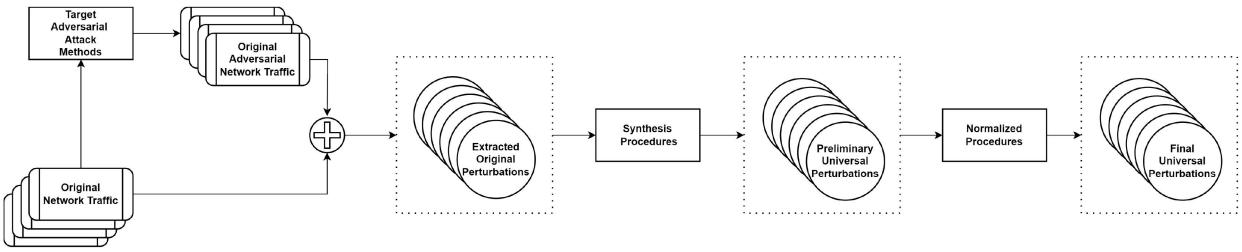


Fig. 3. Workflow of UAPG-IIoT.

LR, DT, NB, RF, etc. For each $\phi_j \in \mathbf{\Phi}$, a substitute NN is trained on the test set to simulate the decision boundary of the corresponding target classifier. Then, the UAPG-IIoT is utilized on each substitute NN to obtain the universal adversarial perturbations. After a simple aggregation, the best universal adversarial perturbations are selected from all the produced universal adversarial perturbations. The workflow of the EUAPG-IIoT is shown in Fig. 4.

The generation process of EUAPG-IIoT can be formulated as (16), where $\boldsymbol{\delta}_{\text{Best}}$ is the final best universal adversarial perturbations in Fig. 4, δ_i is the universal adversarial perturbations constructed on the target classifier ϕ_i, and the function 1 returns 1 when the condition is satisfied or returns 0. The main idea of EUAPG-IIoT is to leverage UAPG-IIoT to generate different universal adversarial perturbations on different ML classifiers. Then, the best universal adversarial perturbations are sifted from all the produced universal adversarial perturbations. Besides the universal adversarial perturbations crafted on different ML classifiers, the average and the sum of all the universal adversarial perturbations are also obtained as the candidates through the aggregation procedure in Fig. 4. Only five common ML algorithms are displayed in Fig. 4 for simplification. Based on the proposed framework, more ML algorithms can be incorporated into EUAPG-IIoT to produce universal adversarial perturbations with better transferability. Theoretically, the more ML algorithms are incorporated, the better transferability can be achieved. The pseudocode of EUAPG-IIoT is shown in Algorithm 2

$$\begin{aligned} \boldsymbol{\delta}_{\text{Best}} &= \boldsymbol{\delta}_k \\ k &= \max_i \sum_j \sum_{(\boldsymbol{x}, y^*)} 1\big(\phi_j(\boldsymbol{x} + \boldsymbol{\delta}_i) == y^*\big). \end{aligned} \tag{16}$$

Lines 4 to 7 in Algorithm 2 utilize UAPG-IIoT to obtain the universal adversarial perturbations against each target classifier. The average and sum of the generated universal adversarial perturbations are computed in lines 8 to 11, serving as potential candidates for the final universal adversarial perturbations. The rest part of Algorithm 2 shows the selection process of the best universal adversarial perturbations. The functions meane and sume compute the corresponding values column-wisely. However, the function mean computes the average of all the elements.

As shown in Algorithms 1 and 2, the proposed attack methods are both straightforward and easy to implement. The underlying details of the gradient-based or optimization-based attack methods are not required for the construction of the universal adversarial perturbations, which makes the proposed attack methods highly available for the adversarial tasks in IIoT. Because the internal information of the target IDSs in IIoT is normally unknown to the adversaries, closed-box adversarial attack methods are more useful in this area. The universal adversarial perturbations show excellent performance and transferability in our experiments, which means that the universal adversarial perturbations produced by the proposed attack methods, together with the substitute model techniques, can be used to evaluate the robustness of ML-based IDSs more effectively in a closed-box way.

V. Experiments

This section presents the details of the evaluation experiments for the proposed attack methods. Specifically, the effectiveness and generalizability of the proposed methods against diverse ML-based IDSs across various real-world scenarios are comprehensively assessed in this section.

A. Experimental Setup

To comprehensively evaluate the effectiveness of the proposed attack methods, the commonly used ML classifiers

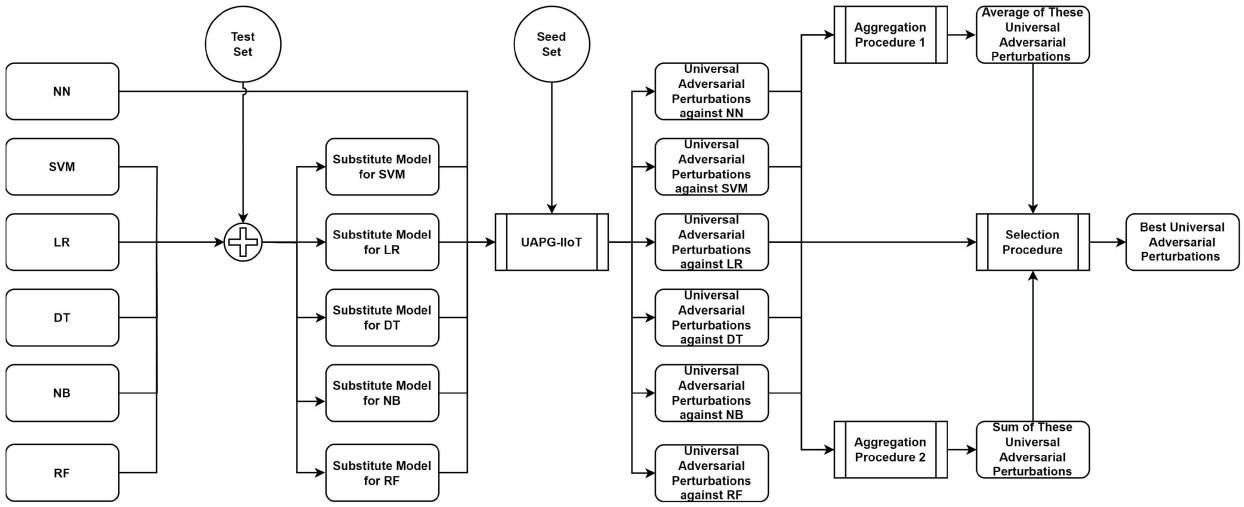


Fig. 4. Workflow of EUAPG-IIoT. The EUAPG-IIoT is based on the UAPG-IIoT.

are employed to build the target ML-based IDSs. Three mainstream intrusion detection data sets for IIoT are leveraged to evaluate the effectiveness of the proposed methods across different real-world scenarios. The mainstream gradient-based or optimization-based adversarial attack methods are adopted as the target attack methods to produce the universal adversarial perturbations. All the target attack methods are also employed as the baseline methods to validate the superiority of the proposed methods. Besides, a mainstream UAPAM is also adopted as the baseline method to further validate the superiority of the proposed methods in generating universal adversarial perturbation. The experimental setups are described in detail in the subsequent parts of this section.

Three widely utilized intrusion detection data sets [6], [15], [17], [18], [19], [32], [33], [34], [35], i.e., the NSL-KDD, Gas Pipeline, and edge-IIoTset, are adopted to evaluate the attack performance of the proposed methods across diverse real-world scenarios. The details of these data sets are shown in Section III-C. The NSL-KDD data set is used to simulate the network traffic in DPL. We mainly employ the DoS and Probe data to evaluate the proposed methods owing to their massive appearance in the real world. The gas pipeline data set is mainly leveraged to simulate the network traffic in DAL. Similarly, the CMRI and MPCI attacks are massively present in the real IIoT. Consequently, we utilize the CMRI and MPCI data in the gas pipeline data set to verify the proposed methods in the subsequent evaluation experiments. The edge-IIoTset data set covers the traffic data in IIoT with cloud/edge configuration, which can be utilized to assess the effectiveness of the proposed attack methods in this scenario. Specifically, the backdoor and ransomware attacks of malware threats in edge-IIoTset are mainly employed to validate the effectiveness of the proposed attack methods. The chosen data sets cover the management network traffic in IIoT, the industrial control traffic, and the IIoT traffic under the cloud/edge configuration, through which the effectiveness of the proposed attack methods under diverse scenarios of IIoT is verified. All the traffic data are scaled into [0, 1] through (17), where x' represents the scaled feature value, x is the original feature value before scaling, $x_{\min}$ denotes the minimum of the corresponding feature in the data set, and $x_{\max}$ is the maximum of the feature

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{17}$$

The widely used ML algorithms [10], [15], [17], [18], [19], i.e., NN, SVM, LR, DT, NB, and RF, etc., are employed to build the target ML-based IDSs in IIoT to comprehensively verify the generalizability of the proposed methods against different ML algorithms. Besides, the transferability of the produced universal adversarial perturbations across different ML algorithms is also assessed. The intrusion detection in IIoT is regarded as a binary classification in this article. All the attack data are labeled as 1 and all the normal data are labeled as 0.

The classic adversarial attack methods, i.e., FGSM, PGD, MIM, CW, and DP, are employed by the proposed methods as the target attack methods to construct the universal adversarial perturbations. These adversarial attack methods are chosen as the target attack methods due to the massive adoption [11], [15], [16] in the existing adversarial attack research in cyber-security. Besides, they are the typical representations of gradient-based or optimization-based attack methods. The intrusion detection in IIoT is regarded as a binary classification in this article. The targeted attack and nontargeted attack share the same effect. Consequently, only the targeted attack of all the adversarial attack methods except DP are utilized to produce AEs. Simultaneously, these target attack methods serve as the baseline methods, with which the attack performance of the proposed methods is compared. To better clarify the availability of the proposed attack methods in IIoT, the attack performance of all the methods is compared and analyzed under two settings, i.e., all the traffic features can be revised and only the nonfunctional features can be modified. As discussed above, the modifiable features are limited during the generation of AEs in intrusion detection for IIoT. Accordingly, the variation in attack performance before and after the restriction on the number of modifiable

Algorithm 2 Ensemble UAPG-IIoT

Input: Seed set $\boldsymbol{X}$; the target model set $\boldsymbol{\Phi}$; a gradient-based or optimization-based attack method φ; UAPG-IIoT attack method λ; adversarial target o; test set $\boldsymbol{D}$; transformation parameters $\boldsymbol{\psi}$; the index set of modifiable features $\boldsymbol{\omega}$; the evaluation function fool; the fooling rate threshold r.

Output: Best universal adversarial perturbations $\boldsymbol{\delta}_{\text{Best}}$.

$\boldsymbol{\delta}_{\text{Best}} = 0$
$r^* = 0$
$\boldsymbol{\rho} = []$
for ϕ in $\boldsymbol{\Phi}$ **do**
  $\boldsymbol{\rho}^* = \lambda(\boldsymbol{X}, \varphi, \phi, o, \boldsymbol{D}, \boldsymbol{\psi}, \boldsymbol{\omega}, \text{fool}, r)$
  $\boldsymbol{\rho}.\text{append}(\boldsymbol{\rho}^*)$
end for
$\boldsymbol{\tau} = \text{meane}(\boldsymbol{\rho})$
$\boldsymbol{\xi} = \text{sume}(\boldsymbol{\rho})$
$\boldsymbol{\rho}.\text{append}(\boldsymbol{\tau})$
$\boldsymbol{\rho}.\text{append}(\boldsymbol{\xi})$
for $\boldsymbol{\rho}'$ in $\boldsymbol{\rho}$ **do**
  $\boldsymbol{r}' = []$
  for ϕ in $\boldsymbol{\Phi}$ **do**
    if o **then**
      $\boldsymbol{D}'_{\text{data}} = \boldsymbol{D}_{\text{data}} - \boldsymbol{\rho}'$
      $\boldsymbol{D}_{adv} = V(T(\boldsymbol{D}'_{\text{data}}, \boldsymbol{\psi}), \boldsymbol{\omega})$
      $\boldsymbol{r}'.\text{append}(\text{fool}(\phi, \boldsymbol{D}_{adv}, \boldsymbol{D}_{\text{label}}))$
    else
      $\boldsymbol{D}'_{\text{data}} = \boldsymbol{D}_{\text{data}} + \boldsymbol{\rho}'$
      $\boldsymbol{D}_{adv} = V(T(\boldsymbol{D}'_{\text{data}}, \boldsymbol{\psi}), \boldsymbol{\omega})$
      $\boldsymbol{r}'.\text{append}(\text{fool}(\phi, \boldsymbol{D}_{adv}, \boldsymbol{D}_{\text{label}}))$
    end if
  end for
  if $\text{mean}(\boldsymbol{r}') > r^*$ **then**
    $r^* = \text{mean}(\boldsymbol{r}')$
    $\boldsymbol{\delta}_{\text{Best}} = \boldsymbol{\rho}'$
  end if
end for
return $\boldsymbol{\delta}_{\text{Best}}$

features can intuitively demonstrate the usability of the attack methods. Besides, the UAPAM introduced in Section III-B6 is also employed as the baseline method to demonstrate the superiority of the proposed attack methods to the mainstream UAPAM. The parameter settings for these baseline adversarial attack methods in the subsequent experiments are shown in Table IV. The $\epsilon_{\max}$ is the acceptable maximum distortion for original inputs. The n denotes the iterations for the iterative gradient-based attack methods. The lr is the learning rate for the Adam optimizer. The n_{bs} denotes the iterations of the binary search in CW for seeking the suitable c. The $n_{\max}$ is the maximum iteration that terminates the generation of AEs. The parameter η is used by DP to avoid converging to a point on the decision boundary of the target classifier.

The main evaluation metric used in this article is the detection rate of the target IDS, which is shown in (18), where the $\boldsymbol{x}_{\text{oa}}$ denotes the number of the original attack examples, the $\boldsymbol{x}_{\text{oad}}$ denotes the number of the detected original attack examples, the $\boldsymbol{x}_{\text{aa}}$ denotes the number of the adversarial attack examples, and the $\boldsymbol{x}_{\text{aad}}$ denotes the number of the detected adversarial attack examples. The ODR is short for the original detection rate, which is the detection rate for the original attack examples. The ADR is short for the adversarial detection rate, which is the detection rate for the adversarial attack examples. The effectiveness of the proposed methods is intuitively validated by examining the variations in the detection performance of the target IDS before and after the addition of the adversarial perturbations. The attack methods demonstrate superior performance as the detection rate decreases.

TABLE IV
PARAMETER SETTINGS FOR THE BASELINE ADVERSARIAL ATTACK METHODS

Attack Method	Parameter Setting
FGSM	$\epsilon = 0.3$
PGD	$\epsilon_{max} = 0.3, \alpha = 0.03, n = 10$
MIM	$\epsilon_{max} = 0.3, \alpha = 0.03, \mu = 1.0, n = 10$
CW	$k = 0, lr = 0.01, n_{bs} = 9, n_{max} = 300, c = 0.001$
DP	$\eta = 0.02, n_{max} = 50$
UAPAM	$\epsilon_{max} = 0.3, \eta = 0.02$

All the experiments are conducted on a computer with an AMD Ryzen 7 2700X CPU and 64-GB RAM. Two Nvidia Geforce GTX 1080 GPUs are employed to accelerate the computation. The PyTorch [38] and Scikit-learn [39] are utilized to build up different ML-based IDSs in IIoT and implement UAPG-IIoT and EUAPG-IIoT.

$$\text{ODR} = \frac{\boldsymbol{x}_{\text{oad}}}{\boldsymbol{x}_{\text{oa}}}$$
$$\text{ADR} = \frac{\boldsymbol{x}_{\text{aad}}}{\boldsymbol{x}_{\text{aa}}}. \tag{18}$$

B. Evaluation for the Influence of Seed Set on the Attack Performance

As shown in Algorithms 1 and 2, the selection of the seed set is very important for utilizing the proposed methods to craft the universal adversarial perturbations. The quality and quantity of the seed set influence the attack performance of the proposed methods. Therefore, to comprehensively assess the proposed attack methods, we first explore the influence of the seed set on the attack performance of the proposed methods. In the real world, adversarial attacks generally occur in the inference stage of the target classifiers. Consequently, the test set is utilized to assess the attack performance of the adversarial attack methods. We adopt different sampling rates to randomly select the seed set from the test set. Then, the selected seed sets are utilized to craft the universal adversarial perturbations against NN-based IDSs. Finally, the attack performance of the universal adversarial perturbations constructed by the proposed methods on different seed sets is evaluated. The experimental results are shown in Figs. 5 and 6.

From Fig. 5(a) and (c), we can observe that the seed set has little influence on the attack performance of UAPG-IIoT when the modifiable features are not limited because the detection rates all decrease to zero under different sampling

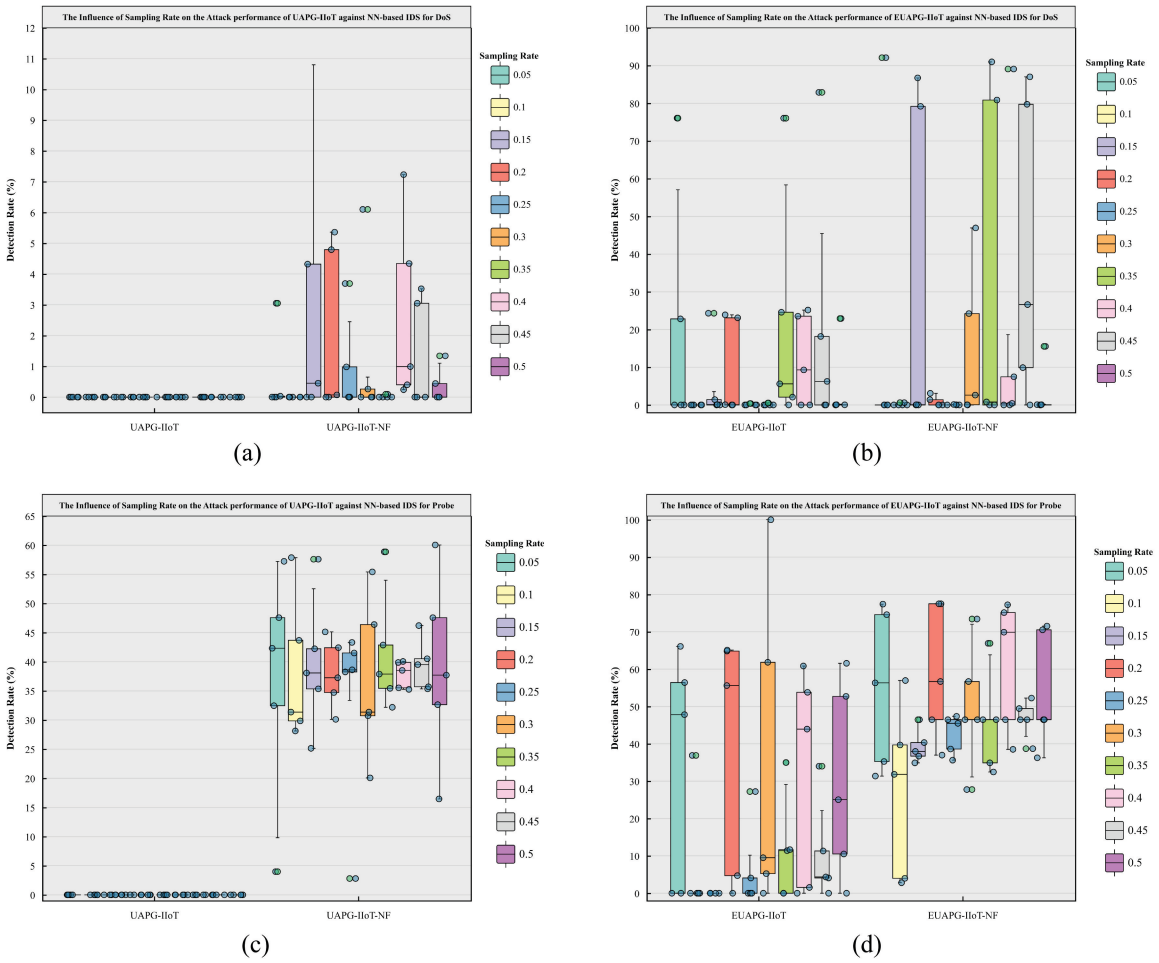


Fig. 5. Evaluation results regarding the influence of sampling rate on the attack performance against NN-based IDSs in DPL. In DPL, the attack performance of UAPG-IIoT and EUAPG-IIoT combined with different target attack methods (FGSM, PGD, MIM, CW, and DP) under different sampling rates ([0,05, 0.5] with an interval 0.05). (a) and (b) Exhibit the evaluation results achieved on DoS. (c) and (d) Show the evaluation results achieved on Probe. The UAPG-IIoT and EUAPG-IIoT denote the adversarial attacks in which all the features of the original inputs can be modified. The UAPG-IIoT-NF and EUAPG-IIoT-NF represent attacks that only can modify the nonfunctional features.

rates on both DoS traffic and probe traffic. After restricting the modifiable features, a low-sampling rate (e.g., 0.05 or 0.1) can already assist the UAPG-IIoT to obtain a good performance. From Fig. 5(b) and (d), we can see that the sampling rate has a greater impact on the attack performance of EUAPG-IIoT. Similarly, a low-sampling rate (0.05, 0.1, or 0.15) is sufficient for EUAPG-IIoT to produce universal adversarial perturbations with excellent performance.

The sampling rate has little impact on the attack performance of UAPG-IIoT and EUAPG-IIoT against NN-based IDSs in DAL, which can be observed in Fig. 6. When the number of modifiable features is not restricted, the universal adversarial perturbations constructed by UAPG-IIoT and EUAPG-IIoT under different sampling rates all diminish the detection rates of the target IDSs to zero on both CMRI traffic and MPCI traffic. While the detection rates increase after limiting the number of modifiable features, the detection rate is always maintained at a low level (not exceeding 1%). On CMRI traffic, the differences in the attack performance of the proposed methods combined with different target attack methods at the same sampling rate do not exceed 0.02%. On MPCI traffic, the corresponding differences stay within 3%.

According to the experimental results in this section, the sampling rate of UAPG-IIoT and EUAPG-IIoT is set to 0.1 in the subsequent experiments.

C. Adversarial Attacks Against DPL

In this section, the attack performance and transferability of the universal adversarial perturbations crafted by the UAPG-IIoT and EUAPG-IIoT against ML-based IDSs in DPL are evaluated. All the ML-based IDSs are trained on the training set of the NSL-KDD data set. The ten-fold cross-validation and grid search are leveraged to seek the optimal parameters for the target IDSs. 10% of the test set is used as the seed set for UAPG-IIoT, EUAPG-IIoT, and UAPAM to produce universal adversarial perturbations. The rest of the test set is utilized to assess the attack performance of the adversarial attack methods and the detection performance of the target ML-based IDSs. For the EUAPG-IIoT, all the substitute NNs are also trained on the test set. The predicted labels from the target

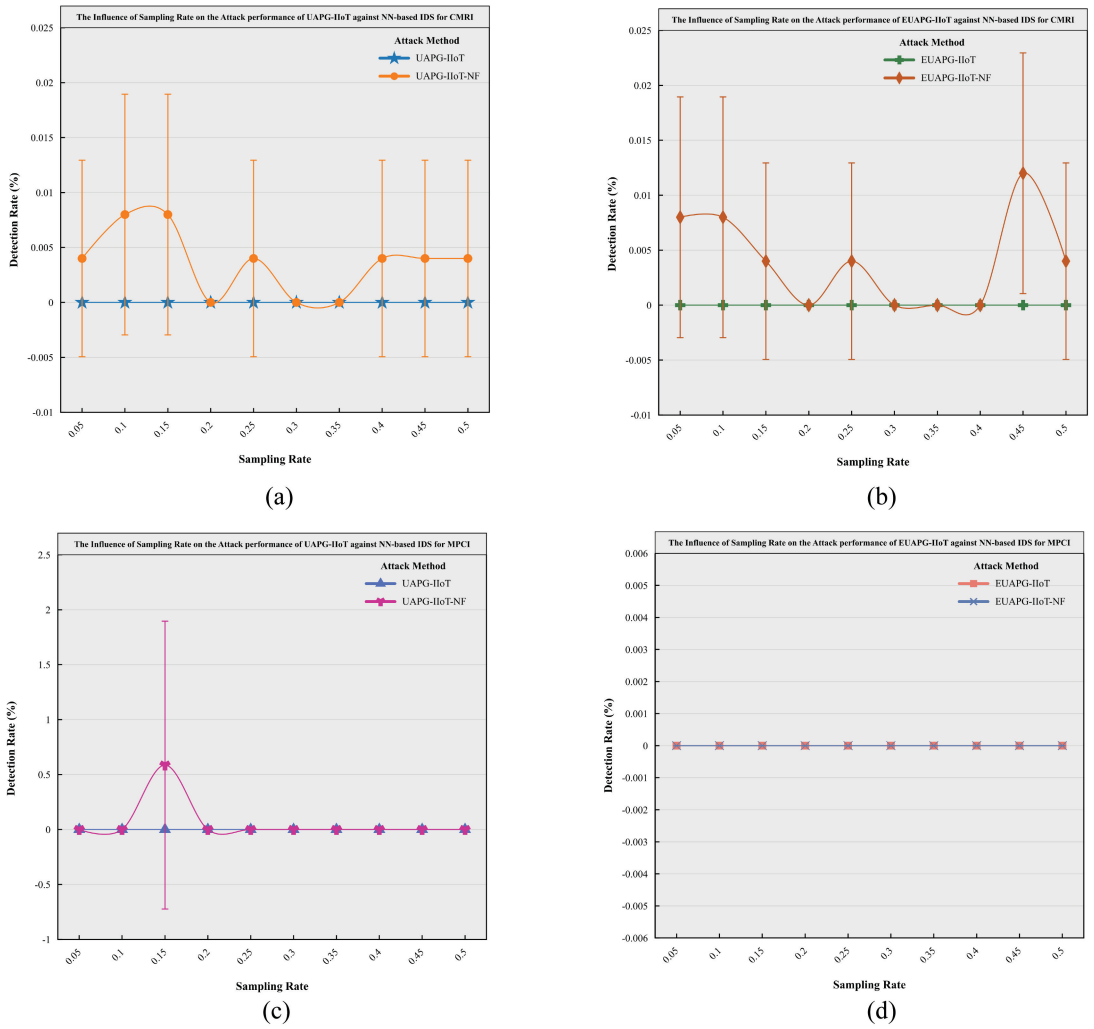


Fig. 6. Evaluation results regarding the influence of sampling rate on the attack performance against NN-based IDSs in DAL. In DAL, the attack performance of UAPG-IIoT and EUAPG-IIoT combined with different target attack methods (FGSM, PGD, MIM, CW, and DP) under different sampling rates ([0,05, 0.5] with an interval 0.05). (a) and (b) Exhibit the evaluation results achieved on CMRI. (c) and (d) Show the evaluation results achieved on MPCI. The UAPG-IIoT and EUAPG-IIoT denote the adversarial attacks in which all the features of the original inputs can be modified. The UAPG-IIoT-NF and EUAPG-IIoT-NF represent attacks that only can modify the nonfunctional features.

ML classifiers are used as the training labels for the original network traffic to obtain the approximate decision boundaries of the target classifiers. For the different attack traffic in NSL-KDD, the functional features are different. The features of network traffic in NSL-KDD can be categorized into four groups: 1) intrinsic; 2) content; 3) time-based traffic; and 4) host-based traffic. Table V displays the functional features of each attack type [19]. The *F* denotes that the features in the group are functional for the corresponding attack category. Accordingly, the *NF* shows that the features are nonfunctional for the attack category. Accordingly, the $\boldsymbol{\omega}$ for UAPG-IIoT and EUAPG-IIoT is set to the indexes of the nonfunctional features in Table V. The setting of $\boldsymbol{\psi}$ includes 0 for the lower bound of clipping and 1 for the upper bound of clipping. The r is set to 1.

We first evaluate the attack performance of the adversarial attack methods against the NN-based IDS in DPL. Then, the AEs produced on the NN-based IDS are directly applied to evaluate the robustness of the other ML-based IDSs, through which the transferability of the crafted AEs across different ML classifiers is also assessed. The experimental results of adversarial attacks against ML-based IDSs in DPL are shown in Tables VI–IX. All the adversarial attack methods without the symbol UAPG-IIoT or EUAPG-IIoT in these tables denote the original attack performance of the classic gradient-based or optimization-based attack methods. The attack methods marked with the symbol UAPG-IIoT or EUAPG-IIoT display the attack performance of the universal adversarial perturbations crafted by the proposed attack methods combined with

TABLE V
FUNCTIONAL FEATURES FOR EACH ATTACK TYPE IN NSL-KDD

Features \ Attacks	DoS	Probe	U2R	R2L
Intrinsic	*F*	*F*	*F*	*F*
Content	*NF*	*NF*	*F*	*F*
Time-based Traffic	*F*	*F*	*NF*	*NF*
Host-based Traffic	*NF*	*F*	*NF*	*NF*

TABLE VI
ADVERSARIAL ATTACKS AGAINST NN-BASED IDSS IN DPL WITHOUT RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	DoS		Probe	
	ODR(%)	ADR(%)	ODR(%)	ADR(%)
FGSM	99.06	19.53	86.71	25.77
UAPG-IIoT(FGSM)		**0.00**		**0.00**
EUAPG-IIoT(FGSM)		**0.00**		**0.00**
PGD		0.24		4.61
UAPG-IIoT(PGD)		**0.00**		**0.00**
EUAPG-IIoT(PGD)		**0.00**		**0.00**
MIM		0.89		3.98
UAPG-IIoT(MIM)		**0.00**		**0.00**
EUAPG-IIoT(MIM)		**0.00**		**0.00**
CW		0.02		**0.00**
UAPG-IIoT(CW)		**0.00**		**0.00**
EUAPG-IIoT(CW)		**0.00**		**0.00**
DP		0.94		13.29
UAPG-IIoT(DP)		**0.00**		**0.00**
EUAPG-IIoT(DP)		**0.00**		42.95
UAPAM		2.18		50.45

TABLE VII
ADVERSARIAL ATTACKS AGAINST NN-BASED IDSS IN DPL AFTER RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	DoS		Probe	
	ODR(%)	ADR(%)	ODR(%)	ADR(%)
FGSM-NF	99.06	75.25	86.71	58.14
UAPG-IIoT-NF(FGSM)		0.61		38.61
EUAPG-IIoT-NF(FGSM)		**0.00**		**2.35**
PGD-NF		72.29		59.40
UAPG-IIoT-NF(PGD)		0.91		38.70
EUAPG-IIoT-NF(PGD)		0.12		4.07
MIM-NF		73.84		57.69
UAPG-IIoT-NF(MIM)		0.07		20.43
EUAPG-IIoT-NF(MIM)		**0.00**		31.74
CW-NF		99.04		86.71
UAPG-IIoT-NF(CW)		**0.00**		50.81
EUAPG-IIoT-NF(CW)		**0.00**		37.25
DP-NF		96.20		86.71
UAPG-IIoT-NF(DP)		**0.00**		37.52
EUAPG-IIoT-NF(DP)		**0.00**		70.07
UAPAM-NF		85.68		73.60

the corresponding target attack methods. The best experimental results in each ADR column of all the tables are boldly represented.

Table VI displays the attack performance of the adversarial attack methods against NN-based IDS without limiting the number of modifiable features. Except for EUAPG-IIoT(DP), the other target attack methods combined with the proposed methods show better performance than the original ones on both DoS and probe traffic. The DP combined with EUAPG-IIoT shows excellent attack performance on DoS traffic, while the attack performance on probe traffic is not ideal. The proposed methods demonstrate their superiority to the classic universal perturbation attack method (namely, UAPAM) on both DoS and probe traffic.

A similar trend can be observed in Table VII. After limiting the number of modifiable features, i.e., only modifying the nonfunctional features in Table V, the attack performance of the target attack methods against NN-based IDS degrades obviously on both DoS traffic and probe traffic. On DoS, the target attack methods combined with the proposed methods maintain excellent performance, most of which diminish the detection rate of the target classifiers close to 0. On Probe, the restriction on the number of modifiable features has a greater influence on the attack performance. This is because the network traffic in Probe possesses fewer modifiable nonfunctional features. Most target attack methods combined with EUAPG-IIoT show better attack performance than the ones combined with UAPG-IIoT. From the results in Table VII, it can be observed that restricting the number of modifiable features imposes a more significant impact on the attack performance of UAPAM than on that of the proposed methods.

Table VIII exhibits the transferability of the produced AEs across heterogeneous ML-based IDSs in IIoT without limiting the number of modifiable features. The AEs crafted by the target attack methods combined with the proposed methods show the best transferability on all the ML-based IDSs. However, the AEs generated by FGSM and MIM also show good transferability on DT-based IDSs for Probe. Simultaneously, the AEs produced by FGSM and MIM show good transferability on RF-based IDSs for DoS. The universal adversarial perturbations produced by the proposed attack methods show better transferability on all target ML-based IDSs compared to those crafted by the UAPAM.

Restricting the number of modifiable features also imposes a great impact on the transferability of the AEs, which can be concluded from the experimental results in Table IX. The attack performance of most adversarial attack methods degrades. After only modifying the nonfunctional features, the transferability of AEs produced by the target attack methods is significantly diminished. AEs constructed by the proposed methods show better transferability on all the target ML-based IDSs. Likewise, the transferability of universal adversarial perturbations constructed by the UAPAM significantly degrades on different ML-based IDSs when the modifiable features are limited, which demonstrates that the classic universal perturbation attack method with excellent performance in computer vision can not effectively evaluate the robustness of ML-based IDSs in IIoT. The universal adversarial perturbations produced by the proposed attack methods are more effective in assessing ML-based IDSs in IIoT. The presence of certain outliers can be observed in Tables VIII and IX, e.g., the adversarial perturbations produced by DP on the NN-based IDS instead improve the detection performance of SVM-based IDS against adversarial probe traffic in Table VIII. In Table IX, the universal adversarial perturbations produced by UAPAM on the NN-based IDS instead improve the detection performance of LR-based IDS against adversarial probe traffic

D. Adversarial Attacks Against DAL

The adversarial attacks against ML-based IDSs in DAL are evaluated in this section on the gas pipeline data set. 70% of the data set is used as the training set to train all the target ML-based IDSs in DAL. Likewise, the ten-fold cross-validation and grid search are employed to train the optimal classifiers. The rest of the data set is utilized as the test set, of which 10% is leveraged as the seed set to produce the universal adversarial perturbations by UAPG-IIoT, EUAPG-IIoT, and UAPAM. The remaining portion of the test set is used to evaluate the attack performance of the adversarial attack methods and the detection performance of the target

TABLE VIII
TRANSFERABILITY OF THE AEs ACROSS DIFFERENT ML-BASED IDSs IN DPL WITHOUT RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	SVM(%)				LR(%)				DT(%)				NB(%)				RF(%)			
	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR
FGSM		77.18		63.11		77.37		74.68		27.14		9.49		80.11		60.85		0.21		44.21
UAPG-IIoT(FGSM)		95.14		0.09		95.02		40.78		80.91		**2.89**		27.80		41.50		6.24		36.62
EUAPG-IIoT(FGSM)		0.14		**0.00**		21.32		26.22		14.23		23.69		81.38		41.23		31.98		36.98
PGD		61.37		58.23		78.94		72.97		16.57		14.65		76.12		63.56		10.19		44.39
UAPG-IIoT(PGD)		9.34		**0.00**		88.09		44.39		**0.84**		53.07		**0.00**		43.13		**0.00**		**0.00**
EUAPG-IIoT(PGD)		0.24		**0.00**		2.39		31.74		81.45		87.25		**0.00**		11.39		**0.00**		17.18
MIM		75.60		63.83		78.28		72.97		18.72		9.76		78.52		55.42		0.16		44.85
UAPG-IIoT(MIM)	93.59	83.78	97.74	0.09	93.52	86.59	98.19	30.47	99.69	8.40	99.55	14.74	81.88	49.26	86.80	**6.60**	99.74	7.56	99.55	2.89
EUAPG-IIoT(MIM)		**0.00**		11.84		**0.00**		85.99		97.11		64.29		26.42		13.47		4.88		34.99
CW		92.70		95.93		93.29		97.02		18.34		26.49		81.57		86.80		13.69		74.68
UAPG-IIoT(CW)		78.80		98.28		81.50		97.02		90.05		12.93		**0.00**		50.45		17.49		4.25
EUAPG-IIoT(CW)		80.82		10.13		82.01		17.09		63.04		28.48		**0.00**		11.39		5.73		12.57
DP		93.24		97.83		93.45		97.74		23.99		34.09		81.50		86.80		12.61		70.16
UAPG-IIoT(DP)		100.00		99.19		100.00		93.49		40.72		59.40		**0.00**		10.49		0.52		12.57
EUAPG-IIoT(DP)		28.60		**0.00**		80.33		**0.00**		43.04		85.53		27.00		45.93		2.26		16.73
UAPAM		56.51		100.00		86.70		99.91		11.01		2.08		79.78		52.08		4.27		40.14

TABLE IX
TRANSFERABILITY OF THE AEs ACROSS DIFFERENT ML-BASED IDSs IN DPL AFTER RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	SVM(%)				LR(%)				DT(%)				NB(%)				RF(%)			
	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR	DoS ODR	DoS ADR	Probe ODR	Probe ADR
FGSM-NF		88.94		78.74		93.42		95.84		21.34		55.06		80.46		62.48		93.26		99.55
UAPG-IIoT-NF(FGSM)		99.98		41.77		99.69		79.48		**11.77**		53.98		27.42		50.36		80.82		96.56
EUAPG-IIoT-NF(FGSM)		36.25		7.05		92.34		**70.61**		11.84		54.88		27.17		42.59		90.12		97.29
PGD-NF		80.67		77.67		91.52		96.56		18.27		54.88		81.38		62.66		93.59		99.55
UAPG-IIoT-NF(PGD)		88.33		2.35		94.90		86.44		13.39		55.15		23.08		22.78		91.87		98.01
EUAPG-IIoT-NF(PGD)		15.62		**0.00**		93.49		73.60		37.85		54.88		26.39		**11.39**		91.69		**95.93**
MIM-NF		85.40		71.25		92.44		95.84		18.59		54.88		81.00		62.30		92.77		99.55
UAPG-IIoT-NF(MIM)	93.59	91.74	97.74	43.40	93.52	94.16	98.19	89.51	99.69	13.12	99.55	55.15	81.88	76.33	86.8	42.50	99.74	95.72	99.55	98.10
EUAPG-IIoT-NF(MIM)		**0.00**		68.81		**22.59**		98.55		39.73		55.15		26.39		20.16		45.74		97.74
CW-NF		93.47		97.02		93.49		98.01		88.89		**55.06**		81.66		86.71		99.20		99.55
UAPG-IIoT-NF(CW)		88.28		99.28		88.59		99.10		12.49		55.15		24.68		62.21		35.20		99.55
EUAPG-IIoT-NF(CW)		85.12		66.55		90.75		92.04		12.58		55.15		25.73		58.50		86.97		99.55
DP-NF		93.47		98.01		93.50		98.28		47.31		55.15		81.64		86.62		98.75		99.55
UAPG-IIoT-NF(DP)		100.00		99.64		100.00		99.55		12.49		55.15		**0.00**		49.55		**20.15**		97.92
EUAPG-IIoT-NF(DP)		90.49		65.55		93.07		92.04		13.99		54.88		66.80		64.83		86.67		99.55
UAPAM-NF		88.66		99.91		91.10		99.82		18.97		55.15		81.38		63.29		93.14		99.55

TABLE X
FUNCTIONAL FEATURES FOR CMRI AND MPCI IN THE GAS PIPELINE DATA SET

Attacks	Functional Features
CMRI	command_address,response_address,gain,reset, dead_band,cycletime,rate
MPCI	command_address,response_address,control_mode,pump, solenoid

TABLE XI
ADVERSARIAL ATTACKS AGAINST NN-BASED IDSS IN DAL WITHOUT RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	CMRI		MPCI	
	ODR(%)	ADR(%)	ODR(%)	ADR(%)
FGSM	98.78	**0.00**	98.34	53.89
UAPG-IIoT(FGSM)		**0.00**		**0.00**
EUAPG-IIoT(FGSM)		**0.00**		**0.00**
PGD		0.77		45.45
UAPG-IIoT(PGD)		**0.00**		**0.00**
EUAPG-IIoT(PGD)		**0.00**		**0.00**
MIM		**0.00**		13.95
UAPG-IIoT(MIM)		**0.00**		**0.00**
EUAPG-IIoT(MIM)		**0.00**		**0.00**
CW		**0.00**		10.06
UAPG-IIoT(CW)		**0.00**		**0.00**
EUAPG-IIoT(CW)		**0.00**		**0.00**
DP		1.22		1.66
UAPG-IIoT(DP)		**0.00**		**0.00**
EUAPG-IIoT(DP)		**0.00**		**0.00**
UAPAM		2.47		31.19

ML-based IDSs. The substitute NNs for EUAPG-IIoT are also trained on the test set using the predicted labels from the target classifiers. The functional features of CMRI and MPCI are shown in Table X [30]. For UAPG-IIoT and EUAPG-IIoT, the $\boldsymbol{\omega}$ is set to the indexes of the nonfunctional features not in Table X. The settings for $\boldsymbol{\psi}$ and r are the same as those in Section V-C. The bold font is also used to highlight the best results of each ADR column in the subsequent experimental result presentation.

The results of adversarial attacks against NN-based IDSs in DAL are shown in Tables XI and XII. Compared to the target attack methods, the proposed methods exhibit superior attack performance when the modifiable features are not limited. The detection rate of the NN-based IDS is decreased to zero by the proposed attack methods on both CMRI and MPCI traffic. Likewise, after only modifying the nonfunctional features, the proposed methods still show strong attack performance. Most target attack methods show better performance after being combined with the proposed methods. The FGSM and MIM also show excellent attack performance against the NN-based IDS for CMRI. The proposed methods still exhibit better attack performance against NN-based IDSs in DAL than that of the UAPAM in both attack settings.

Likewise, the transferability of AEs across different ML-based IDSs in DAL is assessed by applying the AEs generated on the NN-based IDS to the other ML-based IDSs. The corresponding evaluation results of the transferability are shown in Tables XIII and XIV. Whether restricting the number of modifiable features or not, the universal adversarial perturbations produced by the proposed methods exhibit superior attack performance against most ML-based IDSs on both CMRI and MPCI traffic except for the NB-based IDS

TABLE XII
ADVERSARIAL ATTACKS AGAINST NN-BASED IDSS IN DAL AFTER RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	CMRI		MPCI	
	ODR(%)	ADR(%)	ODR(%)	ADR(%)
FGSM-NF	98.78	**0.00**	98.34	73.84
UAPG-IIoT-NF(FGSM)		**0.00**		**0.00**
EUAPG-IIoT-NF(FGSM)		0.02		**0.00**
PGD-NF		0.96		48.34
UAPG-IIoT-NF(PGD)		**0.00**		**0.00**
EUAPG-IIoT-NF(PGD)		**0.00**		**0.00**
MIM-NF		**0.00**		14.74
UAPG-IIoT-NF(MIM)		**0.00**		**0.00**
EUAPG-IIoT-NF(MIM)		0.02		**0.00**
CW-NF		0.41		25.94
UAPG-IIoT-NF(CW)		**0.00**		**0.00**
EUAPG-IIoT-NF(CW)		**0.00**		**0.00**
DP-NF		0.81		1.66
UAPG-IIoT-NF(DP)		**0.00**		**0.00**
EUAPG-IIoT-NF(DP)		**0.00**		**0.00**
UAPAM-NF		3.03		31.01

for MPCI. After limiting the number of modifiable features, the degradation of the attack performance of the target attack methods can be observed from these two tables, especially on ML-based IDSs for MPCI. On the contrary, the target attack methods combined with the proposed methods retain similar attack performance. The UAPAM obtains limited attack effect on ML-based IDSs for MPCI. The transferability of the universal adversarial perturbations produced by the UAPAM is also not ideal on the DT-based IDS for CMRI and the RF-based IDS for CMRI. Like the results in Tables VIII and IX, the presence of certain outliers can be observed in Tables XIII and XIV. For example, the adversarial perturbations produced by CW instead improve the detection performance of DT-based IDS against adversarial CMRI traffic. Besides, as evidenced by comparing the results in the same position of Tables XIII and XIV, restricting the number of modifiable features instead improves the attack performance of FGSM against LR-based IDS for MPCI. Another intriguing phenomenon is that restricting the number of modifiable features improves the attack performance of most attack methods against NB-based IDS for CMRI. The NB-based IDS has achieved a 100% detection rate in detecting adversarial MPCI traffic constructed by different attack methods.

E. Adversarial Attacks Against IIoT With Cloud/Edge Configuration

In this section, the edge-IIoTset data set is utilized to further validate the effectiveness of the proposed attack methods in IIoT with cloud/edge configuration. Similarly, the data set is partitioned into training and test sets in a 7:3 ratio. The ten-fold cross-validation and grid search are used on the training set to obtain the best performance classifiers. The test set is further divided into two parts in a 9:1 ratio. The 10% part is utilized as the seed set by the universal adversarial attack methods (i.e., UAPG-IIoT, EUAPG-IIoT, and UAPAM) to craft the universal adversarial perturbations. The 90% part is utilized to assess the adversarial attack methods and target ML-based IDSs. The test set, in conjunction with the predicted labels generated by the target ML-based IDSs, is utilized for training the substitute NNs. Table XV illustrates the functional

TABLE XIII
TRANSFERABILITY OF THE AEs ACROSS DIFFERENT ML-BASED IDSs IN DAL WITHOUT RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	SVM(%)			LR(%)			DT(%)			NB(%)			RF(%)		
	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR
FGSM		**0.00**		14.39		**0.00**		14.26		94.86		95.76		61.88		**100.00**		96.82		90.24
UAPG-IIoT(FGSM)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		9.97		98.76		**100.00**		**0.00**		5.34
EUAPG-IIoT(FGSM)		**0.00**		**0.00**		**0.00**		**0.00**		95.33		9.97		**0.00**		**100.00**		97.44		**0.00**
PGD		62.54		70.47		0.15		45.28		92.52		96.11		64.78		**100.00**		61.60		95.93
UAPG-IIoT(PGD)		98.76		**0.00**		**0.00**		**0.00**		94.16		10.10		98.76		**100.00**		88.70		5.42
EUAPG-IIoT(PGD)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		98.76		**100.00**		**0.00**		**0.00**
MIM		**0.00**		45.41		**0.00**		14.22		95.20		95.58		21.60		**100.00**		86.23		91.56
UAPG-IIoT(MIM)	98.78	**0.00**	98.34	**0.00**	98.76	**0.00**	98.34	**0.00**	91.51	**0.00**	96.33	9.97	98.76	98.76	98.16	**100.00**	91.94	**0.00**	96.41	**0.00**
EUAPG-IIoT(MIM)		**0.00**		**0.00**		**0.00**		**0.00**		94.16		9.89		**0.00**		**100.00**		96.82		**0.00**
CW		98.76		69.25		**0.00**		26.73		92.35		68.55		100.00		**100.00**		97.33		68.50
UAPG-IIoT(CW)		**0.00**		**0.00**		**0.00**		**0.00**		100.00		0.39		**0.00**		**100.00**		64.46		**0.00**
EUAPG-IIoT(CW)		**0.00**		**0.00**		**0.00**		**0.00**		100.00		0.39		**0.00**		**100.00**		75.10		**0.00**
DP		**0.00**		5.77		**0.00**		2.14		99.08		6.78		63.09		**100.00**		74.71		4.72
UAPG-IIoT(DP)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		0.13		**0.00**		**100.00**		**0.00**		**0.00**
EUAPG-IIoT(DP)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		0.13		**0.00**		**100.00**		**0.00**		**0.00**
UAPAM		**0.00**		98.34		**0.00**		14.22		95.33		96.11		**0.00**		**100.00**		97.44		96.28

TABLE XIV
TRANSFERABILITY OF THE AEs ACROSS DIFFERENT ML-BASED IDSs IN DAL AFTER RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	SVM(%)			LR(%)			DT(%)			NB(%)			RF(%)		
	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR	CMRI ODR	CMRI ADR	MPCI ODR	MPCI ADR
FGSM-NF		**0.00**		14.39		**0.00**		13.95		94.86		95.76		**0.00**		**100.00**		96.82		95.32
UAPG-IIoT-NF(FGSM)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		9.97		**0.00**		**100.00**		**0.00**		5.34
EUAPG-IIoT-NF(FGSM)		**0.00**		**0.00**		**0.00**		**0.00**		95.33		9.97		**0.00**		**100.00**		97.44		**0.00**
PGD-NF		65.03		84.16		0.15		45.89		92.52		96.11		64.78		**100.00**		61.60		96.11
UAPG-IIoT-NF(PGD)		98.76		**0.00**		**0.00**		**0.00**		94.16		10.10		98.76		**100.00**		88.70		5.42
EUAPG-IIoT-NF(PGD)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		98.76		**100.00**		**0.00**		**0.00**
MIM-NF		**0.00**		88.85		**0.00**		13.95		95.20		95.58		**0.00**		**100.00**		86.23		95.49
UAPG-IIoT-NF(MIM)	98.78	**0.00**	98.34	**0.00**	98.76	**0.00**	98.34	**0.00**	91.51	**0.00**	96.33	9.97	98.76	**0.00**	98.16	**100.00**	91.94	**0.00**	96.41	**0.00**
EUAPG-IIoT-NF(MIM)		**0.00**		**0.00**		**0.00**		**0.00**		94.16		9.89		**0.00**		**100.00**		96.82		**0.00**
CW-NF		98.76		69.25		**0.00**		26.82		92.35		68.55		**0.00**		**100.00**		97.33		68.50
UAPG-IIoT-NF(CW)		**0.00**		**0.00**		**0.00**		**0.00**		100.00		0.39		**0.00**		**100.00**		64.46		**0.00**
EUAPG-IIoT-NF(CW)		**0.00**		**0.00**		**0.00**		**0.00**		100.00		0.39		**0.00**		**100.00**		75.10		**0.00**
DP-NF		**0.00**		5.56		**0.00**		2.14		99.08		6.78		0.02		**100.00**		74.71		4.72
UAPG-IIoT-NF(DP)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		0.13		**0.00**		**100.00**		**0.00**		**0.00**
EUAPG-IIoT-NF(DP)		**0.00**		**0.00**		**0.00**		**0.00**		**0.00**		0.13		**0.00**		**100.00**		**0.00**		**0.00**
UAPAM-NF		**0.00**		98.34		**0.00**		13.95		95.33		96.11		**0.00**		**100.00**		97.44		96.28

TABLE XV
FUNCTIONAL FEATURES FOR BACKDOOR AND RANSOMWARE ATTACK TRAFFIC IN THE EDGE-IIOTSET DATA SET

Attacks	Functional Features
Backdoor	arp.hw.size, icmp.seq_le, tcp.ack, tcp.ack_raw, tcp.connection.rst, tcp.flags, tcp.seq, udp.time_delta, dns.qry.name, dns.qry.qu, dns.qry.type, mqtt.conflags, mqtt.hdrflags, mqtt.proto_len
Ransomware	icmp.checksum, tcp.ack, tcp.connection.synack, tcp.len, dns.retransmission, dns.retransmit_request_in, mqtt.len, mqtt.msg_decoded_as, mqtt.msgtype, mqtt.proto_len, mbtcp.unit_id

features of the backdoor and ransomware attack traffic. For UAPG-IIoT and EUAPG-IIoT, the $\boldsymbol{\omega}$ is set to the indexes of the nonfunctional features not in Table XV. The other parameter settings for the adversarial attack methods are the same as those in Section V-C. In the subsequent experimental result presentation, the best experimental results in each ADR column of all the tables are boldly represented for convenient comparison.

The experimental results of adversarial attacks against NN-based IDSs in IIoT with cloud/edge configuration are illustrated in Tables XVI and XVII. It can be observed that the baseline attack methods (FGSM, PGD, MIM, and CW) achieve similar attack performance to the proposed attack methods on both backdoor and ransomware traffic (they all diminish the detection rates of the target NN-based IDS to zero) when the modifiable features are not limited. However, the proposed attack methods demonstrate their superiority over the baseline attack methods when the modifiable features are restricted. The restriction of the number of modifiable features imposes more influence on the attack performance of the baseline attack methods than on that of the proposed methods (the detection rates of the NN-based IDS against AEs constructed by these attack methods increase more significantly). The combination of the proposed attack methods with the target attack methods can effectively improve the performance of the original attack methods (e.g., the UAPG-IIoT(DP) and EUAPG-IIoT(DP) outperform the DP in Table XVI). It can be also observed that the limitation of the number of modifiable features exhibits a more significant impact on CW-related and DP-related attack methods in the context of adversarial backdoor attacks (the detection rates of the target IDS against CW-related and DP-related attack methods recover to the original level). Compared to the UAPAM, the universal adversarial perturbations produced by the proposed attack methods have exhibited better attack performance in Tables XVI and XVII.

The experimental results of the transferability of the produced AEs across different ML-based IDSs in IIoT with cloud/edge configuration are shown in Tables XVIII and XIX. The evaluation results in these two tables are also achieved by applying the AEs generated on the NN-based IDS to ML-based IDSs in the tables directly. It can be observed from these two tables that the combination of the proposed attack methods with the target attack methods can effectively improve the

TABLE XVI
ADVERSARIAL ATTACKS AGAINST NN-BASED IDSS IN IIOT WITH CLOUD/EDGE CONFIGURATION BEFORE RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	Backdoor		Ransomware	
	ODR(%)	ADR(%)	ODR(%)	ADR(%)
FGSM	92.96	**0.00**	91.13	**0.00**
UAPG-IIoT(FGSM)		**0.00**		**0.00**
EUAPG-IIoT(FGSM)		**0.00**		**0.00**
PGD		**0.00**		**0.00**
UAPG-IIoT(PGD)		**0.00**		**0.00**
EUAPG-IIoT(PGD)		**0.00**		**0.00**
MIM		**0.00**		**0.00**
UAPG-IIoT(MIM)		**0.00**		**0.00**
EUAPG-IIoT(MIM)		**0.00**		**0.00**
CW		**0.00**		**0.00**
UAPG-IIoT(CW)		**0.00**		**0.00**
EUAPG-IIoT(CW)		**0.00**		**0.00**
DP		54.00		66.86
UAPG-IIoT(DP)		**0.00**		**0.00**
EUAPG-IIoT(DP)		**0.00**		**0.00**
UAPAM		50.98		5.96

TABLE XVII
ADVERSARIAL ATTACKS AGAINST NN-BASED IDSS IN IIOT WITH CLOUD/EDGE CONFIGURATION AFTER RESTRICTING THE NUMBER OF MODIFIABLE FEATURES

Attack Method	Backdoor		Ransomware	
	ODR(%)	ADR(%)	ODR(%)	ADR(%)
FGSM-NF	92.96	21.57	91.13	**0.00**
UAPG-IIoT-NF(FGSM)		2.53		**0.00**
EUAPG-IIoT-NF(FGSM)		**0.00**		0.07
PGD-NF		42.59		10.17
UAPG-IIoT-NF(PGD)		3.07		**0.00**
EUAPG-IIoT-NF(PGD)		2.43		**0.00**
MIM-NF		1.64		**0.00**
UAPG-IIoT-NF(MIM)		**0.00**		**0.00**
EUAPG-IIoT-NF(MIM)		1.47		**0.00**
CW-NF		92.96		48.48
UAPG-IIoT-NF(CW)		92.97		**0.00**
EUAPG-IIoT-NF(CW)		91.91		**0.00**
DP-NF		92.96		91.13
UAPG-IIoT-NF(DP)		93.21		**0.00**
EUAPG-IIoT-NF(DP)		92.81		**0.00**
UAPAM-NF		89.53		12.26

transferability of AEs in both attack settings (with or without limiting the number of modifiable features). For instance, the AEs generated by UAPG-IIoT(DP) (ADR 6.16%) and EUAPG-IIoT(DP) (ADR 0.00%) exhibit better transferability on RF-based IDS than those produced by DP (ADR 18.01%) when the modifiable features are not restricted in the context of adversarial Ransomware attacks, as shown in Table XVIII. In Table XIX, the AEs generated by UAPG-IIoT-NF(FGSM) (ADR 28.11%) and EUAPG-IIoT-NF(FGSM) (ADR 7.14%) still exhibit better transferability on SVM-based IDS than those produced by FGSM-NF (ADR 46.83%) when the modifiable features are restricted in the context of adversarial backdoor attacks. Besides, it can be observed that the proposed attack methods obtain the optimal ADRs most frequently on all the target ML-based IDSs for both backdoor and ransomware attack traffic. Similarly, some outliers are present in these two tables. For example, the adversarial perturbations generated by MIM on the NN-based IDS instead improve the detection rate of the DT-based IDS against adversarial backdoor attacks, as illustrated in Table XVIII. The limitation of the number of modifiable features may improve the attack performance of some adversarial attack methods. This can be

TABLE XVIII
TRANSFERABILITY OF THE AEs ACROSS DIFFERENT ML-BASED IDSs IN IIoT WITH CLOUD/EDGE CONFIGURATION BEFORE RESTRICTING THE NUMBER OF MODIFIABLE FEATURES (THE BD IN THIS TABLE IS SHORT FOR BACKDOOR AND THE RW IS SHORT FOR RANSOMWARE)

Attack Method	SVM(%)				LR(%)				DT(%)				NB(%)				RF(%)			
	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR
FGSM		**0.00**		**0.00**		**0.00**		**0.00**		100.00		99.73		**0.00**		0.03		**0.00**		**0.00**
UAPG-IIoT(FGSM)		**0.00**		0.07		**0.00**		**0.00**		97.90		56.56		**0.00**		0.07		**0.00**		**0.00**
EUAPG-IIoT(FGSM)		**0.00**		6.88		**0.00**		**0.00**		53.73		96.54		**0.00**		0.07		**0.00**		**0.00**
PGD		27.70		21.40		3.45		3.18		69.48		65.05		0.01		0.07		0.25		0.21
UAPG-IIoT(PGD)		0.74		1.51		**0.00**		**0.00**		51.68		**0.00**		**0.00**		0.07		**0.00**		**0.00**
EUAPG-IIoT(PGD)		**0.00**		0.07		**0.00**		**0.00**		80.90		83.74		**0.00**		**0.00**		**0.00**		**0.00**
MIM		0.14		0.14		**0.00**		**0.00**		99.97		85.86		**0.00**		0.07		**0.00**		**0.00**
UAPG-IIoT(MIM)	96.41	**0.00**	81.65	0.07	97.88	**0.00**	86.55	**0.00**	98.05	97.74	97.26	56.56	100.00	**0.00**	100.00	0.03	98.03	**0.00**	97.26	**0.00**
EUAPG-IIoT(MIM)		**0.00**		0.07		**0.00**		**0.00**		97.74		56.56		**0.00**		0.03		**0.00**		**0.00**
CW		90.40		78.43		97.88		86.55		97.85		86.99		98.21		96.92		50.43		40.81
UAPG-IIoT(CW)		2.85		5.10		48.07		50.74		65.45		53.65		**0.00**		0.07		4.55		2.74
EUAPG-IIoT(CW)		0.28		0.68		**0.00**		**0.00**		**1.03**		1.03		**0.00**		0.07		**0.00**		**0.00**
DP		96.12		81.51		94.26		80.97		94.66		71.14		41.64		23.86		14.63		18.01
UAPG-IIoT(DP)		2.94		5.96		48.07		50.74		86.58		78.06		**0.00**		100.00		5.12		6.16
EUAPG-IIoT(DP)		1.98		0.68		48.07		**0.00**		98.06		0.75		**0.00**		0.07		1.21		**0.00**
UAPAM		49.98		5.55		48.07		50.74		100.00		97.30		100.00		0.07		56.04		**0.00**

TABLE XIX
TRANSFERABILITY OF THE AEs ACROSS DIFFERENT ML-BASED IDSs IN IIoT WITH CLOUD/EDGE CONFIGURATION AFTER RESTRICTING THE NUMBER OF MODIFIABLE FEATURES (THE BD IN THIS TABLE IS SHORT FOR BACKDOOR AND THE RW IS SHORT FOR RANSOMWARE)

Attack Method	SVM(%)				LR(%)				DT(%)				NB(%)				RF(%)			
	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR	BD ODR	BD ADR	RW ODR	RW ADR
FGSM-NF		46.83		**0.00**		10.78		**0.00**		98.24		76.75		**0.00**		0.07		81.28		**0.00**
UAPG-IIoT-NF(FGSM)		28.11		0.07		**0.00**		**0.00**		97.88		**52.69**		**0.00**		0.07		94.54		1.47
EUAPG-IIoT-NF(FGSM)		7.14		20.99		**0.00**		**0.00**		100.00		96.54		**0.00**		0.07		49.06		**0.00**
PGD-NF		60.68		17.32		27.11		19.92		92.64		68.61		0.06		0.27		54.12		5.96
UAPG-IIoT-NF(PGD)		75.74		1.16		0.50		**0.00**		97.91		53.85		**0.00**		0.07		69.95		2.64
EUAPG-IIoT-NF(PGD)		**4.76**		0.07		**0.00**		**0.00**		97.73		83.74		**0.00**		**0.03**		**0.12**		**0.00**
MIM-NF		33.67		**0.00**		0.26		**0.00**		97.40		80.83		**0.00**		0.07		65.30		**0.00**
UAPG-IIoT-NF(MIM)	96.41	41.99	81.65	0.07	97.88	**0.00**	86.55	**0.00**	98.05	97.73	97.26	83.64	100.00	**0.00**	100.00	0.07	98.03	93.32	97.26	**0.00**
EUAPG-IIoT-NF(MIM)		46.20		0.07		**0.00**		**0.00**		97.73		83.43		**0.00**		0.07		93.32		**0.00**
CW-NF		96.41		78.43		97.88		86.55		97.87		84.25		98.49		96.92		98.03		40.71
UAPG-IIoT-NF(CW)		96.41		5.10		97.88		50.74		98.24		53.65		100.00		0.07		98.24		8.01
EUAPG-IIoT-NF(CW)		96.37		0.68		48.07		**0.00**		98.06		53.51		**0.00**		0.07		97.40		**0.00**
DP-NF		96.41		81.51		97.47		81.31		98.17		81.27		44.52		32.18		98.27		41.53
UAPG-IIoT-NF(DP)		96.23		5.48		94.26		50.74		**51.99**		53.89		**0.00**		100.00		98.17		6.13
EUAPG-IIoT-NF(DP)		96.41		0.68		70.29		**0.00**		98.05		53.61		**0.00**		0.07		97.35		**0.00**
UAPAM-NF		94.97		6.30		48.07		50.74		100.00		53.78		100.00		0.07		98.53		**0.00**

concluded by comparing the evaluation results of the FGSM-based adversarial Ransomware traffic against DT-based IDS in Tables XVIII and XIX.

F. Experimental Analysis

Based on the preliminary experimental results above, the following analysis and conclusions can be drawn.

1) The ML-based IDSs in IIoT are vulnerable to AEs. While the modification to the original input space is normally limited, the universal adversarial perturbations are feasible in IIoT. The widely adopted ML algorithms in IIoT are all vulnerable to universal adversarial perturbations even when the modifiable features are restricted to guarantee the validity of the attack traffic. Once the universal adversarial perturbations are produced, they can be applied to different original inputs that are not involved in their generation to attack different ML classifiers. Consequently, the proposed attack methods improve the efficiency of producing AEs in IIoT. Besides, the work in this article reveals the relationship between the vulnerabilities of diverse ML algorithms on the high-dimensional decision boundaries under the same data distribution.
2) The proposed attack methods, i.e., UAPG-IIoT and EUAPG-IIoT, extract the original input-dependent adversarial perturbations from gradient-based or optimization-based attack methods to construct the input-independent universal adversarial perturbations with better performance and transferability. The experimental results in Tables VI–XIX demonstrate the effectiveness and generalizability of the proposed attack methods against diverse ML-based IDSs across various real-world scenarios (i.e., the management network of IIoT, the traditional industrial control network of IIoT, and the IIoT with cloud/edge configuration). Even when the modifiable features are restricted to ensure the validity of the attack traffic, the AEs crafted by the proposed attack methods still show better performance and transferability against various ML-based IDSs on different data sets than the baseline attack methods. In the real world, attacks against ML-IDSs in IIoT can normally be conducted only by altering the nonfunctional features of the attack traffic. Consequently, the proposed attack methods are more suitable for evaluating the ML-based IDSs in IIoT.
3) When assessing the transferability of the constructed AEs, intriguing phenomena can be observed from the experimental results in Tables VIII–XIX. When the AEs generated on the NN-based IDSs are applied to other ML-based IDSs, there is a reversal of the attack effect on some ML classifiers.
    a) The transference of the adversarial perturbations instead improves the detection performance of the target ML-based IDSs. For example, in Table XIII, the original detection rate of the DT-based IDS for CMRI is 91.51%. The AEs crafted by CW on the NN-based IDS for CMRI increase the detection rate of the DT-based IDS to 92.35%. Similar phenomena can also be observed in Table IX. The original detection rate of the LR-based IDS for Probe is 98.19%. The universal adversarial perturbations crafted by UAPG-IIoT-NF(CW) increase the detection rate of the LR-based IDS to 99.10%. In Table XVIII, the AEs generated by FGSM increase the detection rate of the DT-based IDS against adversarial backdoor traffic to 100.00% (the ODR is 98.05%).
    b) After restricting the number of modifiable features, the transference of the adversarial perturbations improves the attack performance of the attack methods. For example, by comparing the ADRs of NB-based IDS for CMRI in Tables XIII and XIV, most ADRs are diminished to zero after limiting the number of modifiable features. By comparing the ADRs of DT-based IDS against adversarial Ransomware traffic constructed by FGSM-related attack methods in Tables XVIII and XIX, it can be observed that the restriction on the number of modifiable features improves the attack performance of the corresponding target attack methods (the ADRs against FGSM and UAPG-IIoT(FGSM) before restriction is 99.73% and 56.56%, respectively. The ADR against FGSM-NF and UAPG-IIoT-NF(FGSM) after restriction diminishes to 76.75% and 52.69%, respectively).

    This goes against the existing findings. Restricting the modifiable features generally implies a limitation on the manipulable space available to the adversary. Therefore, the attack performance of the attack methods will normally degrade when the modifiable features are limited. The preliminary and intuitive explanation for these findings is that the same features in the original inputs possess different meanings for different ML classifiers. Specifically, certain features exhibit positive effects on the detection performance of specific ML classifiers while demonstrating negative impacts on the detection performance of other ML classifiers. This explains why the transference of the same feature modification to another ML classifier improves its detection rate (corresponding to a) in this paragraph). Similarly, when the positive features are restricted, the transference of the feature modification can undermine the detection performance and improve the attack effect (corresponding to b) in this paragraph). We believe that these findings are important for further exploration of the cause and the fix of adversarial attacks. Further research on these findings will be left to our future work.
4) Although the experimental results demonstrate that the proposed attack methods surpass the baseline methods, the attack performance and transferability of the proposed methods on some target ML classifiers need to be further improved. For example, the UAPG-IIoT and EUAPG-IIoT perform not well on RF-based IDS for Probe and LR-based IDS for Probe when the modifiable features are limited. The attack performance

and transferability of universal adversarial perturbations constructed by UAPG-IIoT and EUAPG-IIoT on NB-based IDS for MPCI need to be enhanced. After limiting the modifiable features, the attack performance of the proposed methods combined with CW or DP in some scenarios can be further enhanced. For example, the attack performance of adversarial probe traffic produced by the UAPG-IIoT-NF(CW) and EUAPG-IIoT-NF(DP) against NN-based IDS can be further improved. The attack effect of adversarial backdoor traffic produced by the EUAPG-IIoT-NF(CW) and EUAPG-IIoT-NF(DP) against NN-based IDS can be further explored. The preliminary analysis for this is that the CW and DP is originally designed to produce smaller adversarial perturbations than FGSM, PGD, and MIM. When the modifiable features are limited, the degradation of perturbation strength of the CW and DP is more significant than FGSM, PGD, and MIM. Consequently, the limitation of the number of modifiable features imposes a bigger impact on the attack performance of CW-related or DP-related attack methods. Further exploration of this phenomenon will be left to our future work.

5) The universal adversarial perturbations produced by the proposed attack methods demonstrate their superiority to those crafted by the mainstream universal adversarial perturbation attack methods (namely, UAPAM) in most scenarios of IIoT in the evaluation experiments. The experimental results in this article indicates that the proposed attack methods are more suitable for assessing the robustness of the ML-based IDSs in IIoT against adversarial attacks compared to the UAPAM originally designed for computer vision. Another advantage of the proposed attack methods is that the proposed methods are based on a flexible framework that masks the underlying details of the target attack methods, which makes the proposed methods able to be combined with different gradient-based or optimization-based attack methods. This supplies more choices for specific scenarios. Instead, the UAPAM can only utilize DP to construct the universal adversarial perturbations, which leads to the limitation of the attack effect of the UAPAM in some scenarios.
6) As shown in Figs. 5 and 6, the quantity of seed samples does not significantly impact the attack performance of UAPG-IIoT and EUAPG-IIoT. A low-sampling rate is enough to generate high-quality universal adversarial perturbations for the proposed attack methods. This is because the termination of the generation process occurs when the produced universal adversarial perturbations in an iteration satisfy the requirement of the threshold r, as shown in line 29 of Algorithm 1. Accordingly, not every seed sample will be used during the generation of universal adversarial perturbations. This mechanism ensures the feasibility of the proposed attack methods in the real world due to the difficult collection of massive attack samples, which means that the proposed attack methods only need to collect a small high-quality seed set to conduct an effective evaluation of the target classifier in the real world.
7) While they have been demonstrated to be effective in producing high-quality universal adversarial perturbations against ML-based IDSs in IIoT, the proposed methods can only be combined with the glass-box adversarial attack methods that utilize the gradient information of original inputs to generate adversarial perturbations. This is because the proposed methods extract gradient-based adversarial perturbations from original AEs to construct the universal adversarial perturbations. This limitation implies that the proposed attack methods can not be combined with closed-box adversarial attack methods, such as those based on generative adversarial networks or heuristic optimization methods. Consequently, the proposed attack methods can not be directly applied in a closed-box attack. Generally, they can only conduct closed-box attacks with the assistance of the substitute model technology (which may influence their attack effect in some closed-box scenarios).

VI. Conclusion

This article demonstrates that the ML-based IDSs in IIoT are vulnerable to universal adversarial perturbations. Although universal adversarial perturbations are normally susceptible to restrictions on the number of modifiable features, universal adversarial perturbations produced by the proposed attack methods are demonstrated to be feasible in IIoT and show better attack performance and transferability than the mainstream adversarial attack methods. The findings in this article reveal the relationship between the vulnerabilities of different ML algorithms on the high-dimensional decision boundaries under the same data distribution, which can facilitate the advancement of artificial intelligence security research.

While they have been demonstrated to be effective in producing high-quality universal adversarial perturbations against ML-based IDSs in IIoT, the proposed methods are limited to being combined with glass-box attack methods, which may influence their attack effect in some closed-box scenarios. In our future work, we will further focus on the problems that occur in this article. In particular, we will pay more attention to researching the meaning variation of the same feature across diverse types of ML classifiers to further explore the cause and the fix of the adversarial attacks and improve the attack performance and transferability of the proposed methods on some ML classifiers (e.g., RF and LR).

References

[1] K. Tange, M. De Donno, X. Fafoutis, and N. Dragoni, "A systematic survey of Industrial Internet of Things security: Requirements and fog computing opportunities," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2489–2520, 4th Quart., 2020.

[2] N. Bugshan, I. Khalil, M. S. Rahman, M. Atiquzzaman, X. Yi, and S. Badsha, "Toward trustworthy and privacy-preserving federated deep learning service framework for Industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1535–1547, Feb. 2023.

[3] M. Abdel-Basset, V. Chang, H. Hawash, R. K. Chakrabortty, and M. Ryan, "Deep-IFS: Intrusion detection approach for Industrial Internet of Things traffic in fog environment," *IEEE Trans. Ind. Informat.*, vol. 17, no. 11, pp. 7704–7715, Nov. 2021.

[4] A. Telikani, J. Shen, J. Yang, and P. Wang, "Industrial IoT intrusion detection via evolutionary cost-sensitive learning and fog computing," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 23260–23271, Nov. 2022.

[5] R. F. Mansour, "Blockchain assisted clustering with intrusion detection system for Industrial Internet of Things environment," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422012209

[6] Y. Li et al., "Robust detection for network intrusion of industrial IoT based on multi-CNN fusion," *Measurement*, vol. 154, Mar. 2020, Art. no. 107450. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026322411931317X

[7] H. Gu, Y. Lai, Y. Wang, J. Liu, M. Sun, and B. Mao, "DEIDS: A novel intrusion detection system for industrial control systems," *Neural Comput. Appl.*, vol. 34, no. 12, pp. 9793–9811, Jun. 2022. [Online]. Available: https://doi.org/10.1007/s00521-022-06965-4

[8] J. Ahmad, S. A. Shah, S. Latif, F. Ahmed, Z. Zou, and N. Pitropakis, "DRaNN_PSO: A deep random neural network with particle swarm optimization for intrusion detection in the Industrial Internet of Things," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 10, pp. 8112–8121, Nov. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157822002701

[9] C. Szegedy et al., "Intriguing properties of neural networks," 2014, *arXiv:1312.6199*.

[10] J. Chen, X. Gao, R. Deng, Y. He, C. Fang, and P. Cheng, "Generating adversarial examples against machine learning-based intrusion detector in industrial control systems," *IEEE Trans. Depend. Secure Comput.*, vol. 19, no. 3, pp. 1810–1825, May 2022.

[11] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, and M. Qiu, "Adversarial attacks against network intrusion detection in IoT systems," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10327–10335, Jul. 2021.

[12] B. Esmaeili, A. Azmoodeh, A. Dehghantanha, H. Karimipour, B. Zolfaghari, and M. Hammoudeh, "IIoT deep malware threat hunting: From adversarial example detection to adversarial scenario detection," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 8477–8486, Dec. 2022.

[13] L. Zeng, D. Qiu, and M. Sun, "Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks," *Appl. Energy*, vol. 324, Oct. 2022, Art. no. 119688. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261922009850

[14] O. Gungor, T. Rosing, and B. Aksanli, "STEWART: STacking ensemble for white-box adversaRial attacks towards more resilient data-driven predictive maintenance," *Comput. Ind.*, vol. 140, Sep. 2022, Art. no. 103660. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0166361522000574

[15] N. Martins, J. M. Cruz, T. Cruz, and P. Henriques Abreu, "Adversarial machine learning applied to intrusion and malware scenarios: A systematic review," *IEEE Access*, vol. 8, pp. 35403–35419, 2020.

[16] H. Mohammadian, A. A. Ghorbani, and A. H. Lashkari, "A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems," *Appl. Soft Comput.*, vol. 137, Apr. 2023, Art. no. 110173. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623001916

[17] E. Alhajjar, P. Maxwell, and N. Bastian, "Adversarial machine learning in network intrusion detection systems," *Expert Syst. Appl.*, vol. 186, Dec. 2021, Art. no. 115782. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421011507

[18] N. Wang, Y. Chen, Y. Xiao, Y. Hu, W. Lou, and Y. T. Hou, "MANDA: On adversarial example detection for network intrusion detection system," *IEEE Trans. Depend. Secure Comput.*, vol. 20, no. 2, pp. 1139–1153, Apr. 2023.

[19] Z. Lin, Y. Shi, and Z. Xue, "IDSGAN: Generative adversarial networks for attack generation against intrusion detection," in *Proc. 26th Pacific-Asia Conf. Knowl. Discov. Data Min.*, 2022, pp. 79–91.

[20] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 86–94.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[23] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9185–9193.

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2574–2582.

[25] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy (SP)*, 2017, pp. 39–57.

[26] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, New York, NY, USA, 2017, pp. 506–519. [Online]. Available: https://dl.acm.org/doi/10.1145/3052973.3053009

[27] W. Wang, R. Wang, L. Wang, Z. Wang, and A. Ye, "Towards a robust deep neural network against adversarial texts: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 3159–3179, Mar. 2023.

[28] H. Kim, J. Park, and J. Lee, "Generating transferable adversarial examples for speech classification," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109286. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322007658

[29] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Security Defense Appl.*, 2009, pp. 1–6.

[30] T. Morris and W. Gao, "Industrial control system traffic data sets for intrusion detection research," in *Critical Infrastructure Protection VIII* (IFIP Advances in Information and Communication Technology), vol. 441, J. Butts and S. Shenoi, Eds., Berlin, Germany: Springer, 2014, pp. 65–78.

[31] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for Centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022.

[32] I. A. Khan, D. Pi, Z. U. Khan, Y. Hussain, and A. Nawaz, "HML-IDS: A hybrid-multilevel anomaly prediction approach for intrusion detection in SCADA systems," *IEEE Access*, vol. 7, pp. 89507–89521, 2019.

[33] K.-D. Lu, G.-Q. Zeng, X. Luo, J. Weng, W. Luo, and Y. Wu, "Evolutionary deep belief network for cyber-attack detection in industrial automation and control system," *IEEE Trans. Ind. Informat.*, vol. 17, no. 11, pp. 7618–7627, Nov. 2021.

[34] A. Alsaedi, Z. Tari, R. Mahmud, N. Moustafa, A. Mahmood, and A. Anwar, "USMD: UnSupervised misbehaviour detection for multi-sensor data," *IEEE Trans. Depend. Secure Comput.*, vol. 20, no. 1, pp. 724–739, Jan. 2023.

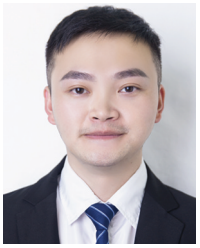
[35] M. Catillo, A. Pecchia, and U. Villano, "CPS-GUARD: Intrusion detection for cyber-physical systems and IoT devices using outlier-aware deep autoencoders," *Comput. Secur.*, vol. 129, Jun. 2023, Art. no. 103210. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404823001207

[36] Z. Chen et al., "Machine learning-enabled IoT security: Open issues and challenges under advanced persistent threats," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–37, Dec. 2022. [Online]. Available: https://doi.org/10.1145/3530812

[37] O. Friha, M. A. Ferrag, M. Benbouzid, T. Berghout, B. Kantarci, and K.-K. R. Choo, "2DF-IDS: Decentralized and differentially private federated learning-based intrusion detection system for industrial IoT," *Comput. Secur.*, vol. 127, Apr. 2023, Art. no. 103097. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740482300007X

[38] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[39] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.



Sicong Zhang was born in Chongqing, China. He received the B.E. degree in electrical engineering and automation from Civil Aviation University of China, Tianjin, China, in 2011, the M.E. degree in computer science and technology from Guizhou Normal University, Guiyang, China, in 2016, and the Ph.D. degree in software engineering from Guizhou University, Guiyang, in 2020.

He is currently a Lecturer and a Postgraduate Supervisor with the School of Cyber Science and Technology, the Key Laboratory of Information and Computing Science Guizhou Province, Guizhou Normal University. His research interests include cyber-security and artificial intelligence.

Yang Xu was born in Shandong, China. He received the Ph.D. degree in computer software and theory from Guizhou University, Guiyang, China, in 2010.

He is currently a Professor and a Postdoctoral Supervisor with the School of Cyber Science and Technology, the Key Laboratory of Information and Computing Science Guizhou Province, Guizhou Normal University. His research interests include cyber-security and artificial intelligence.

Prof. Xu is a Senior Member of the China Computer Federation.

Xiaoyao Xie (Member, IEEE) was born in Guizhou, China. He received the Ph.D. degree in computer application technology from Wuhan University, Wuhan, China, in 2004.

He is currently a Professor and the Ph.D. Supervisor with the School of Cyber Science and Technology, the Key Laboratory of Information and Computing Science Guizhou Province, Guizhou Normal University, Guiyang, China. He is also the Director of the Key Laboratory. His research interests include artificial intelligence, 5G, and IPV6.