A Robust Approach for Privacy Data Protection: IoT Security Assurance Using Generative Adversarial Imitation Learning

Chenxi Huang, Sirui Chen, Yaqing Zhang, Wen Zhou, *Member, IEEE*, Joel J. P. C. Rodrigues, *Fellow, IEEE*, and Victor Hugo C. de Albuquerque, *Senior Member, IEEE*

***Abstract*—With the increasing importance of data security, privacy protection has gradually risen to a strategic position, especially IoT data privacy protection. The concern for data security has become a national strategy. The discovery of potential risks of privacy data is of great significance, such as the risk of data privacy leakage, data security vulnerabilities, etc. In this article, starting from the privacy data protection mechanism in the Industrial Internet of Things (IIoT) scenario, we proposed a method based on generative adversarial imitation learning (GAIL) to discover the privacy data security risks in IIoT by training privacy protection agents using a large amount of expert data on privacy protection. Finally, our proposed method is validated by relevant simulation experiments, and the results show that our proposed method has wide generalizability and reliability to obtain the maximum payoff of the agents and thus, reduce the risk of data security leakage.**

***Index Terms*—Generative adversarial imitation learning (GAIL), Industrial Internet of Things (IIoT), policy optimization, private data protection.**

I. Introduction

THE INDUSTRIAL Internet is a highly integrated global industrial system with advanced computing, analytics, and sensing technologies and the Internet. Through the connection of intelligent machines and ultimately humans and machines, combined with software and big data analytics, it will reshape the global industry, stimulate productivity, and make the world faster, safer, cleaner, and more economical.

Harnessing the massive amounts of data generated by smart devices is an important function of the Industrial Internet. Leveraging capabilities, such as big data, complex analytics, and predictive algorithms, the Industrial Internet provides a way to understand the massive amounts of data generated by smart devices and can help select, analyze, and utilize this data to bring benefits, such as network optimization, maintenance optimization, system recovery, autonomous machine learning, and intelligent decision making, ultimately helping the industrial sector reduce costs, save energy, and drive productivity improvements.

At present, the security issue has become an important factor that hinders the further development of Industrial Internet of Things (IIoT). If its security cannot be fully guaranteed, personal information, commercial secrets, and military secrets in the IIoT system may be stolen or used by unscrupulous elements, which will certainly seriously affect personal privacy, economic security, military security, and national security. The research of IIoT security technology mainly includes the following four aspects: 1) IIoT security architecture; 2) network security protocol of IIoT; 3) network security protection technology; and 4) cryptography and its application in IIoT.

In order to provide ubiquitous personalized services to users, the system of IIoT needs to use their personal information without their awareness or disturbance through automatic sensing functions. For example, in an intelligent medical care system, the user's physiological signs data (heart rate, temperature, blood pressure, etc.) need to be collected in real time. In the IIoT environment, the process of using personal information covers the whole life cycle of users' personal data, including its perception, storage, transmission, and application, and user privacy issues mainly occur in the perception and application phases of these four processes. Data perception has the characteristics of invisibility and wide coverage, and it is a system behavior, and since the perceived personal information is private to the user, the user requires privacy protection for this process; in addition, the application processing with the goal of service is essentially a process in which personal information is shared by other entities interacting with the system, and the information is not controllable to the individual, and the user also requires privacy protection for this process. In these cases, if the system privacy protection

Manuscript received 30 August 2021; revised 22 October 2021; accepted 3 November 2021. Date of publication 16 November 2021; date of current version 7 September 2022. This work was supported in part by FCT/MCTES through national funds and co-funded EU funds under Project UIDB/50008/2020; in part by the Brazilian National Council for Scientific and Technological Development–CNPq under Grant 313036/2020-9; in part by the National Natural Science Foundation of China under Grant 61902003; and in part by the Doctoral Scientific Research Foundation of Anhui Normal University. *(Corresponding author: Wen Zhou.)*

Chenxi Huang and Yaqing Zhang are with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: supermonkeyxi@xmu.edu.cn; 24320172203255@stu.xmu.edu.cn).

Sirui Chen is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: imchen_doris@163.com).

Wen Zhou is with the School of Computer and Information, Anhui Normal University, Wuhu 241000, China (e-mail: zhouwen327@gmail.com).

Joel J. P. C. Rodrigues is with the Fecomeércio - IPDC & Senac Faculty of Ceará, Fortaleza 60125-100, Brazil, also with the Instituto de Telecomunicações, 6201-001 Covilhã, Portugal, and also with the Federal university of Ceará, Fortaleza, Brazil (e-mail: joeljr@ieee.org).

Victor Hugo C. de Albuquerque is with the Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza 60811-905, Brazil (e-mail: victor.albuquerque@ieee.org).

Digital Object Identifier 10.1109/JIOT.2021.3128531

mechanism is missing, the user's private information will be potentially threatened, which raises the issue of privacy protection in IIoT.

There have been quite a few studies on the security and privacy protection issues of IIoT.

1) Among the IIoT security mechanisms, the research on key management and authentication technologies for IIoT is relatively mature. The current key management and authentication methods mainly adopt two kinds: a) the centralized management method with the Internet as the core [1] and b) the distributed management method with the respective heterogeneous networks as the center.
2) For the node capture problem, there is still little research on privacy that put leakage technology for IIoT anti-node capture, but the node anti-capture attack in ad hoc networks can be studied, thus indirectly playing a role in the protection of privacy data [2].
3) RFID, one of the most important supporting technologies for IIoT, will cause leakage of personal privacy data when personal information is combined with tag information within a certain distance. The National Institute of Standards and Technology (NIST) released a standard guideline for smart tags in 2007; this standard guideline recommends the use of firewalls to isolate RFID databases from other IT systems and databases, and the report recommends that users use auditing, logging, and timestamping methods to detect security threats and establish tag handling and reuse procedures to update devices and destroy old data.
4) For anti-node capture attack techniques in wireless mobile communication networks, most of them have been implemented using detection and traditional encryption and authentication techniques [3], [4]. All the above are traditional encryption, authentication, and detection techniques, but this is far from enough, to prevent the leakage of their private information after the capture of IIoT nodes, and anti-copying and necessary tracking techniques are the key issues to be considered.
5) Privacy homomorphism techniques can be widely applied in several fields, including data privacy security protection for IIoT. Privacy homomorphism techniques were proposed by Brickell and Yacobi in 1978 [5] as cryptographic transformations that allow direct manipulation of ciphertexts. Domingo and Ferrer [6], [7] proposed two algebraic privacy homomorphism approaches for known plaintext attacks.
6) A.C. Yao proposed the concept of secure multiparty computation (SMC) [8] in 1982, and the SMC technique is one of the hot spots of research in the international cryptography community and one of the important security techniques in IIoT privacy protection.
7) Protecting the location privacy of users is one of the important issues of IIoT applications, and the data processing process in IIoT involves location-based services and data privacy protection in the information processing process. Information was established in 2008 by ACM to work on the theory and application of spatial information [9].

The international academic community has published papers on the privacy protection of IIoT, and some important research results have been published in IEEE series of conferences and ACM series of conferences. The research on data privacy protection technology in IIoT has also received attention and focus from domestic academia, which is mainly focused on the research based on data distortion or data encryption technology. However, there are few studies on how to train privacy-preserving agents to discover the privacy data security risks in industrial IIoT.

Generative adversarial imitation learning (GAIL) is a kind of inverse reinforcement learning method that combines the ideas of the generative adversarial network (GAN). It is a backward reinforcement learning method that combines the ideas of GAN. GAIL is characterized by a GAN framework for solving imitation learning problems, in which the training process of the discriminator is analogous to the learning process of the reward function and the training process of the generator is analogous to the learning process of the policy. Compared with traditional imitation learning methods, GAIL has better robustness, characterization ability, and computational efficiency. Therefore, it is capable of handling complex large-scale problems and can be extended to practical applications. An intelligent body is a very important concept in the field of artificial intelligence. Any independent entity that can think and interact with its environment can be abstracted as an intelligent body. The concept of an intelligent body was introduced by Minsky, a leading computer scientist at MIT and one of the founders of the discipline of artificial intelligence, who introduced the concept of society and social behavior to computing systems in his book "Society of Mind." There has been work combining machine learning with privacy preservation, and Papernot *et al.* [10] proposed the PATE algorithm (private aggregation of teacher ensembles, PATE) in 2018, which will be useful for researchers who know how to train supervised machine learning models, and differential privacy for machine learning. The PATE framework implements privacy learning by carefully coordinating the activities of several different machine learning models, and as long as the procedures specified in the PATE framework are followed, the resulting models are privacy preserving.

The remainder of this article is structured as follows. Section I introduces the background of this article. Section II describes the relevant work of this article. Section III gives a detailed description of our proposed method. Section IV describes our experiments and the results, and finally, Section V makes a summary of the experiments.

II. Related Work

With the improvement of the computer hardware manufacturing process and cost reduction, the computational power and storage resources available for machine learning algorithms are becoming more and more abundant. Among the many machine learning algorithms, deep learning was the most significant one to enhance the learning effect with the help of hardware resources [11]. Deep learning has been surprisingly valuable in various fields due to its excellent automated

feature representation characteristics. The success of classical deep learning algorithms relies on a large number of learning samples. Often, training with the same deep learning model as the goal requires the use of sample data generated or stored in multiple sources. Machine intelligence requires people to apply large amounts of data to train machines to make them smarter. Among them, generative algorithms are often considered as a measure of how well the machine "understands" the training data. Generative algorithms generate information distributions by learning a set of information and then mapping the high-dimensional rich perceptron inputs to category labels through discriminative models based on backpropagation and Dropout algorithms in the field of deep learning [12], as a way to discriminate the correctness of the distribution generated by the generation algorithm, and if the generated distribution is correct, then the machine must have also understood the information correctly.

However, many intractable probabilistic computational problems in maximum-likelihood estimation and correlation strategies in general, and the inability to take full advantage of segmented linear units in generating contexts, have led to the lack of widespread adoption of deep generative models. However, the research aiming at generating or reproducing samples indistinguishable from real samples remains a hot topic in statistical signal processing and machine learning. In particular, obtaining generative models for high-dimensional data distributions is a challenging but important task because of their importance for various applications.

Imitation learning based on GANs (GANs-IL), developed from IRL-IL, is a class of imitation learning methods that incorporate GANs [13]. The main difference between the two is the reward function, the representation model of the strategy, and the training method of the model. GANs-IL uses two neural networks to represent the reward function and the strategy in IRL-IL and optimizes the parameters of these two networks in an adversarial way.

GANs were first proposed by Ian Goodfellow [14] and inspired by the two-person zero-sum game in the game theory. GAN models are composed of a generative model and a discriminative model. The generative model [15] refers to a model that can produce the desired sample output, which learns the probability distribution of real data, input random noise, and transforms it into a picture [16] or speech [17], etc., that is close to the distribution of real data, while the discriminative model was usually a two-class classification network, where the input was real data and the generated samples from the generative model, and the output was the judgment of the probability that the input comes from real data probability.

The earliest and most representative GANs-IL method is GAIL proposed by Ho and Ermon [13] in 2016. If the policy was characterized as a generative model from state input to action output, then the process of imitation learning was to learn a policy based on expert samples, which is the training process of the generative model. In GAIL, the policy that outputs actions based on input states can be analogous to a generator, and the reward function that outputs reward values based on input expert samples or generative samples can be analogous to a discriminator. Thereby, GAIL analogizes the process of solving the reward function to the training process of the discriminator and the learning process of the strategy to the training process of the generator

Deep learning enables the deployment of end-to-end learning systems that use multiple nonlinear feature transformations, i.e., processing layers consisting of a multilayer perceptron (MLP), to learn representations of the data. It is this high abstraction that makes the parameters and intermediate results in it not easy to understand and analyze, and the high degree of model fitting to the data will make the model parameters and detailed prediction results retain more data features, which are the source of privacy threats.

The academic research on privacy began in the 1960s [18]. In data privacy protection, it is extremely important to ensure the balance between availability and privacy of the data set. GANs take advantage of themselves by adding noise to the latent space instead of directly to the data, reducing the overall information loss while ensuring privacy. Huang *et al.* [19] proposed a context-aware privacy model in combination with GANs by subtly adding noise to achieve private data release. Meanwhile, Triastcyn and Faltings [20] proposed a method for generating artificial data sets by adding a Gaussian noise layer to the discriminator of GANs so that the output and gradient have different privacy concerning the training data, and then synthesizing artificial data sets with confidentiality using the generator component, which not only preserves the statistical properties of real data but also provides differential privacy for these data protection. Papernot *et al.* [21] proposed a privacy-preserving approach for deep networks in the teacher–student model, using a deep model for teachers and a model for student GANs, by training and thus, protecting the training data set. Frigerio *et al.* [22] proposed a privacy-preserving data publishing framework by differential privacy definition, from time series to continuous data and discrete data generation, all of which can be easily adapted to different use cases to ensure that the user's personality is protected while publishing new open data.

Although such a deep learning training model has tremendous advantages, it also introduces new data security issues. For example, when multiple hospitals need to jointly train a deep learning model, the deep learning algorithm requires a large amount of data as training samples. Each hospital needs to share its private medical data to complete the training of a global deep learning model with higher accuracy [23]. However, sharing medical data as such would violate patient privacy, and medical information has always been confidential to hospitals out of respect for patients. The confidentiality of training data prevents joint deep learning algorithms from being applied in such situations. Specifically, actually, in aforementioned methods, in general, a patient's records are used to train some machine learning model; furthermore, this model usually is used to determine the appropriate drug dose for a disease or to discover the genetic basis of the disease. In this way, the attending doctors can be sufficient to conclude whether the patient has the disease. Due to the widespread use of machine learning as a service, privacy attacks under black boxes will be more relevant. Fredrikson *et al.* [24] initiated model inversion attacks (MIAs) against deep models,

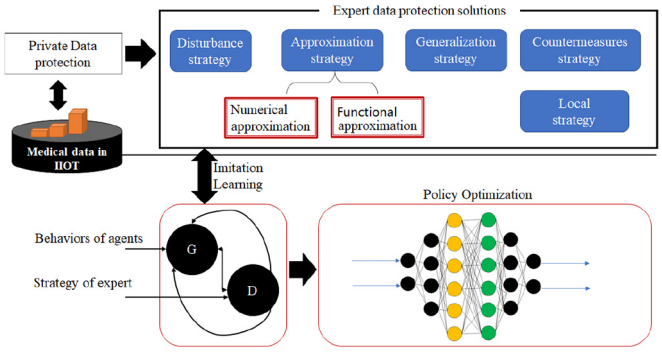

Fig. 1. Architecture of private medical data protection in IIoT scenario.

using the output of the model to infer certain features of the input data. In one of the models with patient data as training data [24], the correlation between drug doses and patient genes was pointed out, but it was also argued that this attack did not result in a privacy breach due to the inherent medical facts of both. Jones *et al.* [25] combined GANs and differential privacy to propose a differential privacy-assisted classification generation adversarial network for generating medical clinical data.

III. Methods

In this article, we put forward a framework based on the protection of medical big data in the IIOT scenario as shown in Fig. 1. The framework mainly consists of three parts; the existing private data are processed by the expert data protection solutions module and input to GANs for training and learning, and GANs generate a large amount of data through training and input to the policy optimization module for training. Since privacy data protection is oriented to different scenarios that require different methods, our network can automatically select specific methods according to the scenarios and finally, obtain the optimal data protection policy.

A. Privacy Data Protection Technology

There are five main privacy data protection techniques we use, which are: 1) disturbance strategy; 2) approximation strategy; 3) generalization strategy; 4) countermeasures strategy; and 5) local strategy, as shown in Fig. 3. There is a certain coupling relationship between these different techniques, and these techniques are introduced separately next.

1) Disturbance Strategy: Data disturbance techniques are effective solutions adopted in the privacy-preserving microdata publishing. These strategies interfere with the original data through data transformation and generalization, and afterward mine the interfered data to obtain the desired patterns and rules. We write the data disturbance function as

$$\hat{X} = g(X, Y) \tag{1}$$

where X is the original data, Y is the added disturbance, and g denotes the disturbance function.

Since we perform data disturbance to prevent the attacks instead of corrupting the data, we still need to recover the perturbed data to the original data. Hence, we obtain the recovered data as

$$\hat{Y} = h(g(X, Y)) \tag{2}$$

where h is the recovery function. Now, combining the recovery function h and the disturbance Y, the loss function becomes

$$L(h, g) = \mathbb{E}\big[l(h(g(X, Y)), Y)\big]. \tag{3}$$

In addition, to make the obtained data more robust, a data disturbance optimization function is introduced

$$\min_{g(\cdot)} \max_{h(\cdot)} L(h, g). \tag{4}$$

The original data then after disturbance and recovery will certainly have a certain loss, because the original data cannot be destroyed, to control the disturbance within a certain range, so the definition of D as a disturbance adjustment parameter as

$$\mathbb{E}\big[d(g(X, Y), X)\big] \le D. \tag{5}$$

2) Approximation Strategy: The approximation strategy is mainly divided into numerical and function approximations. The numerical approximation method is used to construct a simple function $g(x)$ to approximate or replace the original function $f(x)$ using a function table of discrete data. Algebraic interpolation is to find an interpolation function to approximate the target function in the presence of the target function. The main quadratic interpolation methods are the Newton form of interpolation, the Lagrange form, the form of successive linear interpolation, etc. The main methods of function approximation are polynomial approximation and connected fractional approximation. The transcendental functions in mathematics, such as $\exp(x)$, $\ln(x)$, and $\sin(x)$, are often calculated by Taylor series expansion, which is to use polynomial to approximate the function.

3) Generalization Strategy: The generalization strategy is mainly divided into generalized output and generalized model.

We can use a function to approximate the function. We only need to store the key coefficients of the function, which greatly reduces the storage volume.

The generalized linear model (GLM) is an extension of the linear model that establishes the relationship between the mathematical expectations of the response variables and the linear combination of the predictor variables through a linking function.

4) Countermeasures Strategy: The countermeasures strategy is mainly divided into counter disturbance and regularization.

During data transmission, the data code needs to be made more resistant to interference because attenuation or interference can cause abrupt changes in the data code. This must be done by adding a few bits of binary code length to the original binary code length so that the corresponding data have a certain degree of redundancy.

Regularization is achieved by adding a regularization term to the original loss function. By attaching some rules, i.e., constraints, to the model parameters, the model is prevented from overfitting the training data.

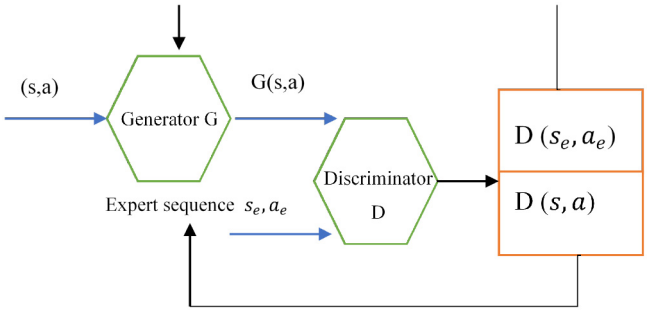

Fig. 2. Imitation learning process.

5) Local Strategy: Local strategy is mainly divided into tag integration, associative learning, and secure multiparty.

Tag integration is a way to convert a multitag classification problem into a problem constructed from a single-tag model and then merge the models.

We can implement associative learning using the priori algorithm and the FP-Growth algorithm. The priori algorithm uses a bottom-up approach where frequent subsets are expanded one object at a time (a step known as candidate set generation) and the candidate sets are examined by the data. The FP-Growth algorithm recursively divides the transactional data set into multiple smaller conditional transactional data sets to mine the frequent item sets.

The SMC method uses cryptographic techniques and secure computation technology to obtain global data mining results through collaborative distributed computation for multiple entities involved in association rule mining in a distributed environment. We can base on secure dot product cryptography protocols for vertically partitioned distributed privacy-preserving association rule mining. Alternatively, we can use secure merge operation for vertically partitioned distributed privacy-preserving association rule mining.

B. Generative Adversarial Networks Generates Sample of Simulation Experts

Imitation learning simulates the learning goal of obtaining reward values for state-behavior (s, a) sequences (where s is the state and a is the behavior) by learning from existing expert data, thus allowing an intelligent body to automatically identify data privacy security risks.

$D(s, a)$ is the output of the discriminant model, which represents the probability that the input (s, a) is real data. $G(s, a)$ is the output of the generative model, and the output is fake data.

However, the direction of the gradient change is to be changed for the prediction results of the discriminant model. The gradient update direction is to be changed when the discriminant model considers the output of $G(s, a)$ as the real data set and when it considers the output as noisy data.

The process is represented as shown in Fig. 2.

For the generative model, what we want to do is to make the data generated by $G(s, a)$ as much as possible the same as the data in the data set. It is the so-called same data distribution. Then, what we want to do is to minimize the error of the generative model, i.e., to pass only the error generated by

Algorithm 1 Training Process of GANs

for number of training iterations **do**
  for ***m*** steps **do**
    Update the generator by descending its stochastic gradient:

$$\underset{c\in C}{Max}\left(\underset{\pi\in\prod}{min}\, E_\pi[c(s,a)] - E_\pi\left[-\log\pi(a|s)\right]\right) - E_{\pi_E}[c(s,a)]$$

  end for
  Update the discriminator by ascending its stochastic gradient:

$$\widehat{E}_{\tau_i}\left[\nabla_w \log(D_w(s,a))\right] + \widehat{E}_{\tau_e}\left[\nabla_w \log(1 - D_w(s,a))\right]$$

end for

$G(s, a)$ to the generative model. This leads to

$$E_\pi[c(s,a)] = E\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \tag{6}$$

where the terms c and γ are represented as the state–action Q function and the discounted factor, respectively, besides, $s_0 : p_0$, $a_t : \prod(g|s_t)$, $s_(t+1) : P(g|s_t, a_t)$, $c(s, a)$: obtains the obtained Q values for the sequence of state behaviors.

During the training process, the parameters of one of the models are generally fixed and the other model is updated. We first update the generative model G. When fixing the discriminative model D and updating the generative model G, the discriminative model should make a wrong judgment for the false sample (s, a) generated by the generative model G, and it is difficult to distinguish whether the input data are true or not. The above game process is summarized into the objective function that can be obtained from (7), which is the cost function optimization function Ψ

$$\underset{c\in C}{\text{Max}}\left(\min_{\pi\in\prod} E_\pi[c(s,a)] - E_\pi\left[-\log\pi(a|s)\right]\right) - E_{\pi_E}[c(s,a)]. \tag{7}$$

The training is continuously compared in (7) to update the parameters, thus getting closer to the expected value for the network. Such a training process is less time consuming and more robust, which improves the relevance and learning motivation more highly. After that, we fix the generative model G to update the discriminative model D. The output probability given by the discriminative model should be close to 0 for samples from the true distribution x, and close to 1 for samples from the true distribution. The discriminant model should give an output probability close to 0 for samples from the generative model G and close to 1 for samples x from the true distribution, i.e., the discriminant model should be able to correctly discriminate between true and false samples. The cross-entropy function as shown in (8) is the parameter update function of the discriminator D. For a sequence of sample trajectories $\tau_i, \ldots, \tau_n$

$$\widehat{E}_{\tau_i}\left[\nabla_w \log(D_w(s,a))\right] + \widehat{E}_{\tau_e}\left[\nabla_w \log(1 - D_w(s,a))\right]. \tag{8}$$

The parameters of the recognizer network are updated. After that, a large number of samples can be obtained to stimulate the policy optimization function afterward. The method is shown in Algorithm 1. Specifically, the input and output of

the proposed method is the numerous training policy data and the policy network with reasonable parameters, respectively.

C. Strategy Optimization Training Network to Get the Optimal Strategy

The method of using policy gradients to optimize the policy model is called the policy gradient algorithm. To ensure that the overall effect of the model becomes stronger, for all the boosting integration algorithms, the effect of the integrated model will be better than before with each evaluator constructed. That is, as the iteration proceeds, the overall effect of the model must be gradually improved, and finally, the optimal effect of the integrated model must be achieved. We define the gradient evaluator as shown in

$$\hat{g} = \hat{E}_t\Big[\nabla_t \log \pi_\theta(a_t|s_t)\hat{A}_t\Big]. \tag{9}$$

The dominance function $\hat{A}_t$ is defined as shown in (12). The objective function based on the trust region is expressed as shown in

$$\max_\theta \hat{E}_t = \left[\frac{\pi_\theta(a_\theta|s_t)}{\pi_{\theta_{\text{old}}}(a_\theta|s_t)}\hat{A}_t\right]. \tag{10}$$

Subject to

$$\hat{E}_t = \Big[KL\Big[\pi_{\theta_{\text{old}}}(\bullet|s_t), \pi_\theta(\bullet|s_t)\Big]\Big] \le \delta \tag{11}$$

where KL denotes the KL divergence, which is used to measure the difference or distance between these two distributions. The term t is in the range $[0, T]$ at time timestep. Furthermore, the parameter δ is an experimental value to constrain the KL distance; in this article, we set this term $\delta = 0.1$

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t-1} + \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T). \tag{12}$$

Equation (12) is a Markovian time series, where t is the current time, r_t is the reward at time point $t-1$, T is the reward of the state after a period of time, and γ is the discount rate, and V is the summation of the state behavior values. After domain trimming of the dominance function, (7) can be transformed as follows:

$$\hat{A}_t = \delta_t + (\gamma\delta_{t-1}) + \cdots + \gamma^{T-t+1}\delta_{T-1} \tag{13}$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_T). \tag{14}$$

Equations (5) and (6) can be converted into the linear equation shown in (15), and the final optimal strategy can be solved after fitting in the following equation:

$$L(\theta) = \hat{E}_t\left[\frac{\pi_\theta(a_\theta|s_t)}{\pi_{\theta_{\text{old}}}(a_\theta|s_t)}\hat{A}_t - \beta KL\Big[\pi_{\theta_{\text{old}}}(\bullet|s_t), \pi_\theta(\bullet|s_t)\Big]\right]. \tag{15}$$

In summary, the method is shown in Algorithm 2. Additionally, the input and output of proposed method are policy trajectories data and the optimal policy, respectively.

Finally, the training process of our private medical data protection scheme in the IIOT scenario can be represented by the flowchart shown in Fig. 3.

Algorithm 2 Select Optimal Stratagy

for number of training iterations **do**
  Run policy $\boldsymbol{\pi_\theta}$ for **T** timesteps, collecting $\{\boldsymbol{S_t, a_t, r_t}\}$
  Estimate advantages $\hat{A}_t = -V(s_t) + r_t + \gamma r_{t-1} + \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T)$.
  $\pi_{old} \leftarrow \pi_\theta$
  for m steps **do**

$$L(\theta) = \hat{E}_t\left[\frac{\pi_\theta(a_\theta|s_t)}{\pi_{\theta_{old}}(a_\theta|s_t)}\hat{A}_t - \beta KL\Big[\pi_{\theta_{old}}(\bullet|s_t), \pi_\theta(\bullet|s_t)\Big]\right].$$

    Update θ by a gradient method w.r.t $L(\theta)$
  end for
  if $\mathrm{KL}[\pi_{old}|\pi_\theta] > \beta_{high}\mathrm{KL}_{target}$ **then**
    $\lambda \leftarrow \alpha\lambda$
    elseif $\mathrm{KL}[\pi_{old}|\pi_\theta] > \beta_{low}\mathrm{KL}_{target}$ **then**
    $\lambda \leftarrow \lambda/\alpha$
  end if
end for

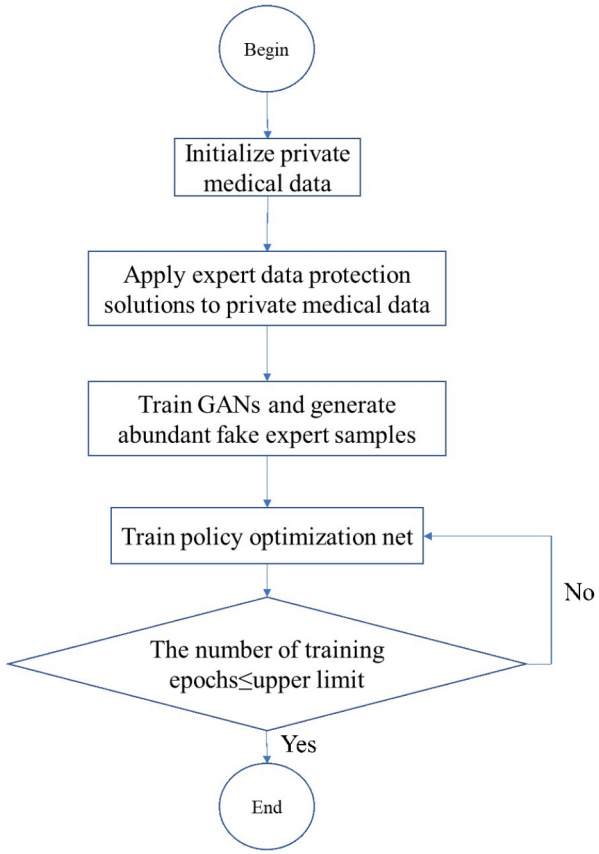

Fig. 3. Training process of our scheme.

IV. Experiments

A. Environment

The approach presented in this chapter is implemented in the Tensorflow 2.0 framework. The experiments in this chapter are completed on Ubuntu systems. For the experiment, the processor used is Intel Core i7 3.5 GHZ, 8-GB RAM, and the graphics card is NVIDIA GeForce GTX 1080, visualized using the TensorBoard framework.

We mainly focus on Cumulative Reward, LOSS, and Policy for quantitative analysis, where LOSS indicators are GAIL Loss, Policy Loss, Pretraining Loss, and Value Loss; Policy indicators are Beta, Entropy, Epsilon, GAIL Expert Estimate,

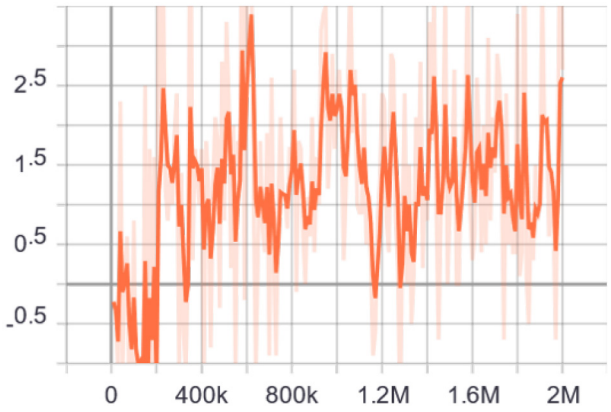

Fig. 4. Cumulative Reward metric in training stage.

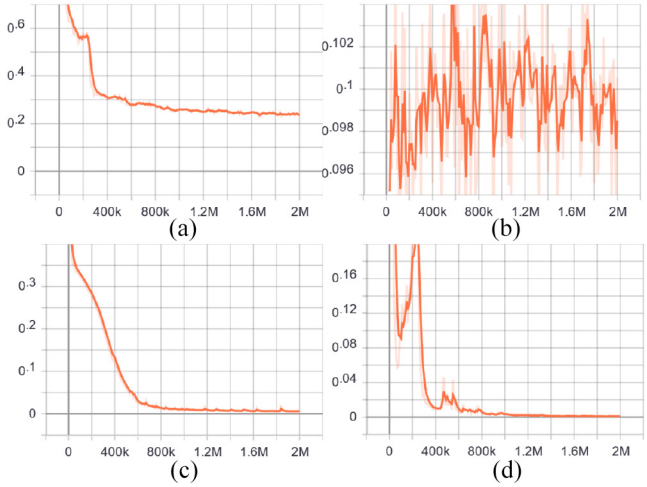

Fig. 5. Four different loss metrics in training stage. (a) GAIL loss. (b) Policy loss. (c) Pretraining loss. (d) Value loss.

GAIL Grad Mag Loss, GAIL Policy Estimate, Gail Reward, Gail Value Estimate, Learning Rate, and Total Score. The training of our experiments was conducted after a total of 200 sessions. The following series of images show the results of our experiments, which have been trained for a total of 2 million cycles.

B. Results Analysis

Cumulative Reward: This scalar is a good or bad metric, and the ultimate goal of the agent is to maximize the cumulative reward of the whole process as much as possible. It can be seen from Fig. 4 that as the training proceeds, the cumulative reward for strategy optimization oscillates continuously, reaching a higher value of about 2.5 at the end of the training.

From Fig. 5(a), we see that the GAIL loss gradually decreases as the number of training rounds increases, indicating that the loss of the data in each state is small, proving that the gap between the real data and the fake data is already small and our GANs network training scheme works well. Policy Loss [Fig. 5(b)] can be used to evaluate the goodness of the action chosen by the actor, and then guide the actor to make a better choice next time. It oscillates up and down during the training process, and then gradually converges to around 0.099.

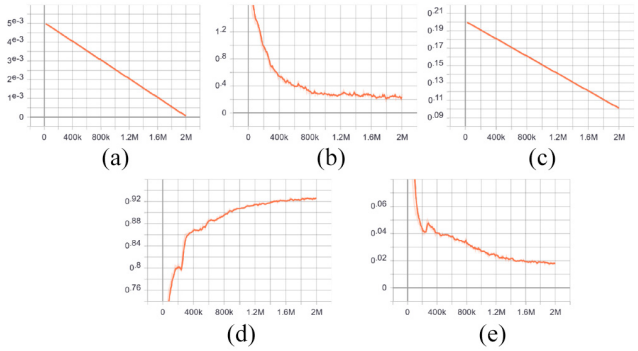

Fig. 6. During training epoches, our proposed method is evaluated on five different indicators, i.e., (a) Beta, (b) Entropy, (c) Epsilon, (d) GMAIL expert estimate, and (e) GAIL grad mag loss. The result is shown as above.

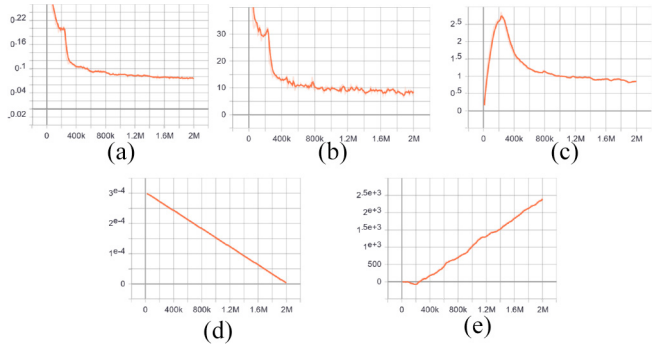

Fig. 7. Five different policy metrics in training stage. (a) GAIL policy estimate. (b) GAIL reward. (c) GAIL value estimate. (d) Learning rate. (e) Total score.

Pretraining Loss: This metric tests the generalization ability of the network. A pretrained network has a smaller error and is more robust as the depth of the network increases. From Fig. 5(c), we can see that Pretraining Loss gradually decreases as the training progresses and finally converges to 0. The value loss [Fig. 5(d)] is the loss of the fake data generated by GANs, which also decreases gradually as the training progresses and finally converges to 0. After the training, we get the smallest value loss, which corresponds to the output of the action with the largest value.

According to the objective function-constraint formula, it is seen from Fig. 6(a) that Beta becomes smaller and smaller as the training proceeds. The change in Beta value indicates that the constraint becomes smaller and smaller, which proves that the generated data gradually satisfy the condition. *Entropy:* The coefficient of the entropy loss term is a very important hyperparameter, which has a direct impact on the convergence speed and final performance. As the training proceeds, the variance of the action distribution of the policy output becomes smaller and smaller, which is reflected in the statistical index that the entropy becomes smaller and smaller. A reasonable entropy coefficient ensures that the model is fully explored early in the training to move in the right direction, and also allows the model to fully utilize the learned skills later in the training to achieve high performance. It can also be seen from Fig. 6(b) that the entropy of our experiments decreases gradually with increasing training rounds, as expected. *Epsilon:* This metric is the threshold value for the

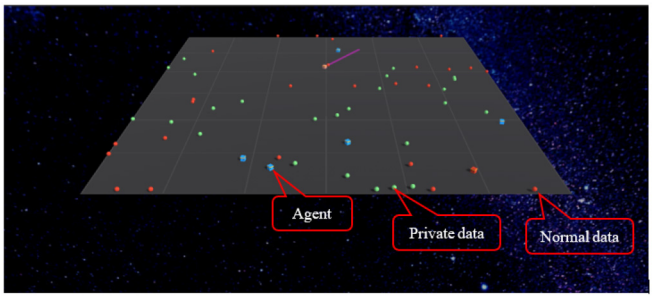

Fig. 8. Strategy optimization network visualization simulation platform.

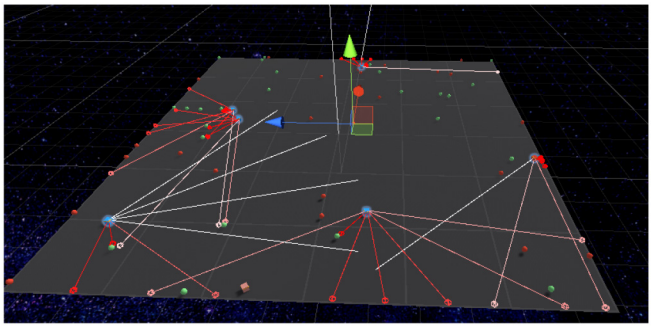

Fig. 9. Policy optimization network discovers privacy data security risks in industrial IoT.

choice of exploring or exploiting. From Fig. 6(c), it can be seen that the threshold decreases as the training proceeds, which indicates that the generated data gradually meet the conditions and the network automatically selects the most desired action. The GAIL Expert Estimate was evaluated on the expert data, and as seen in Fig. 6(d), it gradually increased as the training progressed and finally stabilized, reaching a high value of about 0.93. This reflects that the policy optimization network learns different parameterized policies for each different expert and finally achieves the optimal policy. *GAIL Grad Mag Loss:* The gradient grad is determined by the loss function loss, and the larger the loss function is, the larger the gradient is. The parameters to be optimized are determined by the gradient grad and the learning rate together. After the parameters are updated, the loss function will also be reduced, thus further reducing the gradient until the loss function is minimized and the gradient is 0, at which time the optimal solution is obtained. From Fig. 6(e), the GAIL Grad Mag Loss gradually decreases as the training progresses and finally converges to around 0.02, indicating that the optimal strategy is obtained.

GAIL Policy Estimate: As shown in Fig. 7(a), this parameter gradually decreases with training and finally converges to about 0.08, indicating that the policy has minimized its JS dispersion from the expert policy. *Gail Reward:* GAIL's discriminator provides a reward for strategy learning, which is used to distinguish the expert strategy from the learned strategy, in the same way as the discriminator training in GAN. Fig. 7(b) reflects the gradual decrease of Reward, which shows that the expert strategies and the learned strategies are getting closer. *Gail Value Estimate:* The regional stability of the value function indicates that the network is generating fake data that are getting closer to the real data. From Fig. 7(c), it can be seen that although the indicator rises at the beginning of the training, it falls back as the training proceeds and finally, converges to around 0.8. *Learning Rate:* Fig. 7(d) reflects that the learning rate decreases as training proceeds, ensuring that the model does not fluctuate too much in the later stages of training and thus, gets closer to the optimal solution. *Total Score:* This metric reflects our evaluation of the network as a whole. From Fig. 7(e), we see that the total score of the strategies gets higher as the training progresses, indicating that we get the optimal strategy. Next is our simulation experiment platform for private data protection. A qualitative comparison of our strategies is presented below. Fig. 8 shows a demonstration of our simulated experimental platform, where blue represents the agents, green represents private data, i.e., private data that have been corrupted, and red represents normal data, i.e., private data that are protected and not corrupted. Below is a visualization of our experiment in 3-D space, where the red line indicates that the corrupted privacy data has been found, and the white line is a process of aimless searching, reflecting that the target has not been found yet. Fig. 9 illustrates that our strategy can find those corrupted data, but sometimes it also captures normal data in the process of exploration, which still exists with some errors.

C. Discussion

Through the above quantitative evaluation, we can observe that as the number of training rounds increases, all evaluation indicators of the GANs module and policy optimization module have stable and good results, which are in line with our expected judgment. Through the above qualitative evaluation, we can intuitively see the training process. The simulation platform results show that our solution provides a feasible privacy data protection strategy. In general, the experimental results show that our scheme is suitable for medical privacy data protection in the IIoT scenario. Our scheme is effective in terms of training time and maintains high accuracy. In addition, it also received a high-scoring learning strategy at the end.

V. Conclusion

In this article, based on the serious data leakage problem in IIOT scearios, we proposed an adversarial simulation learning GAIL-based approach. The approach aims to protect medical big data by training privacy-preserving agents to discover privacy data security risks in the industrial IoT. GANs first learns expert data processed by various expert data protection methods, and then inputs a large amount of fake data to the policy optimization network. The experiments show that the trained policy optimization network is able to choose the optimal privacy data protection policy according to different application scenarios. In the future, we will further try to increase the complexity of the network in the hope of obtaining more optimal privacy data protection strategies; meanwhile, we will also try to propose more optimal loss functions to reduce the training time.

References

[1] K. Wang, J. Bao, M. Wu, and W. Lu, "Research on security management for Internet of Things," in *Proc. ICCASM*, vol. 15, Oct. 2010, pp. 133–137.

[2] S. Radomirovic, "Towards a model for security and privacy in the Internet of Things," in *Proc. 1st Int. Workshop Security Internet of Things*, Dec. 2010, pp. 6–10.
[3] Y.-S. Choi and S. H. Shin, "A study on sensor node capture defense protocol for ubiquitous sensor network," in *Proc. ICCIT*, Dec. 2007, pp. 400–405.
[4] M. Conti, R. Pietro, L. Mancini, and A. Mei, "Mobility and cooperation to thwart node capture attacks in MANETs," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, p. 13, Dec. 2009.
[5] E. Brickell and Y. Yacobi, *On Privacy Homomorphisms* (Lecture Notes in Computer Science), vol. 304. Heidelberg, Germany: Springer, Jan. 1988, pp. 117–125.
[6] J. Domingo-Ferrer, *A Provably Secure Additive and Multiplicative Privacy Homomorphism**. Heidelberg, Germany: Springer, Sep. 2002.
[7] J. D. I. Ferrer, "New privacy homomorphism and applications," *Inf. Process. Lett.*, vol. 60, no. 5, pp. 277–282, Dec. 1996.
[8] A. Yao, "Protocols for secure computation," in *Proc. 23rd Annu. Symp. Found. Comput. Sci.*, Dec. 1982, pp. 160–164.
[9] K. B. Frikken and M. J. Atallah, "Privacy preserving electronic surveillance," in *Proc. WPES*, Jan. 2003, pp. 45–52.
[10] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with PATE," Feb. 2018, *arXiv:1802.08908*.
[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539
[12] H. Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. Courville, "Modulating early visual processing by language," Jul. 2017, *arXiv:1707.00683*.
[13] J. Ho and S. Ermon, "Generative adversarial imitation learning," Jun. 2016, *arXiv:1606.03476*.
[14] I. Goodfellow *et al.*, "Generative adversarial nets," Jun. 2014, *arXiv:1406.2661*.
[15] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Aug. 2006.
[16] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," May 2016, *arXiv:1605.05396*.
[17] A. Oord *et al.*, "WaveNet: A generative model for raw audio," Sep. 2016, *arXiv:1609.03499*.
[18] F. Kerlinger, *Foundations of Behavioral Research*. New Delhi, India: Sarjeet, 2003.
[19] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Generative adversarial privacy," 2019, *arXiv:1807.05306*.
[20] A. Triastcyn and B. Faltings, "Generating differentially private datasets using GANs," in *Proc. ICLR*, Mar. 2018, pp. 1–12.
[21] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," Oct. 2016, *arXiv:1610.05755*.
[22] L. Frigerio, A. Oliveira, L. Gomez, and P. Duverger, "Differentially private generative adversarial networks for time series, continuous, and discrete open data," in *Proc. SEC*, Jun. 2019, pp. 151–164.
[23] G. Litjens *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Sci. Rep.*, vol. 6, no. 1, May 2016, Art. no. 26286.
[24] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. USENIX Security Symp.*, Aug. 2014, pp. 17–32.
[25] B. B. Jones *et al.*, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation Cardiovasc. Qual. Outcomes*, vol. 12, no. 7, Jul. 2019, Art. no. e005122.

Chenxi Huang received the Ph.D. degree from the College of Electronics and Information Engineering, Tongji University, Shanghai, China, in 2019.

Since 2019, he has been with the School of Informatics, Xiamen University, Xiamen, China, where he is currently an Assistant Professor. His current research interests include deep learning, object detection, and medical image analysis.

Sirui Chen is currently pursuing the B.Sc. degree with Tongji University, Shanghai, China.

Her research interests include machine learning, image processing, and reconstruction.

Yaqing Zhang is currently pursuing the B.Sc. degree with Xiamen University, Xiamen, China.

Her research interests include machine learning, image processing, and reconstruction.

Wen Zhou (Member, IEEE) received the Ph.D. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2018.

Since 2018, he has been with the School of Computer and Information, Anhui Normal University, Wuhu, China, where he is currently a Lecturer. His research interests include WebVR visualization, virtual reality, sketch-based retrieval, and machine learning.

Dr. Zhou is a member of Chinese Computer Federation.

Joel J. P. C. Rodrigues (Fellow, IEEE) received the Ph.D. degree from the Departamento de Informática, Universidade da Beira Interior, Covilha, Portugal, in 2006.

He is with Senac Faculty of Ceará, Fortaleza, Brazil, Head of Research, Development, and Innovation; and a Senior Researcher with the Instituto de Telecomunicações, Covilhã, Portugal. He has authored or coauthored more than 1000 papers in refereed international journals and conferences, three books, two patents, and one ITU-T recommendation.

Mr. Rodrigues is the Editor-in-Chief of the *International Journal of E-Health and Medical Communications* and editorial board member of several journals. Top ranking for Computer Science in Brazil (Research.com). He is the Leader of the Next Generation Networks and Applications Research Group (CNPq), an IEEE Distinguished Lecturer, and a Member Representative of the IEEE Communications Society on the IEEE Biometrics Council. He was the Director for Conference Development–IEEE ComSoc Board of Governors, the Past-Chair of the IEEE ComSoc TCs on eHealth and on Communications Software, and a Steering Committee Member of the IEEE Life Sciences Technical Community. He is a member of the Internet Society and a Senior Member of ACM and AAIA .

Victor Hugo C. de Albuquerque (Senior Member, IEEE) received the B.S.E. degree in mechatronics engineering from the Federal Center of Technological Education of Ceará (CEFETCE), Fortaleza, Brazil, in 2006, the M.Sc. degree in teleinformatics engineering from the Department of Teleinformatics Engineering/Graduate Program in Teleinformatics Engineering (PPGETI), Federal University of Ceará (UFC), Fortaleza, Brazil, in 2007, and the Ph.D. degree in mechanical engineering from the Federal University of Paraíba, João Pessoa, Brazil, in 2010.

He is a Professor and Senior Researcher with the PPGETI, UFC. He specializes in image data science, IoT, machine/deep learning, pattern recognition, automation and control, and robotics.