基于安全联邦蒸馏 GAN 的工业 CPS 协作入侵检测系统

梁俊威[1]，杨耿[1]，马懋德[2]，Muhammad Sadiq[1]
（1. 深圳信息职业技术学院软件学院，广东 深圳 518172；
2. 南洋理工大学电子与电气工程学院，新加坡 639798）

摘 要：针对敏感信息保密必要性导致的数据孤岛问题，提出了一种适用于工业信息物理系统（CPS）的安全协作入侵检测系统（PFD-GAN）。具体来说，首先通过融入 Wasserstein 距离和标签条件，改进外部分类器生成对抗网络（EC-GAN），构建了一种新型半监督入侵检测模型，以产生能够实用的生成数据来增强分类性能。同时，在改进 EC-GAN 的训练中，融合本地差分隐私技术，防止敏感信息的泄露、保障协作过程的隐私安全。此外，设计了基于去中心化联邦蒸馏的协作方式，允许多个工业 CPS 共同构建一个综合的入侵检测系统，以识别整个网络系统下的威胁，而无须共享统一的模板模型。对真实工业 CPS 数据集的实验评估和理论分析表明，PFD-GAN 可以在免受隐私泄露风险的同时，高效地检测针对工业 CPS 的各种类型攻击。

关键词：入侵检测系统；信息物理系统；生成对抗网络；本地差分隐私；去中心化联邦蒸馏
中图分类号：TP393.08
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023216

Secure federated distillation GAN for CIDS in industrial CPS

LIANG Junwei[1], YANG Geng[1], MA Maode[2], Muhammad Sadiq[1]
1. College of Software Engineering, Shenzhen Institute of Information Technology, Shenzhen 518172, China
2. School of Electronic and Electrical Engineering, Nanyang Technological University, Singapore 639798, Singapore

Abstract: Aiming at the data island problem caused by the imperativeness of confidentiality of sensitive information, a secure and collaborative intrusion detection system (CIDS) for industrial cyber physical systems (CPS) was proposed, called PFD-GAN. Specifically, a novel semi-supervised intrusion detection model was firstly developed by improving external classifier-generative adversarial network (EC-GAN) with Wasserstein distance and label condition, to strengthen the classification performance through the use of synthetic data. Furthermore, local differential privacy (LDP) technology was incorporated into the training process of developed EC-GAN to prevent sensitive information leakage and ensure privacy and security in collaboration. Moreover, a decentralized federated distillation (DFD)-based collaboration was designed, allowing multiple industrial CPS to collectively build a comprehensive intrusion detection system (IDS) to recognize the threats under the entire cyber systems without sharing a uniform template model. Experimental evaluation and theory analysis demonstrate that the proposed PFD-GAN is secure from the threats of privacy leaking and highly effective in detecting various types of attacks on industrial CPS.

Keywords: intrusion detection system, cyber physical system, generative adversarial network, local differential privacy, decentralized federated distillation

收稿日期：2023−08−07；**修回日期：**2023−12−11
通信作者：杨耿，yangg@sziit.edu.cn
基金项目：广东省青年创新人才基金资助项目（No.2022KQNCX233）；公共大数据国家重点实验室基金资助项目（No.PBD2022-14）；深圳市自然科学基金资助项目（No.20220820003203001）
Foundation Items: The Guangdong Provincial Research Platform and Project (No.2022KQNCX233), The Foundation of State Key Laboratory of Public Big Data (No.PBD2022-14), The Shenzhen Natural Science Foundation (No.20220820003203001)

0　引言

工业信息物理系统（CPS, cyber physical system）是一个涉及计算、控制、通信和物理过程的综合体，广泛应用于智能电网、自主运输和无人工厂等领域[1]。由于在交通网络、能源系统、水/气分配等网络中扮演着重要的角色以及高度互联性，工业 CPS 已经成为攻击者的主要目标之一。截至 2022 年年底，针对工业 CPS 的攻击事件年增长率达 25%。新一代网络和计算技术融合扩大了网络威胁的范围，暴露了多个潜在漏洞和威胁[2]。工业 CPS 的通用体系结构如图 1 所示。

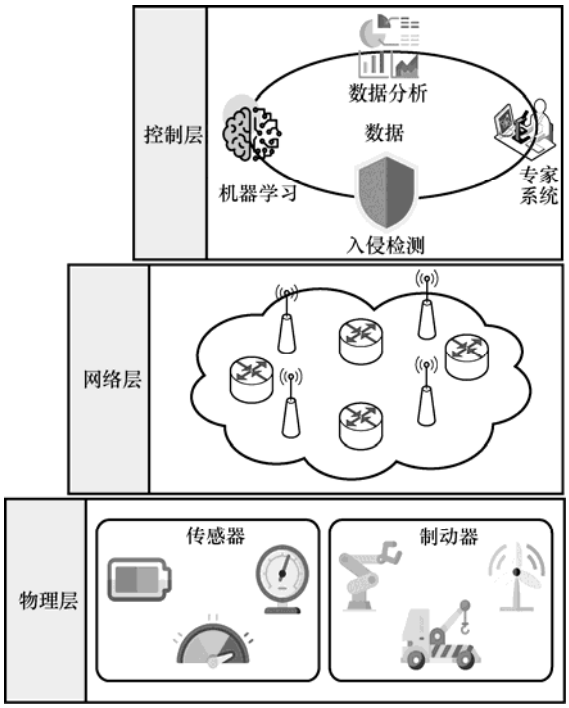


图 1　工业 CPS 的通用体系结构

入侵检测系统（IDS, intrusion detection system）是保障信息安全的关键技术，通过分析网络流量检测来自内部与外部的威胁。然而，由于工业 CPS 的无线通信、分布式、时延敏感等特性，传统 IDS 在此环境下效果有限。目前，针对工业 CPS 的分布式网络簇结构，专家学者对入侵检测技术展开了研究，已取得了一定的理论与实践成果[3-17]，包括 IDS 的基于规则与基于异常入侵检测分类算法/分类器的性能提升[3-10]和工业 CPS 下 IDS 的协作策略[11-17]，相关研究内容将在第 1 节进行讨论。

尽管多种集中/分布式 IDS 解决方案已部署在工业 CPS 中，但仍有一些瓶颈问题尚未解决。1) 老旧的数据集，如 KDD'99、NSL-KDD 等，仍被广泛使用[3-10]，无法充分反映当前网络攻击现状，导致新型攻击识别效果不佳[11]；2) 分布式 IDS 模型间的协作学习有望解决数据集受限和数据孤岛问题[12-18]，但众包过程可能暴露隐私数据；3) 差分隐私联邦协作技术[13-15]能够对协作学习中的隐私数据进行有效的保护，但聚合中心和统一模版缺陷限制了其在工业 CPS 环境下的实际使用；4)工业 CPS 数据集大多分类不均衡，真实数据集中某些攻击类型样本量不足 2%[19]。这些瓶颈问题显然限制了 IDS 在工业 CPS 中检测性能和广泛应用，但目前未有 IDS 方法解决这些挑战。

针对上述问题，提出了一种基于安全联邦蒸馏生成对抗网络（GAN, generative adversarial network）的工业 CPS 协作入侵检测方法，称为 PFD-GAN（privacy-preserving federated distillation GAN）。首先，开发了半监督学习的 IDS 模型，这是在入侵检测中采用外部分类器 GAN（EC-GAN, external classifier-GAN）的检测模型，并进一步采用 Wasserstein 距离替代 Jensen-Shannon 散度作为目标函数，为稳定梯度下降过程提供平滑的度量，同时引入标签条件作为潜在空间的扩展，以改进提高 EC-GAN 的检测性能；此外，在改进的 EC-GAN（Dev EC-GAN）的梯度中添加精心设计的高斯噪声，实现局部差分隐私（LDP, local differential privacy）保护；再者，为工业 CPS 节点设计去中心化联邦蒸馏（DFD, decentralized federated distillation）的协作训练策略，其中生成器通过调度其生成的数据与分布式判别器的反馈损失进行联合训练，而分类器和判别器通过知识蒸馏执行网络权值更新，不需要共享统一的神经网络模板模型。

1　相关工作

1.1　IDS 检测性能优化研究

分类算法/分类器作为入侵检测系统的核心模块是近年的热门研究方向，主要包含两类：基于规则匹配和基于模式识别。1) 基于规则匹配的分类器。吕思才等[3]采用了信誉评分和基于规则的 IDS 进行入侵检测。Nespoli 等[4]和 Rajasoundaran 等[5]分别设计了 2 个用于入侵检测的监视装置。前者监控所有的数据包以确定是否

有攻击行为的出现，而后者同时监控 MAC 层终端的请求发送/清除发送（R2S/C2S, request to send/clear to send）请求的数量来进行异常检测。尽管基于规则的 IDS 具有较高的检测精度和效率，但它只能检测已知的攻击，容易受到未知威胁的影响。2) 基于模式识别的分类器。Naqash 等[6]、Ieracitano 等[7]和 Aldribi 等[8]提出了基于统计方法的入侵检测算法，他们声称所提出的检测算法可以准确有效地检测攻击。实际上，统计类方法必须事先知道异常数据的概率分布，这在实际应用中非常困难。Yao 等[9]提出了一种基于分层时空特征的 IDS，通过卷积神经网络先学习网络流量的低级空间特征，然后使用长短期记忆网络（LSTM, long short-term memory）学习高级时间特征。Chadza[10]等将马尔可夫模型用于预测终端的未来行为以检测恶意活动。但是，在预测之前，必须收集终端的许多历史行为，这意味着在收集历史行为期间，分类器处于无法工作的状态。

通过分析近几年的研究，发现关键问题不在于入侵检测技术，而在于使用过时数据集，导致 IDS 无法适应工业物联网络。真实工业 CPS 数据集的分类不均衡问题也导致 IDS 对稀疏类数据训练不足，成为攻击者的潜在漏洞[11]。

1.2 IDS 分布式协作策略研究

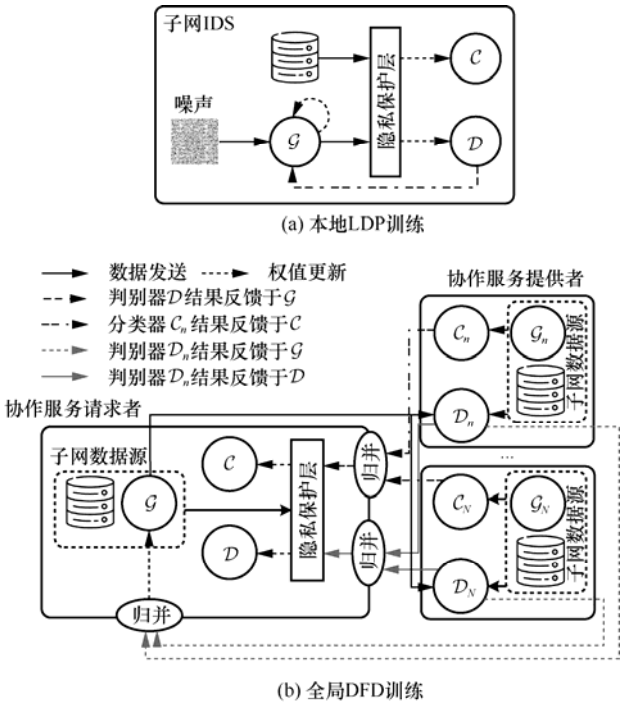
随着网络结构变复杂，工业 CPS 中网络资源分散，暴露了传统集中式 IDS 的局限。现有研究趋向于建立分布式 IDS，通过多个实体监控网络不同部分并相互协作完成检测任务[11]。Shu 等[12]提出了一种基于分布式软件定义网络（SDN, software defined networking）的联邦入侵检测系统，通过在每个路侧单元（RSU, road side unit）上放置一个分布式的 SDN 控制器，让多个 SDN 控制器能够联合训练整个网络的全局入侵检测算法模型。基于 SDN 的 IDS 协作式框架的有效性必须建立在所有的 RSU 都是可信的假设上，一旦某个或多个 RSU 受到 DoS 攻击，IDS 的合作训练过程会立即瘫痪。Ruzafa-Alcázar 等[13]对不同隐私要求和聚合函数的 2 种差分隐私联邦入侵检测模型（即 FedAvg 和 Fed+[14]）进行了系统的比较，评估结果显示基于高斯噪声的差分隐私入侵检测模型有较高的准确性水平，在低隐私要求的情况下，差分隐私对检测性能的影响几乎不可感知。Zhang[15]等设计了一种差分隐私保护的联邦 SecFedNIDS，通过使用基于梯度的模型参数选择方法来筛选有效的低纬聚合参数，并采用在线无监督的异常检测方法，确保模型聚合的可信性。差分隐私联邦协作检测技术是一种解决数据孤岛问题的有效手段，然而现行的方法[13-15]大部分依赖一个聚合中心来完成模型的联邦学习，这无疑带来了单点故障、通信和计算负担等问题，此外，采用统一的模板模型也带来了模型单一化、可解释性与适用性低和管理协调等问题。Abdel-Basset 等[16]提出了一种新型的边缘智能区块链（EIB, edge intelligence blockchain）框架，该框架结合了区块链和移动边缘计算（MEC, mobile edge computing）范式来处理网络节点中的数据交换，从而提供了一种有效、安全和分散的 IDS 协作框架。为了防止 IDS 协作框架中常见的注入攻击和成员推断攻击，Shende 等[17]构建基于区块链的对等联合训练框架用于入侵检测。然而，严格且复杂的共识机制直接导致公有链处理数据的速度超出可容忍范围，通常一个链上信息的更新需要 3~4 个区块的上链才能被最终确认。

分布式 IDS 协作策略确实有利于提高工业 CPS 的安全性和有效性，但协作过程可能向参与者暴露敏感数据，导致子网间形成数据孤岛[18]。

2 PFD-GAN 系统架构与威胁模型

2.1 系统架构

PFD-GAN 的系统架构大致分为 2 个部分：1) 工业 CPS 内部的 IDS 本地 LDP 训练，如图 2(a)所示；2) 协作服务请求者和提供者之间的 IDS 全局 DFD 训练，如图 2(b)所示。在进行全局训练之前，每个独立 CPS 分别利用本地数据集进行本地训练。本地训练完成后，任何面向协作的 CPS，即协作服务请求者，都可以通过与多个对等方（即协作服务提供者）联网来启动 IDS 的全局训练过程，以构建综合的 IDS 模型。在协作架构中有一个协作服务请求者 u 和 N 个对应的协作服务提供者 $\{v_n\}_n^N$。u 和第 n 个协作服务提供者 v_n 分别配备了本地数据集 $\boldsymbol{S}$ 和 $\boldsymbol{S}_n$，完整数据集由所有本地数据集组成，即 $\bigcup_{n=1}^{N}\boldsymbol{S}_n \cup \boldsymbol{S}$。生成器 $\mathcal{G}$、判别器 $\mathcal{D}$ 和外部分类器 $\mathcal{C}$ 存储在具有隐私保护层的 u 中，而 v_n 以点对点（P2P, peer to peer）方式运行自己的判别器 $\mathcal{D}_n$ 和分类器 $\mathcal{C}_n$。$\mathcal{G}$、$\mathcal{D}$ 和 $\mathcal{C}$ 的权重分别为 $\boldsymbol{\varphi}$、$\boldsymbol{\omega}$ 和 $\boldsymbol{\theta}$。相应地，v_n 的权值分别为 $\boldsymbol{\varphi}_n$、$\boldsymbol{\omega}_n$ 和 $\boldsymbol{\theta}_n$。



(a) 本地LDP训练

(b) 全局DFD训练

图 2　PFD-GAN 系统架构

2.2　威胁模型

1) 工业 CPS 的网络威胁。工业 CPS 所面临的风险不仅包括传统计算机系统常见的网络威胁，如拒绝服务（DoS/DDoS）攻击，还需要应对一系列专门针对工业 CPS 设计的新型网络威胁，如命令注入和响应注入攻击。重点关注以下内容。

① 侦察攻击，旨在收集工业 CPS 的有价值信息，绘制网络架构图，以及识别设备特征，如制造商、型号号码、支持的网络协议和设备地址。

② 响应注入攻击，包括 NMRI （naive malicious response injection）攻击和 CMRI （complex malicious response injection）攻击，目的是干扰监视和报告状态，通过伪造返回响应提供错误的系统状态信息。

③ 命令注入攻击，包括 MSCI（malicious state command injection）攻击、MPCI （malicious parameter command injection）攻击和 MFCI（malicious function command injection）攻击，通过注入虚假命令误导工业控制系统，可能导致设备配置、过程设置或通信目的地的未经授权修改。

④ DoS 和 DDoS 攻击通常通过以极高的频率向目标发送大量无用请求，耗尽工业 CPS 中服务器系统的资源，这可能会导致服务器瘫痪或阻止合法请求得到响应。

2) 联邦协作的网络威胁。假设所有的工业 CPS 所有者都是有条件诚实的，且严格遵循设计的协议，但可能关注其他工业 CPS 的数据资源。此外，还需要考虑恶意窃听者或其他外部攻击者可能截取通信链路，以尝试访问每个工业 CPS 的数据资源和入侵检测模型的参数。在这种情况下，考虑以下 2 种网络威胁。

① 数据资源窃听。工业 CPS 的数据资源，尤其是攻击样本，具有极高的敏感性，甚至可能影响国家的重大利益。如果这些资源被恶意第三方所窃听，可能会导致严重的商业损失或国家安全风险。

② 模型参数窃听。入侵检测模型的参数包含数据资源的关键信息。未经授权获取这些参数，可能会泄露有关的数据资源，如网络威胁类型或其示例分布。

3　PFD-GAN 本地 LDP 训练

PFD-GAN 中的 Dev EC-GAN 模型是为分布式工业 CPS 的协作入侵检测特别设计的，它利用带有标签条件的 WGAN 和半监督外部分类器来提高异常检测的性能，其通用神经网络结构如图 3 所示。在实际应用中，不同的 CPS 可以根据自身需求定制各自的 Dev EC-GAN 模型。Dev EC-GAN 模型主要由生成器、判别器和分类器 3 个模块组成，每个模块都有独立的网络结构。

1) 生成器 $\mathcal{G}(\cdot)$。生成器由全连接（FC, fully-connected）层、Dt（dropout）层和 LReLU（leakyReLU）激活函数组成，其中，FC 层从小到大依次排序，且最末尾的 FC 层包含与输入数据维度相同数量的神经元。在运作时，潜在空间中随机采样的噪声向量 $\boldsymbol{z} \sim p(\boldsymbol{z})$ 和类标签信息 l（即标签条件概率）一并被输入生成器 $\mathcal{G}(\cdot)$ 中，以生成稀疏类的补充数据样本 $\hat{\boldsymbol{x}} = \mathcal{G}(\boldsymbol{z} \mid l)$ 用于半监督学习。

2) Wasserstein 判别器 $\mathcal{D} \triangleq f_\omega(\cdot)$。判别器使用与生成器类似的神经网络架构，但将 FC 组合层从最大尺度排列到最小尺度，且没有加入 Dt 层，并以 1×1维 FC 层结束，以将生成样本与真实数据区分开来。判别执行后，$f_\omega(\boldsymbol{x} \mid y) = 1$ 和 $f_\omega(\hat{\boldsymbol{x}} \mid l) = 1$ 表示条件输入（包括真实条件输入 $(\boldsymbol{x}, y) \sim p(\boldsymbol{S})$ 和生成条件输入 $(\hat{\boldsymbol{x}}, l)$）均被判别器 $f_\omega(\cdot)$ 归类为真实数据；$f_\omega(\boldsymbol{x} \mid y) = 0$ 或 $f_\omega(\hat{\boldsymbol{x}} \mid l) = 0$ 时，两方条件输入被认为是生成的数据。

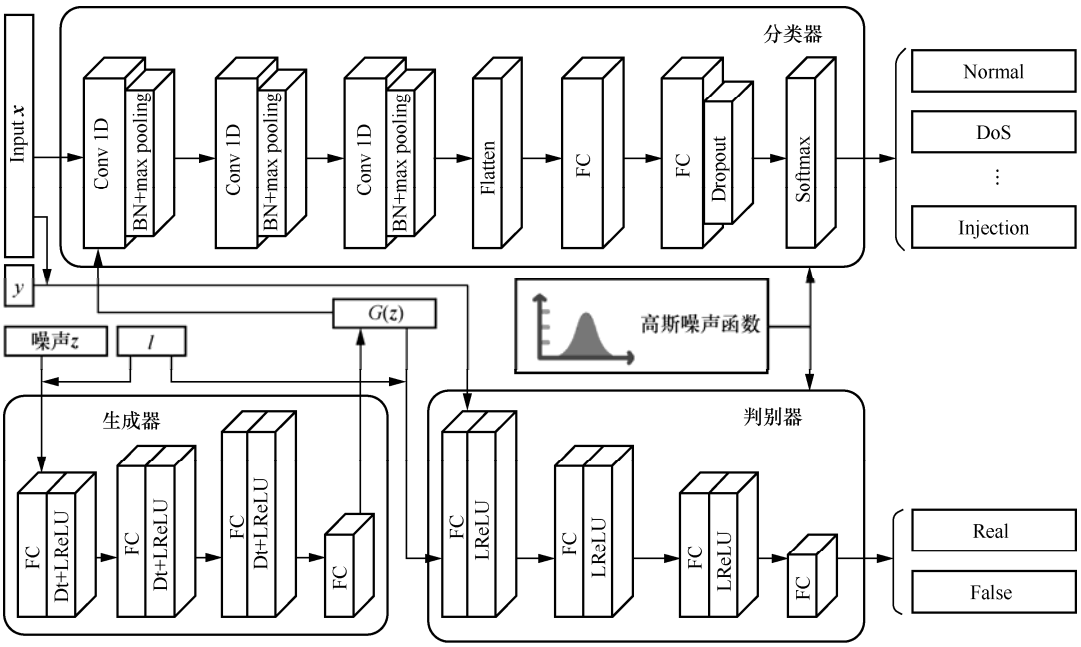


图 3　Dev EC-GAN 模型的通用神经网络结构

3) 分类器$\mathcal{C}(\cdot)$。分类器主要由3个卷积块组成，其次是一个展平层（Flattern）、2 个 FC 层、一个 Dropout 层和一个 Softmax 层。每个卷积块由一个一维卷积（Conv 1D）层、一个批归一化（BN, batch normalization）层和一个最大池化（max pooling）层组成。在分类过程中，根据$y=\mathcal{C}(\boldsymbol{x})$或$y=\mathcal{C}(\mathcal{G}(\boldsymbol{z}\,|\,l))$的结果，$\mathcal{C}(\cdot)$利用 Softmax 层将归一化输出映射到概率最大的预测类上，例如，正常流量（Normal）、拒绝攻击（DoS）、注入攻击（Injection）等。

算法 1 展示了协作服务请求者的本地训练过程（图 2(a)）。协作服务提供者的本地训练过程可以由此类比。首先，在算法第 1)行初始化$\mathcal{G}$、$\mathcal{D}$和$\mathcal{C}$，权值分别为$\boldsymbol{\varphi}$、$\boldsymbol{\omega}$和$\boldsymbol{\theta}$。当$t_1 \leqslant T_\varphi$时，其中T_φ是$\mathcal{G}$的迭代数，在算法第 3)～15)行进行$\mathcal{G}$和$\mathcal{D}$的本地训练，以产生可用于半监督学习的生成样本。在算法第 5～10)行的循环中，使用真实样本$\{\boldsymbol{x}^{(b)},y^{(b)}\}_b^B$和$\{\boldsymbol{z}^{(b)}\}_b^B$产生的生成样本迭代训练$\mathcal{D}$，直到$t_2>T_\omega$，其中$T_\omega$为$\mathcal{D}$的迭代数。具体来说，在算法第 7)~8)行计算$\mathcal{D}$相对于真实样本$(\boldsymbol{x}^{(b)},y^{(b)})$和随机噪声$\boldsymbol{z}^{(b)}$的梯度时，通过注入精心修剪高斯噪声，确保灵敏度控制在ε范围内。$\mathcal{D}$对于第b个样本的梯度可以通过式(1)计算。

$$g_\omega(\boldsymbol{x}^{(b)},y^{(b)},\boldsymbol{z}^{(b)})=\nabla_\omega[f_\omega(\boldsymbol{x}^{(b)}\,|\,y^{(b)})-f_\omega(\mathcal{G}(\boldsymbol{z}^{(b)}\,|\,y^{(b)})\,|\,y^{(b)})] \tag{1}$$

其中，$f_\omega(\cdot)$和$\mathcal{G}(\cdot)$以类标签$y^{(b)}$为条件。在算法第 9)行中，RMSProp(·)是一个优化函数，可以根据梯度$\tilde{g}_\omega$的大小自适应地调整$\boldsymbol{\omega}_{t_2}$（$t_2$迭代中的$\boldsymbol{\omega}$值）。进而在算法 1 第 10)行采用 clip(·)函数来保证$\{f_\omega\}_\omega$遵循K_ω-Lipschitz，并且裁剪每个数据的梯度以确保落在$[-C_\omega,C_\omega]$内。在$\mathcal{D}$训练完成后，计算$\mathcal{G}$相对于$\{\boldsymbol{z}^{(b)},l^{(b)}\}_b^B$的平均梯度，如式(2)所示。

$$\bar{g}_\varphi=-\nabla_\varphi\frac{1}{B}\sum_b^B f_\omega\left(\mathcal{G}(\boldsymbol{z}^{(b)}\,|\,l^{(b)})\,|\,l^{(b)}\right) \tag{2}$$

其中，$f_\omega(\cdot)$和$\mathcal{G}(\cdot)$以随机生成的标签$l^{(b)}$为条件。接下来，在算法 1 第 15)行通过利用 RMSProp(·)、学习率α_φ和平均梯度$\bar{g}_\varphi$将$\boldsymbol{\varphi}_{t_1}$更新为$\boldsymbol{\varphi}_{t_1+1}$。当$T_\varphi<t_1\leqslant T_\theta$时，外部分类器$\mathcal{C}$的本地训练在算法第 16)～24)行中进行。在算法第 20)行中，针对一组真实样本$(\boldsymbol{x}^{(b)},y^{(b)})$和生成数据$(\mathcal{G}(\boldsymbol{z}^{(b)}\,|\,l^{(b)}),l^{(b)})$，通过最小化经验损失函数$\mathcal{L}(\theta)$，能够计算每个$\mathcal{C}$的梯度$g_\theta(\boldsymbol{x}^{(b)},y^{(b)},\boldsymbol{z}^{(b)},l^{(b)})$，如式(3)所示。

$$g_\theta(\boldsymbol{x}^{(b)},y^{(b)},\boldsymbol{z}^{(b)},l^{(b)})=\nabla_\theta[\mathcal{L}_{ce}(\mathcal{C}(\boldsymbol{x}^{(b)}),y^{(b)})+\lambda\mathcal{L}_{\text{ce}}\begin{pmatrix}\mathcal{C}(\mathcal{G}(\boldsymbol{z}^{(b)}|l^{(b)})),\\ \arg\max\left(\mathcal{C}(\mathcal{G}(\boldsymbol{z}^{(b)}|l^{(b)}))\right)\end{pmatrix}>\tau] \tag{3}$$

其中，$\mathcal{L}_{\text{ce}}(\cdot)$是交叉熵损失，$\tau$是伪标签阈值。类似地，为了保护隐私数据，算法 1 第 21)～23)行裁剪每个g_θ的范数以及添加的噪声，并在这个平均噪声

梯度 $\tilde{g}_\theta$ 的负方向进行更新，将 $\boldsymbol{\theta}_{t_1}$ 更新为 $\boldsymbol{\theta}_{t_1+1}$。上述过程迭代运行，直至 $\boldsymbol{\varphi}$、$\boldsymbol{\omega}$ 和 $\boldsymbol{\theta}$ 收敛或达到最大迭代次数。

算法 1　本地 LDP 训练

输入　$\alpha_\varphi, \alpha_\omega, \alpha_\theta$　/*$\mathcal{G}$、$\mathcal{D}$ 和 $\mathcal{C}$ 的学习率*/

B　/*数据批量抽样数目*/

$T_\varphi, T_\omega, T_\theta$　/*$\mathcal{G}$、$\mathcal{D}$ 和 $\mathcal{C}$ 的迭代数*/

$\sigma_\omega, \sigma_\theta, C_\omega, C_\theta$　/*噪声和裁剪范围*/

输出　差分隐私保护的已训练 $\mathcal{G}$、$\mathcal{D}$ 和 $\mathcal{C}$

1) 初始化 $\mathcal{G}$、$\mathcal{D}$ 和 $\mathcal{C}$ 的权值 $\boldsymbol{\varphi}$、$\boldsymbol{\omega}$ 和 $\boldsymbol{\theta}$

2) for $t_1 = 1,2,\cdots,T_\varphi,\cdots,T_\theta$ do

3)　if $t_1 \leqslant T_\varphi$ then

4)　for $t_2 = 1,2,\cdots,T_\omega$ do

5)　抽样 $\{\boldsymbol{z}^{(b)}\}_b^B \sim p(\boldsymbol{z})$

6)　抽样 $\{\boldsymbol{x}^{(b)}, y^{(b)}\}_b^B \sim p(\boldsymbol{S})$

7)　用式(1)求每个 b 的 $g_\omega(\boldsymbol{x}^{(b)}, y^{(b)}, \boldsymbol{z}^{(b)})$

8)　$\tilde{g}_\omega \leftarrow \dfrac{\sum_b g_\omega(\boldsymbol{x}^{(b)}, y^{(b)}, \boldsymbol{z}^{(b)}) + \boldsymbol{N}(0, \sigma_\omega^2 C_\omega^2 I)}{B}$

9)　$\boldsymbol{\omega}_{t_2+1} \leftarrow \boldsymbol{\omega}_{t_2} + \alpha_\omega \mathrm{RMSProp}(\boldsymbol{\omega}_{t_2}, \tilde{g}_\omega)$

10)　$\boldsymbol{\omega}_{t_2+1} \leftarrow \mathrm{clip}(\boldsymbol{\omega}_{t_2+1}, -C_\omega, C_\omega)$

11)　end for

12)　抽样 $\{\boldsymbol{z}^{(b)}\}_b^B \sim p(\boldsymbol{z})$

13)　随机产生 $\{l^{(b)}\}_b^B$

14)　根据式(2)计算 $\bar{g}_\varphi$

15)　$\boldsymbol{\varphi}_{t_1+1} \leftarrow \boldsymbol{\varphi}_{t_1} - \alpha_\varphi \mathrm{RMSProp}(\boldsymbol{\varphi}_{t_1}, \bar{g}_\varphi)$

16)　else

17)　抽样 $\{\boldsymbol{z}^{(b)}\}_b^B \sim p(\boldsymbol{z})$

18)　抽样 $\{\boldsymbol{x}^{(b)}, y^{(b)}\}_b^B \sim p(\boldsymbol{S})$

19)　随机产生 $\{l^{(b)}\}_b^B$

20)　用式(3)求每个 b 的 $g_\theta(\boldsymbol{x}^{(b)}, y^{(b)}, \boldsymbol{z}^{(b)}, l^{(b)})$

21)　$\tilde{g}_\theta \leftarrow \dfrac{\sum_b g_\theta(\boldsymbol{x}^{(b)}, y^{(b)}, \boldsymbol{z}^{(b)}, l^{(b)}) + \mathcal{N}(0, \sigma_\theta^2 C_\theta^2 I)}{B}$

22)　$\boldsymbol{\theta}_{t_1+1} \leftarrow \boldsymbol{\theta}_{t_1} - \alpha_\theta \tilde{g}_\theta$

23)　$\boldsymbol{\theta}_{t_1+1} \leftarrow \mathrm{clip}(\boldsymbol{\theta}_{t_1+1}, -C_\theta, C_\theta)$

24)　end if

25)end for

PFD-GAN 的本地 LDP 训练是通过将高斯噪声添加到判别器 $\mathcal{D}$ 和分类器 $\mathcal{C}$ 与训练数据相关的每个梯度中，伴随噪声梯度界范数的裁剪、平均梯度值的计算，达到控制训练数据在随机梯度下降计算中影响的目的，从而实现判别器 $\mathcal{D}$ 和分类器 $\mathcal{C}$ 的 LDP 保护，并自然保证未用于训练数据的隐私，因为替换这些数据不会导致输出分布发生任何变化（即 $\varepsilon = 0$ 的情况）。同时，生成器 $\mathcal{G}$ 的 LDP 保护可以在与 $\mathcal{D}$ 的博弈训练中实现，这是因为 LDP 技术有一个后处理特性，其证明了 LDP 输出的任何映射都不会泄露敏感数据。所以，生成器 $\mathcal{G}$ 在本地训练之后的生成数据是局部差分隐私安全的，因为这里的映射实际上是 $\mathcal{G}$ 的计算权值，而输出是 $\mathcal{D}$ 的 LDP 保护的权值。

4　PFD-GAN 全局 DFD 训练

算法 2 展示了协作服务请求者 u 的生成器 $\mathcal{G}$ 的全局 DFD 训练伪代码（图 2(b)）。对于每个全局迭代 t，$\mathcal{G}$ 首先生成一批大小为 B 的生成样本，即 $\hat{\boldsymbol{X}}_{n,t} = \{\mathcal{G}(\boldsymbol{z}^{(b)} \mid l^{(b)})\}_b^B$ 和生成的随机标签 $\boldsymbol{L}_{n,t} = \{l^{(b)}\}_b^B$。然后，工业 CPS 中的每个协作服务提供者 v_n 都会被发送 $\mathcal{G}$ 生成的 $(\hat{\boldsymbol{X}}_{n,t}, \boldsymbol{L}_{n,t})$，用以计算 $\mathcal{G}$ 的梯度。一旦 v_n 从 u 处接收到对应的 $(\hat{\boldsymbol{X}}_{n,t}, \boldsymbol{L}_{n,t})$，便可以计算出误差项 $\{e_{n,t}^{(b)}\}_b^B$，其中第 b 个误差可以通过式(4)计算。

$$e_{n,t}^{(b)} = -\frac{\partial \log\left(f_{\omega_n}(\hat{\boldsymbol{x}}^{(b)} \mid l^{(b)})\right)}{\partial \hat{\boldsymbol{x}}^{(b)}} \tag{4}$$

其中，$\hat{\boldsymbol{x}}^{(b)}$ 是批次 $\hat{\boldsymbol{X}}_{n,t}$ 的第 b 个数据。当 u 从所有协作服务提供者中得到误差项集合 $\left\{\{e_{n,t}^{(b)}\}_b^B\right\}_n^N$ 时，便可更新 $\mathcal{G}$ 的权重，即 $\Delta\varphi_t^{(i)} \leftarrow \dfrac{\left(\sum_{n=1}^{N} \sum_{\hat{\boldsymbol{x}}^{(b)} \in \hat{\boldsymbol{X}}_{n,t}} e_{n,t}^{(b)} \left(\dfrac{\partial \hat{\boldsymbol{x}}^{(b)}}{\partial \varphi_t^{(i)}}\right)\right)}{NB}$，其中 $\varphi_t^{(i)}$ 是 $\boldsymbol{\varphi}_t$ 的第 i 个元素。在计算出 $\Delta\varphi_t^{(i)}$ 之后，利用 Adam 优化器的聚合并行更新方法，将 $\varphi_t^{(i)}$ 更新为 $\varphi_{t+1}^{(i)} \leftarrow \varphi_t^{(i)} + \mathrm{Adam}(\Delta\varphi_t^{(i)})$。上述过程循环迭代，直到 $\boldsymbol{\varphi}_{t+1}$ 收敛或达到最大迭代次数。对于全局 DFD 训练，协作服务请求者的 $\mathcal{G}$ 使用协作服务提供者的判别器 $\{\mathcal{D}_n\}_n$ 和其本地数据集 $\{\boldsymbol{S}_n\}_n$ 进行更新。这是一个 1-N 的博弈，其中 $\mathcal{G}$ 被更新优化以产生可用的生成数据，而 $\mathcal{D}_n$ 试图从 $\mathcal{G}$ 生成的数据与真实数据 $\{\boldsymbol{S}_n\}_n$ 的判别中进行博弈更新。

算法 2　生成器 $\mathcal{G}$ 的全局 DFD 训练

1) repeat:

2)　procedure 协作服务请求者 u

3) for $n=1,\cdots,N$ do
4) 抽样 $\{\boldsymbol{z}^{(b)}\}_b^B \sim p(\boldsymbol{z})$
5) 随机产生 $\boldsymbol{L}_{n,t}=\{l^{(b)}\}_b^B$
6) 向 v_n 发送($\hat{\boldsymbol{X}}_{n,t}=\{\mathcal{G}(\boldsymbol{z}^{(b)}\mid l^{(b)})\}_b^B, \boldsymbol{L}_{n,t}$)
7) end for
8) 从 v_n 中接收 $\left\{\{e_{n,t}^{(b)}\}_b^B\right\}_n^N$
9) for each $\varphi_t^{(i)}\in\boldsymbol{\varphi}_t$ do
10) $\Delta\varphi_t^{(i)}\leftarrow\dfrac{\left(\sum\limits_{n=1}^{N}\sum\limits_{\hat{\boldsymbol{x}}^{(b)}\in\hat{\boldsymbol{X}}_{n,t}} e_{n,t}^{(b)}\left(\dfrac{\partial\hat{\boldsymbol{x}}^{(b)}}{\partial\varphi_t^{(i)}}\right)\right)}{NB}$
11) $\varphi_{t+1}^{(i)}\leftarrow\varphi_t^{(i)}+\mathrm{Adam}(\Delta\varphi_t^{(i)})$
12) end for
13) end procedure
14) procedure 第 n 个协作服务提供者 v_n
15) 从 u 中接收 $(\hat{\boldsymbol{X}}_{n,t},\boldsymbol{L}_{n,t})=\{\hat{\boldsymbol{x}}^{(b)},l^{(b)}\}_b^B$
16) for $b=1,\cdots,B$ do
17) 根据式(4)计算 $e_{n,t}^{(b)}$
18) end for
19) 向 u 发送 $\{e_{n,t}^{(b)}\}_b^B$
20) end procedure
21)until $\boldsymbol{\varphi}_{t+1}$ 收敛 or $++t>T$

算法 3 展示了判别器 $\mathcal{D}$ 和分类器 $\mathcal{C}$ 的全局 DFD 训练过程。

算法 3 判别器 $\mathcal{D}$ 和分类器 $\mathcal{C}$ 的全局 DFD 训练

1) repeat:
2) procedure 第 n 个协作服务提供者 v_n
3) for $\boldsymbol{s}_n^{(b)}=(\boldsymbol{x}_n^{(b)},y_n^{(b)})\in\boldsymbol{S}_n$ do
4) $\boldsymbol{F}_{\theta_n,t}^m=\boldsymbol{F}_{\theta_n,t}^m+\boldsymbol{F}\left(\boldsymbol{\theta}_n,\boldsymbol{s}_n^{(b)}\right)$
5) $\mathrm{cnt}_{\theta_n,t}^m=\mathrm{cnt}_{\theta_n,t}^m+1$
6) $\boldsymbol{F}_{\omega_n,t}^1=\boldsymbol{F}_{\omega_n,t}^1+\boldsymbol{F}\left(\boldsymbol{\omega}_n,\boldsymbol{s}_n^{(b)}\right)$
7) $\mathrm{cnt}_{\omega_n,t}^1=\mathrm{cnt}_{\omega_n,t}^1+1$
8) end for
9) for $\hat{\boldsymbol{s}}_n^{(b)}=(\hat{\boldsymbol{x}}_n^{(b)},l_n^{(b)})\in\hat{\boldsymbol{S}}_n$ do
10) $\boldsymbol{F}_{\omega_n,t}^0=\boldsymbol{F}_{\omega_n,t}^0+\boldsymbol{F}\left(\boldsymbol{\omega}_n,\hat{\boldsymbol{s}}_n^{(b)}\right)$
11) $\mathrm{cnt}_{\omega_n,t}^0=\mathrm{cnt}_{\omega_n,t}^0+1$
12) end for
13) $\bar{\boldsymbol{F}}_{\omega_n,t}^0=\dfrac{\boldsymbol{F}_{\omega_n,t}^0}{\mathrm{cnt}_{\omega_n,t}^0}$, $\bar{\boldsymbol{F}}_{\omega_n,t}^1=\dfrac{\boldsymbol{F}_{\omega_n,t}^1}{\mathrm{cnt}_{\omega_n,t}^1}$
14) 为每个 m 计算 $\bar{\boldsymbol{F}}_{\theta_n,t}^m=\dfrac{\boldsymbol{F}_{\theta_n,t}^m}{\mathrm{cnt}_{\theta_n,t}^m}$
15) 向 u 发送 $\left(\{\bar{\boldsymbol{F}}_{\omega_n,t}^0,\bar{\boldsymbol{F}}_{\omega_n,t}^1\},\{\bar{\boldsymbol{F}}_{\theta_n,t}^m\}_m^{\boldsymbol{M}}\right)$
16) end procedure
17) procedure 协作服务请求者 u
18) 接收所有响应 $\left\{\{\bar{\boldsymbol{F}}_{\omega_n,t}^0,\bar{\boldsymbol{F}}_{\omega_n,t}^1\},\{\bar{\boldsymbol{F}}_{\theta_n,t}^m\}_m^{\boldsymbol{M}}\right\}_n^N$
19) $\bar{\boldsymbol{F}}_t^0\leftarrow\dfrac{\sum\limits_n\bar{\boldsymbol{F}}_{\omega_n,t}^0}{N}$, $\bar{\boldsymbol{F}}_t^1\leftarrow\dfrac{\sum\limits_n\bar{\boldsymbol{F}}_{\omega_n,t}^1}{N}$
20) 为每个 m 计算 $\bar{\boldsymbol{F}}_t^m=\dfrac{\sum\limits_n\bar{\boldsymbol{F}}_{\theta_n,t}^m}{N}$
21) 根据式(11)计算每个 b 的 $g_{\theta_t}(\boldsymbol{s}^{(b)})$
22) 根据式(12)计算每个 b 的 $g_{\omega_t}(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})$
23) 根据式(13)计算每个 g_{θ_t} 的 $\hat{g}_{\theta_t}(\boldsymbol{s}^{(b)})$
24) 根据式(14)求每个 g_{ω_t} 的 $\hat{g}_{\omega_t}(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})$
25) $\tilde{g}_{\theta_t}\leftarrow\dfrac{\sum\limits_b\hat{g}_{\theta_t}(\boldsymbol{s}^{(b)})+\mathcal{N}\left(0,\sigma_\theta^2C_\theta^2I\right)}{B}$
26) $\tilde{g}_{\omega_t}\leftarrow\dfrac{\sum\limits_b\hat{g}_{\omega_t}(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})+\mathcal{N}\left(0,\sigma_\omega^2C_\omega^2I\right)}{B}$
27) $\boldsymbol{\omega}_{t+1}\leftarrow\boldsymbol{\omega}_t-\alpha_\omega\tilde{g}_{\omega_t}$, $\boldsymbol{\theta}_{t+1}\leftarrow\boldsymbol{\theta}_t-\alpha_\theta\tilde{g}_{\theta_t}$
28) end procedure
29)until $\boldsymbol{\omega}_{t+1}$ 和 $\boldsymbol{\theta}_{t+1}$ 收敛 or $++t>T$

为了更好地理解算法 3，需要先对知识蒸馏相关的符号进行简单的说明与解释。$m\in\boldsymbol{M}=\{\text{“Normal”},\text{“DoS”},\text{“Injection”},\cdots\}$ 被定义为数目 $|\boldsymbol{M}|$ 的字母表标签。$\boldsymbol{F}(\boldsymbol{\theta}_n,\boldsymbol{s}_n^{(b)})$（或 $\boldsymbol{F}\left(\boldsymbol{\omega}_n,\boldsymbol{s}_n^{(b)}\right)$）是 Softmax 函数归一化后的局部 logit 向量，其中 $\boldsymbol{\theta}_n$（或 $\boldsymbol{\omega}_n$）和 $\boldsymbol{s}_n^{(b)}=\left\{\boldsymbol{x}_n^{(b)},y_n^{(b)}\right\}\in\boldsymbol{S}_n$ 分别是 $\mathcal{C}_n$（或 $\mathcal{D}_n$）的权重和输入。函数 $\mathcal{L}_{\mathrm{ce}}(\cdot)$ 是用于损失函数和蒸馏正则化的交叉熵损失。γ_θ（或 γ_ω）是 $\boldsymbol{\theta}$（或 $\boldsymbol{\omega}$）的权重参数。$\bar{\boldsymbol{F}}_{\theta_n,t}^m$（或 $\{\bar{\boldsymbol{F}}_{\omega_n,t}^0,\bar{\boldsymbol{F}}_{\omega_n,t}^1\}$）是 $\boldsymbol{\theta}_n$（或 $\boldsymbol{\omega}_n$）在第 t 次迭代时 $y_n^{(b)}=m$（或 $y_n^{(b)}=0$或1）的局部平均 logit 向量。$\bar{\boldsymbol{F}}_t^m$（或 $\{\bar{\boldsymbol{F}}_t^0,\bar{\boldsymbol{F}}_t^1\}$）是全局平均 logit 向量，其等于 $\{\bar{\boldsymbol{F}}_{\theta_n,t}^m\}_n$（或 $\left\{\{\bar{\boldsymbol{F}}_{\omega_n,t}^0\}_n,\{\bar{\boldsymbol{F}}_{\omega_n,t}^1\}_n\right\}$）的平均值。$\mathrm{cnt}_{\theta_n,t}^m$（或 $\left\{\mathrm{cnt}_{\omega_n,t}^0,\mathrm{cnt}_{\omega_n,t}^1\right\}$）是统计 ground-truth 标签等于 m（或 $\{0,1\}$）的样本计数。如算法 3 和图 2(b)所示，在任意的第 t 次迭代中，对协作服务请求者 u 的 $\mathcal{D}$ 和 $\mathcal{C}$ 的全局 DFD 训练主要包含以下步骤。

步骤 1 第 n 个协作服务提供者 v_n 利用其本地数据集 $\boldsymbol{S}_n$ 为每个标签 m（$m\in\boldsymbol{M}$）计算 $\mathcal{C}_n$ 的局部

logit 向量$\left\{\boldsymbol{F}_{\theta_n,t}^m\right\}_m^{\boldsymbol{M}}$，并为每个 m 计算相应的 $\mathrm{cnt}_{\theta_n,t}^m$ 统计值，如式(5)～式(6)所示。

$$\boldsymbol{F}_{\theta_n,t}^m=\boldsymbol{F}_{\theta_n,t}^m+\boldsymbol{F}\left(\boldsymbol{\theta}_n,\boldsymbol{s}_n^{(b)}\right) \tag{5}$$

$$\mathrm{cnt}_{\theta_n,t}^m=\mathrm{cnt}_{\theta_n,t}^m+1 \tag{6}$$

除了真实样本 $\boldsymbol{S}_n$，一组生成样本$\left\{\hat{\boldsymbol{s}}_n^{(b)}\mid\hat{\boldsymbol{s}}_n^{(b)}\in\left\{\hat{\boldsymbol{x}}_n^{(b)},l_n^{(b)}\right\}_b^B\right\}$也作为补充数据样本，用以计算 $\mathcal{D}_n$ 的局部 logit 向量$\left\{\boldsymbol{F}_{\omega_n,t}^0,\boldsymbol{F}_{\omega_n,t}^1\right\}$和真实与生成数据的计数$\left\{\mathrm{cnt}_{\omega_n,t}^0,\mathrm{cnt}_{\omega_n,t}^1\right\}$，如式(7)～式(8)所示。

$$\boldsymbol{F}_{\omega_n,t}^{\{0,1\}}=\boldsymbol{F}_{\omega_n,t}^{\{0,1\}}+\boldsymbol{F}\left(\boldsymbol{\omega}_n,\hat{\boldsymbol{s}}_n^{(b)}\right) \tag{7}$$

$$\mathrm{cnt}_{\omega_n,t}^0=\mathrm{cnt}_{\omega_n,t}^0+1,\mathrm{cnt}_{\omega_n,t}^1=\mathrm{cnt}_{\omega_n,t}^1+1 \tag{8}$$

步骤 2　第 n 个协作服务提供者计算并上传其所有局部平均 logit 向量$\left(\left\{\bar{\boldsymbol{F}}_{\omega_n,t}^0,\bar{\boldsymbol{F}}_{\omega_n,t}^1\right\},\left\{\bar{\boldsymbol{F}}_{\theta_n,t}^m\right\}_m^{\boldsymbol{M}}\right)$，其计算式如式(9)～式(10)所示。

$$\bar{\boldsymbol{F}}_{\theta_n,t}^m=\frac{\boldsymbol{F}_{\theta_n,t}^m}{\mathrm{cnt}_{\theta_n,t}^m} \tag{9}$$

$$\bar{\boldsymbol{F}}_{\omega_n,t}^0=\frac{\boldsymbol{F}_{\omega_n,t}^0}{\mathrm{cnt}_{\omega_n,t}^0},\bar{\boldsymbol{F}}_{\omega_n,t}^1=\frac{\boldsymbol{F}_{\omega_n,t}^1}{\mathrm{cnt}_{\omega_n,t}^1} \tag{10}$$

然后，协作服务请求者根据每个标签分别从所有协作服务提供者上传的局部平均 logit 向量计算平均值，并进而计算所有标签的全局平均 logit 向量，即 $\bar{\boldsymbol{F}}_t^{\{0,1\}}=\dfrac{\sum_n\bar{\boldsymbol{F}}_{\omega_n,t}^{\{0,1\}}}{N}$ 和$\left\{\bar{\boldsymbol{F}}_t^m=\sum_n\dfrac{\bar{\boldsymbol{F}}_{\theta_n,t}^m}{N}\mid m\in\boldsymbol{M}\right\}$。

步骤 3　与第 3 节类似，LDP 技术也被用于协作服务请求者的 $\mathcal{D}$ 和 $\mathcal{C}$ 的更新过程。$\mathcal{C}$ 关于每个 $\boldsymbol{s}^{(b)}\in\boldsymbol{S}$ 样本的梯度 g_{θ_t} 和 $\mathcal{D}$ 针对每个 $(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})$ 样本的梯度可利用$\left(\bar{\boldsymbol{F}}_t^0,\bar{\boldsymbol{F}}_t^1,\left\{\bar{\boldsymbol{F}}_t^m\right\}_m^{\boldsymbol{M}}\right)$计算得出，如式(11)～式(12)所示。

$$g_{\theta_t}(\boldsymbol{s}^{(b)})=\nabla_\theta\left[\mathcal{L}_{\mathrm{ce}}\left(\boldsymbol{F}(\boldsymbol{\theta}_t,\boldsymbol{s}^{(b)}),m\right)+\gamma_\theta\mathcal{L}_{\mathrm{ce}}\left(\boldsymbol{F}(\boldsymbol{\theta}_t,\boldsymbol{s}^{(b)}),\bar{\boldsymbol{F}}_t^m\right)\right] \tag{11}$$

$$\begin{aligned}g_{\omega_t}(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})=\nabla_\omega\Big[&\mathcal{L}_{\mathrm{ce}}\left(\boldsymbol{F}(\boldsymbol{\omega}_t,\boldsymbol{s}^{(b)}),1\right)+\\&\mathcal{L}_{\mathrm{ce}}\left(\boldsymbol{F}(\boldsymbol{\omega}_t,\hat{\boldsymbol{s}}^{(b)}),0\right)+\gamma_\omega\left(\mathcal{L}_{\mathrm{ce}}\left(\boldsymbol{F}(\boldsymbol{\omega}_t,\boldsymbol{s}^{(b)}),\bar{\boldsymbol{F}}_t^1\right)+\right.\\&\left.\mathcal{L}_{\mathrm{ce}}\left(\boldsymbol{F}(\boldsymbol{\omega}_t,\hat{\boldsymbol{s}}^{(b)}),\bar{\boldsymbol{F}}_t^0\right)\right)\Big]\end{aligned} \tag{12}$$

由于梯度大小没有先验界限，L2 范数中的每个 g_{θ_t} 和 g_{ω_t} 都被裁剪为 $\hat{g}_{\theta_t}$ 和 $\hat{g}_{\omega_t}$，其裁剪阈值为 C_θ 和 C_ω，如式(13)~式(14)所示。

$$\hat{g}_{\theta_t}(\boldsymbol{s}^{(b)})=\frac{g_{\theta_t}(\boldsymbol{s}^{(b)})}{\max\left(1,\dfrac{\|g_{\theta_t}(\boldsymbol{s}^{(b)})\|_2}{C_\theta}\right)} \tag{13}$$

$$\hat{g}_{\omega_t}(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})=\frac{g_{\omega_t}(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})}{\max\left(1,\dfrac{\|g_{\omega_t}(\boldsymbol{s}^{(b)},\hat{\boldsymbol{s}}^{(b)})\|_2}{C_\omega}\right)} \tag{14}$$

步骤 4　为了保护数据隐私，协作服务请求者为每个 $\hat{g}_{\theta_t}$ 和 $\hat{g}_{\omega_t}$ 加入精心设计的高斯噪声 $\mathcal{N}\left(0,\sigma_\theta^2C_\theta^2I\right)$ 和 $\mathcal{N}\left(0,\sigma_\omega^2C_\omega^2I\right)$，以获得平均噪声梯度 $\tilde{g}_{\theta_t}$ 和 $\tilde{g}_{\omega_t}$。最后，$\mathcal{D}$ 和 $\mathcal{C}$ 的权重 $\boldsymbol{\omega}_t$ 和 $\boldsymbol{\theta}_t$ 按照 $\tilde{g}_{\omega_t}$ 和 $\tilde{g}_{\theta_t}$ 的负方向更新为 $\boldsymbol{\omega}_{t+1}=\boldsymbol{\omega}_t-\alpha_\omega\tilde{g}_{\omega_t}$ 和 $\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t-\alpha_\theta\tilde{g}_{\theta_t}$。

基于 DFD 的协作 IDS 模型训练不仅提高了请求者生成器的生成泛化能力，而且增强了判别器和外部分类器模型的分类性能，这是因为 PFD-GAN 的预测能提供比作为正则化的独热标签更多的有用信息，这将在第 5 节中证明与讨论。

5　实验及结果分析

5.1　实验环境与实验参数

关于仿真参数（如表 1 所示），根据实验数据集的特征维度，潜在空间 z 的维度设置为 24，其中前 23 维为特征，最后一维为对应标签，各维度值均落在$[-1,1]$的范围内；IDS 采用类似于第 3 节中介绍的 Dev EC-GAN 模型；使用 $B=\dfrac{|\boldsymbol{S}|}{T}$ 作为训练样本的批量采样数，确保所有的训练数据都能够被模型所学习；生成器、判别器和外部分类器的学习率，即 α_φ、α_ω 和 α_θ 被设置为 5×10^{-4}；裁剪阈值 C_ω 和 C_θ 被设置为 0.01；本地迭代次数 $T_{[\varphi,\omega,\theta]}$ 和全局迭代次数 T 被设置为[50,5,100]和 50，其中 $T=T_\varphi=\dfrac{T_\theta}{2}$；无监督损失权重 λ 和伪标签阈值 τ 分别被设置为 0.1 和 0.7；LDP 的噪声尺度 (ε,δ) 被设置为 $(6,1\times10^{-5})$。以上参数的设置是在进行多次实验后统计得出的经验性设置。在 5.4 节中，将对关键参数进行讨论。

表 1　　仿真实验参数

参数名	参数值
潜在空间纬度 $\dim(z)$ /维	24
批量采样数 B /个	$B=\frac{\lvert \boldsymbol{S} \rvert}{T}$
学习率 $\alpha_{\varphi/\omega/\theta}$	5×10^{-4}
裁剪阈值 C_ω, C_θ	0.01
本地迭代次数 $T_{[\varphi,\omega,\theta]}$ /次	[50,5,100]
全局迭代次数 T /次	$T=T_\varphi=\frac{T_\theta}{2}$
无监督损失权重 λ	0.1
伪标签阈值 τ	0.7
LDP 噪声尺度 (ε,δ)	$(6,1\times10^{-5})$

关于实验数据集方面，采用了一个真实的工业 CPS 数据集，即储水罐系统（WSTS，工业 CPS 的一个重要示例）数据集[19]。WSTS 数据集集合中的每个样本均由一个包含 23 个属性的特征向量组成，其中最后一个属性是相应的类别或标签。每个标签确定了样本是正常流量 Benign（73%）或一种攻击类型，即 Reconnaissance（14.6%）、CMRI（5.5%）、DOS（0.5%）、MFCI（0.6%）、MSCI（0.7%）、MPCI（1.5%）或 NMRI（3.6%）。显然，稀疏攻击分类 DOS、MFCI、MSCI、MPCI 和 NMRI 远少于其他分类类别，占比总数均不足 7%，数据集各分类间存在严重不均衡问题。

为了针对更多类型的网络入侵攻击进行分析与检测，同时采用了爱琴海无线入侵数据集 （AWID, Aegean Wi-Fi intrusion dataset）[20]，进一步验证提出的 PFD-GAN 的有效性。AWID 各分类占比分别为 Normal（91%）、Flooding（3.6%）、Impersonation（2.7%）和 Injection（2.7%）。相比正常流量，其他 3 种攻击分类为稀疏类，占比均不足总数的 4%。在仿真实验中，WSTS 和 AWID 数据集分为 2 个主要部分，即 80%的训练数据和 20%的测试数据，训练数据平均给 5 个不同的对等工业 CPS 节点用以协作入侵检测系统的训练，并使用相同的测试数据来评估所有的训练模型，评估结果取所有模型的算术平均值。

5.2　效能评估与结果分析

为了验证提出的 PFD-GAN 及各改进点的有效性，在真实 CPS 数据集 WSTS 上，针对卷积神经网络（CNN, convolutional neural network）、原始 EC-GAN、Dev EC-GAN、所有 CPS 统一采用如图 3 所示的 Fix PFD-GAN 和各 CPS 采用最佳实践定制 Dev EC-GAN 模型的 PFD-GAN 进行对比，验证半监督学习和联邦蒸馏协作对分类性能的提升。结果在 2 种不同的情况下呈现，即使用和不使用 LDP 技术，并在表 2 中根据评估参数 Precision、Recall、F1 值、FAR （false alarm rate）、FLOP（floating point operations per second）和推理时延（Latency）进行比较验证。在 2 种情况下均可观察到，在检测性能方面，PFD-GAN 的 Precision、Recall、F1 值和 FAR 是最有优势的，而 Dev EC-GAN 比 EC-GAN 和 CNN 也有明显的提升；在检测效率方面，所提出的 PFD-GAN 与 CNN 相比，因为仅在训练时利用泛化能力强化模型性能，并未增加模型检测的计算复杂度和推理时延，而且由于采用定制结构提升了 IDS 的检测效率。具体地，图 4 进一步展示了 WSTS 下不同稀疏攻击类的 F1 值。很明显，稀疏攻击类，即 DoS、MFCI、MSCI、MPCI 和 NMRI 的 F1 值在采用 PFD-GAN 后获得了显著的提升。

表 2　　WSTS 数据集下不同 IDS 模型的效能评估

IDS 模型		Precision	Recall	F1 值	FAR	FLOPS/(flop·s^{-1})	Latency/ms
不使用 LDP 技术	CNN	95.62%	94.56%	94.92%	5.41%	89 216	4.29
	EC-GAN	96.55%	95.89%	96.11%	3.93%	89 216	4.20
	Dev EC-GAN	97.13%	96.76%	96.81%	2.20%	89 216	4.26
	Fix PFD-GAN	98.65%	98.62%	98.60%	1.14%	89 216	4.31
	PFD-GAN	99.00%	98.99%	98.97%	0.79%	74 470	3.58
使用 LDP 技术	CNN	93.76%	92.77%	93.00%	7.85%	89 216	4.36
	EC-GAN	95.63%	94.14%	94.68%	4.57%	89 216	4.34
	Dev EC-GAN	96.41%	95.60%	95.82%	3.03%	89 216	4.38
	Fix PFD-GAN	97.94%	97.75%	97.77%	1.54%	89 216	4.29
	PFD-GAN	98.16%	97.97%	97.99%	1.47%	74 470	3.49

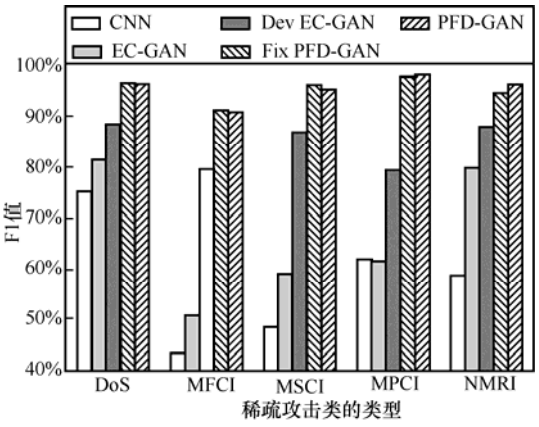


图 4　WSTS 下不同稀疏攻击类的 F1 值

此外，采用 AWID 数据集对 CNN、EC-GAN、Dev EC-GAN、Fix PFD-GAN 和 PFD-GAN 进行对比实验，以针对更多类型的网络入侵攻击进行分析与验证。比较结果如图 5~图 6 所示。Jaccard 系数（即 J_i, i 为分类类别）被用于评估 IDS 在检测不同攻击类别时的分类性能，对于 AWID 数据集中的每个攻击类型，可以计算其与数据集整体之间的 Jaccard 系数，评估 IDS 的准确性和鲁棒性。如果某个攻击类型的 Jaccard 系数较高，则说明入侵检测系统对该类型攻击的检测效果较好。与 CNN 相比，EC-GAN、Dev EC-GAN 和 PFD-GAN 的 $\{J_i\}_i$ 均实现阶梯式的上升，其中 PFD-GAN 在所有类别的度量都取得最佳结果。特别地，对于稀疏攻击类，PFD-GAN 的 J_{flooding}、$J_{\text{injection}}$ 和 $J_{\text{impersonation}}$ 得到了显著的提升。类似地，在检测效率方面，采用定制模型的 PFD-GAN 也得到了整体上的 FLOPS 和 Latency 优化。

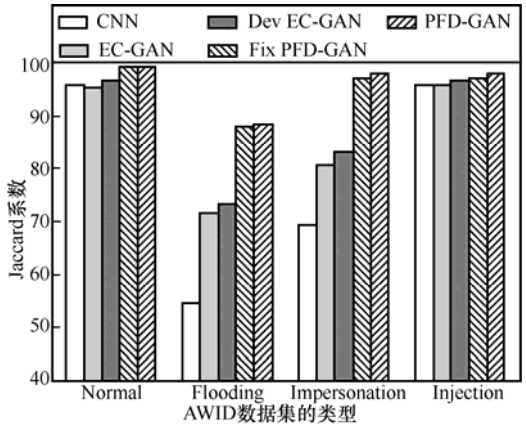


图 5　AWID 下不同 IDS 模型的 Jaccard 系数性能评估

上述结论主要归因于所提出的 PFD-GAN 中 Dev EC-GAN 继承和发展了生成可用数据的能力，以缓解 WSTS 和 AWID 数据集中存在的分类不均衡问题。更重要的是，PFD-GAN 中提出的基于 DFD 的协作训练策略，使多个工业 CPS 能够在完整的数据集上共同构建一个综合的 IDS。此外，可以注意到采用定制结构的 PFD-GAN 的检测性能和效率均优于固定结构的 Fix PFD-GAN，在实际的使用中，不同 CPS 可以根据自身需求，使用定制化的 IDS 结构，摆脱联邦学习需要固定模板的限制。因此可证所提出的 PFD-GAN 非常适合作为工业 CPS 的协作入侵检测方法。

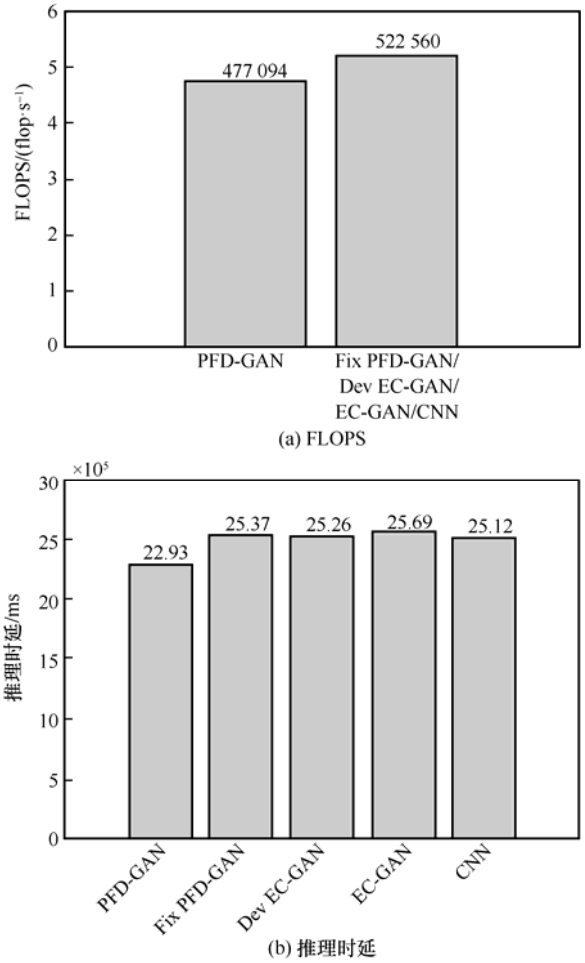


图 6　AWID 下不同 IDS 模型的计算复杂度和实时性对比

5.3　对比实验与结果分析

本节将 2 个具有高引用率和 2 个较新的 IDS 模型，即 DL-GAN[21]、MD-BiGAN[12]、Fed-ANIDS[22]和 SSFL[23]，与提出的 PFD-GAN 在 WSTS 和 AWID 数据集下分别进行了对比实验。对比 IDS 模型的复现采用相关文献中的方法和代码在 5.1 节的实验环境下进行，以控制对比结果中的可变变量仅为不同的 IDS 模型，确保对比结果的有效性。WSTS 数据集下 PFD-GAN 和对比

IDS 模型的对比结果如图 7 所示，当迭代次数 T 从 10 增加到 50 时，PFD-GAN 的 Precision、Recall 和 FAR 均优于 DL-GAN[21]、MD-BiGAN[12]、Fed-ANIDS[22]和 SSFL[23]的相应评估值，此外针对稀疏攻击类的检测 PFD-GAN 的 F1 值也优于其余对比 IDS 模型的评估值。AWID 数据集下的对比结果如表 3 所示，其展示了在 $T=50$ 时 5 种不同的 IDS 模型的针对不同类型攻击的 Jaccard 系数对比。从表 3 可知，PFD-GAN 的分类性能和检测效率均优于其他 4 种对比 IDS 模型，特别是在稀

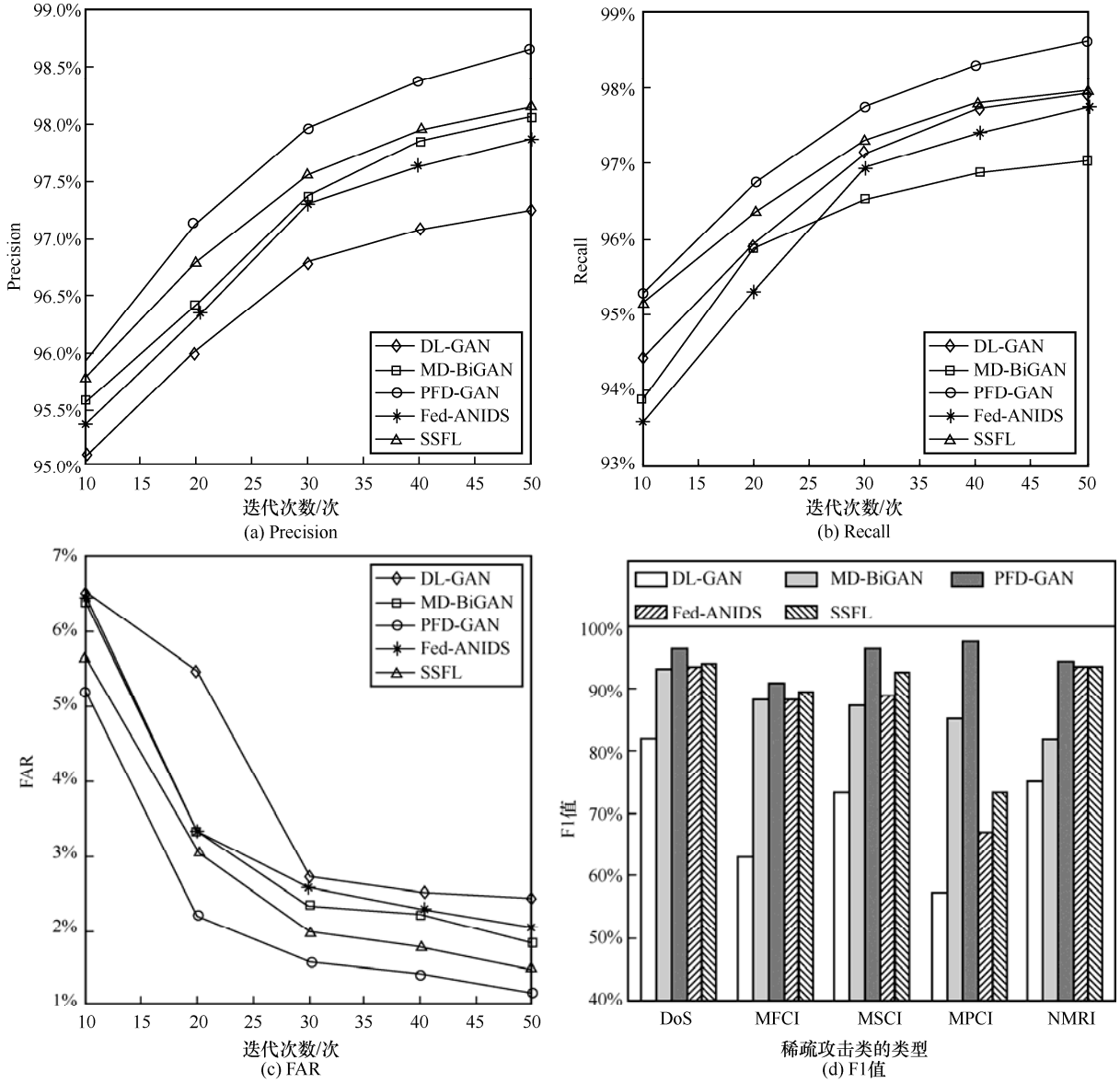


图 7　WSTS 数据集下 PFD-GAN 和对比 IDS 模型的比较

表 3　　AWID 下 PFD-GAN 和对比 IDS 模型的效率和性能比较

IDS 模型	J_{normal}	J_{flooding}	$J_{\text{impersonation}}$	$J_{\text{injection}}$	FLOPS/(flop·s^{-1})	Latency/ms
DL-GAN	96.80%	54.28%	96.39%	95.65%	5.22×10^5	25.12
MD-BiGAN	98.11%	74.35%	82.97%	96.62%	5.14×10^5	24.73
Fed-ANIDS	98.33%	71.55%	80.58%	96.84%	2.51×10^6	120.97
SSFL	98.70%	83.33%	91.21%	96.69%	2.48×10^6	119.40
PFD-GAN	99.30%	87.94%	97.11%	97.02%	4.77×10^5	22.93

疏攻击类 J_{flooding}、$J_{\text{injection}}$ 和 $J_{\text{impersonation}}$ 的检测性能上的表现尤为突出。

这主要是因为 PFD-GAN 的 Dev EC-GAN 不仅采用深度学习网络作为区分攻击与正常流量的分类器，同时也作为生成器产生可用样本，作为稀疏攻击类的补充样本以增强分类性能。更重要的是，在 PFD-GAN 中设计了基于 DFD 的协作训练策略，允许不同工业 CPS 采用定制的检测模型，共同训练综合的 IDS 模型。然而，DL-GAN[21]没有任何协作机制来处理数据集的不均衡问题，Fed-ANIDS[22]和 SSFL[23]的联邦协作过程需要借助统一的检测模型模板，这很大程度地限制了不同检测环境下模型性能的发挥，而 MD-BiGAN[12]的协作仅从多个生成器和编码器到中央判别器单向进行，即只有模型的生成性能得到提升。因此，本文所提出的 PFD-GAN 被证明在所有涉及的评估参数上均比这 4 个代表性 IDS 有着更好的入侵检测性能。值得注意的是，虽然 PFD-GAN 的检测性能优于其他的对比方法，但对 Flooding 攻击的检测率不高，这主要是由于 1) Flooding 攻击（3.6%）是显著的稀疏类，检测模型对这种类别的训练不足，难以学习到稀疏类的特征和规律；2) 相较于其他 2 种攻击类型，Flooding 攻击更类似于正常的网络消息，且可以通过分布式的方式发起，因此防御难度更大，更难以被正确检测和甄别。

此外，表 4 提供了 WSTS 数据集下多种现行的 IDS 解决方案效率和性能的比较，包括 3 个部分：1) 无任何辅助机制的 IDS；2) 带辅助机制的 IDS；3) 差分隐私保护的联邦 IDS。在前两部分的对比实验中，与其他现行的 IDS 解决方案相比，所提出的 PFD-GAN 凭借联邦协作建模的优势，取得了 Precision、Recall 和 F1 值的最高值与 FAR 的最低值，推理时延仅高于假设检验等直接判别法 3 ms。在最后的对比实验部分，由于提出的 PFD-GAN 采用了基于分布式联邦蒸馏的协作建模策略，有效地规避了聚合中心和模板统一的缺陷，获得了在 Precision、Recall、F1 值、Latency 和 FAR 评价指标上的最优表现。因此，由实验可证明 PFD-GAN 的异常检测性能优于表 4 中现行的 IDS 解决方案。

5.4　关键参数分析讨论

为将 LDP 技术应用于提出的 PFD-GAN 中，分别采用最常见的高斯噪声和拉普拉斯噪声进行实验，探讨不同噪声对 PFD-GAN 的性能影响，在选定噪声类型后，进一步探讨不同噪声尺度的影响和尺度参数的选择。高斯噪声采用均值为零（即无偏移）且具有多个标准差 σ 的尺度参数以实现 (ε,δ)-LDP，其中 $\sigma=2\dfrac{\sqrt{\ln\left(\dfrac{5}{4\delta}\right)}}{\varepsilon}$；拉普拉斯噪声主要受隐私预算 ε [24]的影响，隐私预算代表了隐私的泄露程度。

不同噪声类型对 PFD-GAN 的性能影响如图 8 所示。通过对图 8(a)～图 8(b)的纵向对比分析可以直观看出，无论采用哪一个数据集，当 $\delta=10^{-5}$ 时，高斯噪声对 PFD-GAN 的影响最小；通过对图 8(a)～图 8(c)的横向对比分析可知，采用 $(\varepsilon\sim\{1,9\},\delta=1\times10^{-5})$ 的高斯噪声的 PFD-GAN 的检测性能表现要优于采用

表 4　　WSTS 数据集下多种现行 IDS 解决方案的效率和性能的比较

IDS 解决方案		Precision	Recall	F1 值	FAR	Latency/ms
无任何辅助机制的 IDS	GHSOM	96.17%	96.11%	95.99%	5.48%	3.40
	Hypothesis Testing	94.28%	94.00%	93.96%	7.15%	0.43
	Dolphine + SVM	95.99%	95.97%	95.87%	5.59%	1.44
	Markov Model	91.33%	93.17%	92.00%	10.40%	1.56
带辅助机制的 IDS	GHSOM + MOEA	97.43%	97.03%	97.01%	2.70%	2.03
	t-test + Bayesian	96.46%	96.18%	96.08%	5.42%	0.57
	PFD-GAN w/o LDP	98.65%	98.62%	98.60%	1.14%	3.52
差分隐私保护的联邦 IDS	FedAvg	93.20%	92.95%	92.82%	7.96%	5.74
	Fed+	93.49%	93.05%	93.09%	7.68%	5.89
	SecFedNIDS	95.89%	94.82%	95.06%	3.65%	4.28
	PFD-GAN w LDP	97.94%	97.75%	97.77%	1.54%	3.65

$\varepsilon \sim \{0.30,8\}$ 的拉普拉斯噪声的情况，因此高斯噪声更适合成为实现 LDP 技术的噪声类型。

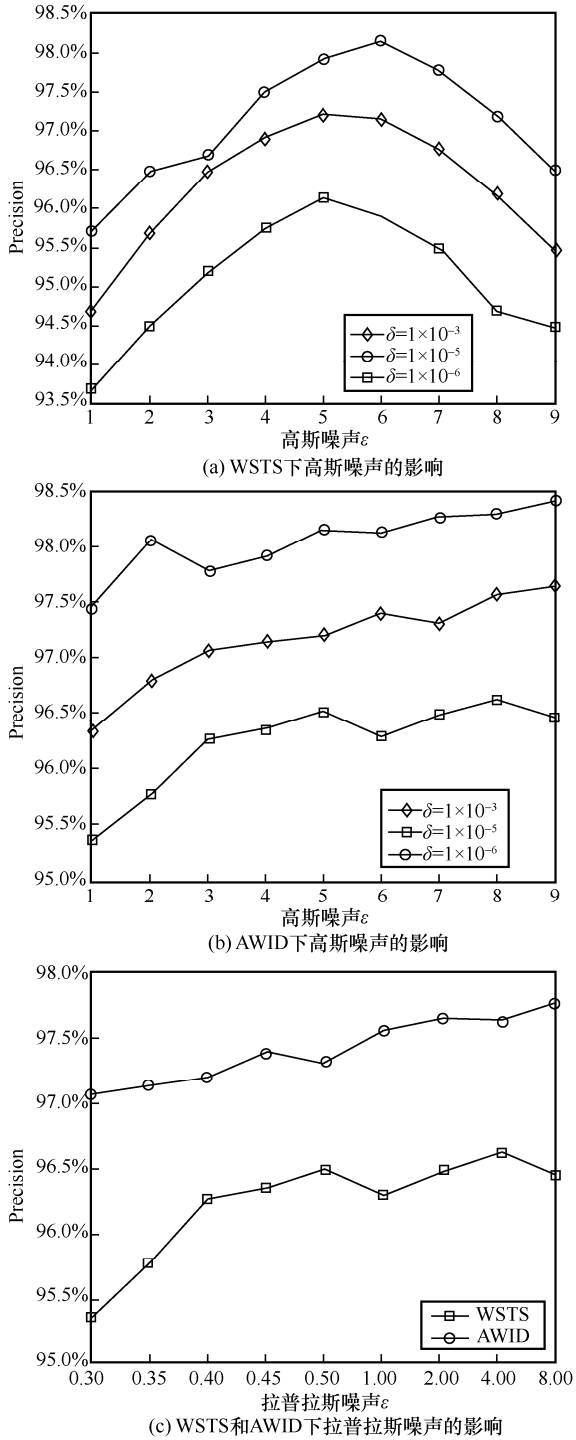


图 8　不同类型噪声对 PFD-GAN 的影响

在选定噪声类型后，需要进一步探讨不同噪声尺度和尺度参数的影响。不同噪声尺度下所提出的 PFD-GAN 的性能曲线（包括受试者特征曲线（ROC）、镜像 ROC 和 Precision 变化曲线）如图 9 所示，其反映了 PFD-GAN 在不同参数阈值下的真阳性率（TPR）、真阴性率（TNR）、假阳性率（FPR）和 Precision 变化。在图 9 中，PFD-GAN 被分配了 3 组不同的 (ε,δ) 参数向量，分别为 $(1, 1\times10^{-5})$、$(6, 1\times10^{-5})$ 和 $(7, 1\times10^{-5})$，用于展示参数变化与性能的关系。与其他两组 (ε, δ) 相比，采用 $(6, 1\times10^{-5})$ 作为噪声尺度的模型获得了更大的 ROC 曲线下面积（AUC）。此外，当 $\varepsilon=6$ 且 $\delta=1\times10^{-5}$ 时，PFD-GAN 获得最高的 Precision。换句话说，噪声尺度 $(6, 1\times10^{-5})$ 是一个适合的 (ε,δ) 向量。

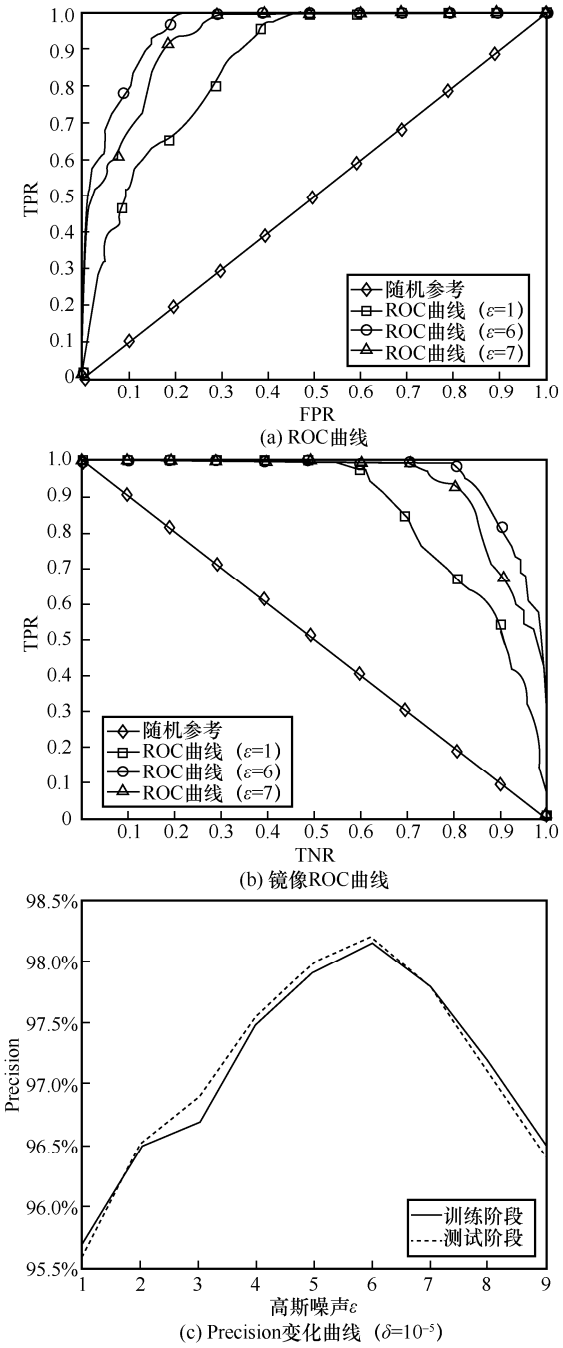


图 9　在不同噪声尺度下 PFD-GAN 的性能曲线

6　结束语

本文提出了基于安全联邦蒸馏生成对抗网络的工业 CPS 协作入侵检测方法，称为 PFD-GAN。首先，提出了一种的半监督 IDS 模型，该模型采用 Wasserstein 距离和标签条件改进了 EC-GAN 模型，以提高其在不均衡数据集的分类性能。同时，利用 LDP 来防止 IDS 在协作训练中敏感信息的泄露。此外，设计了基于 DFD 的协作策略，允许多个工业 CPS 节点构建一个综合的 IDS 以检测整体 CPS 下的异常行为，而无须共享统一的神经网络模板模型。对真实工业 CPS 数据集的实验对比和理论分析证明了 PFD-GAN 在检测异常行为能力方面的有效性以及与当前先进 IDS 解决方案相比的优越性。未来的工作将针对更多类型的网络入侵攻击进行分析与检测。

参考文献：

[1] PIVOTO D G, ALMEIDA L F, RIGHI R, et al. Cyber-physical systems architectures for industrial Internet of things applications in industry 4.0: a literature review[J]. Journal of Manufacturing Systems, 2021, 58: 176-192.

[2] SALAU B A, RAWAL A, RAWAT D B. Recent advances in artificial intelligence for wireless Internet of things and cyber–physical systems: a comprehensive survey[J]. IEEE Internet of Things Journal, 2022, 9(15): 12916-12930.

[3] 吕思才, 张格, 张耀方, 等. 一种面向工控系统的 PU 学习入侵检测方法[J]. 信息安全学报, 2021, 6(4): 72-89.
LV S C, ZHANG G, ZHANG Y F, et al. A PU learning intrusion detection method for industrial control system[J]. Journal of Cyber Security, 2021, 6(4): 72-89.

[4] NESPOLI P, DÍAZ-LÓPEZ D, MÁRMOL F G. Cyberprotection in IoT environments: a dynamic rule-based solution to defend smart devices[J]. Journal of Information Security and Applications, 2021, 60: 102878.

[5] RAJASOUNDARAN S, PRABU A V, SUBRAHMANYAM J B V, et al. WITHDRAWN: secure watchdog selection using intelligent key management in wireless sensor networks[J]. Materials Today: Proceedings, 2021: doi.org/10.1016/j.matpr.2020.12.1027.

[6] NAQASH T, TANVEER M H, SHAH S H, et al. Statistical analysis-based intrusion detection for software defined network[C]//Smart Trends in Computing and Communications. Singapore: Springer, 2022: 279-289.

[7] IERACITANO C, ADEEL A, MORABITO F C, et al. A novel statistical analysis and autoencoder driven intelligent intrusion detection approach[J]. Neurocomputing, 2020, 387: 51-62.

[8] ALDRIBI A, TRAORÉ I, MOA B, et al. Hypervisor-based cloud intrusion detection through online multivariate statistical change tracking[J]. Computers & Security, 2020, 88: 101646.

[9] YAO R Z, WANG N, LIU Z H, et al. Intrusion detection system in the advanced metering infrastructure: a cross-layer feature-fusion CNN-LSTM-based approach[J]. Sensors, 2021, 21(2): 626.

[10] CHADZA T, KYRIAKOPOULOS K G, LAMBOTHARAN S. Analysis of hidden Markov model learning algorithms for the detection and prediction of multi-stage network attacks[J]. Future Generation Computer Systems, 2020, 108: 636-649.

[11] LI W, MENG W. Collaborative intrusion detection in the era of IoT: recent advances and challenges[J]. Security and Privacy in the Internet of Things: Architectures, Techniques, and Applications, 2021: 123-149.

[12] SHU J G, ZHOU L, ZHANG W Z, et al. Collaborative intrusion detection for VANETs: a deep learning-based distributed SDN approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(7): 4519-4530.

[13] RUZAFA-ALCÁZAR P, FERNÁNDEZ-SAURA P, MÁRMOL-CAMPOS E, et al. Intrusion detection based on privacy-preserving federated learning for the industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2021, 19(2): 1145-1154.

[14] YU P Q, WYNTER L, LIM S H. Fed+: a family of fusion algorithms for federated learning[J]. arXiv Preprint, arXiv: 2009.06303, 2020.

[15] ZHANG Z, ZHANG Y, GUO D, et al. SecFedNIDS: robust defense for poisoning attack against federated learning-based network intrusion detection system[J]. Future Generation Computer Systems, 2022, 134: 154-169.

[16] ABDEL-BASSET M, MOUSTAFA N, HAWASH H, et al. Federated intrusion detection in blockchain-based smart transportation systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(3): 2523-2537.

[17] SHENDE O, PATERIYA R K, VERMA P, et al. CEBM: collaborative ensemble blockchain model for intrusion detection in IoT environment[J]. Europe PMC, 2021: doi.org/10.21203/rs.3.rs-702181/v1.

[18] AGRAWAL S, SARKAR S, AOUEDI O, et al. Federated Learning for intrusion detection system: concepts, challenges and future directions[J]. Computer Communications, 2022, 195: 346-361.

[19] MORRIS T, GAO W. Industrial control system traffic data sets for intrusion detection research[C]//International Conference on Critical Infrastructure Protection. Berlin: Springer, 2014: 65-78.

[20] CHATZOGLOU E, KAMBOURAKIS G, KOLIAS C. Empirical eval-

uation of attacks against IEEE 802.11 enterprise networks: the AWID3 dataset[J]. IEEE Access, 2021, 9: 34188-34205.

[21] XU C G, REN J, ZHANG D Y, et al. GANobfuscator: mitigating information leakage under GAN via differential privacy[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(9): 2358-2371.

[22] ZHAO R J, WANG Y J, XUE Z, et al. Semisupervised federated-learning-based intrusion detection method for Internet of things[J]. IEEE Internet of Things Journal, 2023, 10(10): 8645-8657.

[23] IDRISSI M J, ALAMI H, MAHDAOUY A E, et al. Fed-ANIDS: federated learning for anomaly-based network intrusion detection systems[J]. Expert Systems With Applications, 2023, 234: 121000.

[24] PHAN N, WU X T, HU H, et al. Adaptive laplace mechanism: differential privacy preservation in deep learning[C]//Proceedings of 2017 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2017: 385-394.

[作者简介]

梁俊威（1992– ），男，广东深圳人，博士，深圳信息职业技术学院讲师，主要研究方向为信息安全、人工智能、无线通信网络等。

杨耿（1986– ），男，广西贵港人，博士，深圳信息职业技术学院高级工程师，主要研究方向为无线通信网络、模式识别等。

马懋德（1964– ），男，加拿大人，博士，南洋理工大学教授、博士生导师，主要研究方向为无线通信网络、信息安全、人工智能等。

Muhammad Sadiq（1982– ），男，巴基斯坦人，博士，深圳信息职业技术学院助理教授，主要研究方向为人工智能、模式识别、信息安全等。