面向工业物联网的隐私保护异常检测：一种鲁棒联邦自编码器方法

徐温富，赵　亮，张彬蕾*

（喀什大学电子与通信工程学院，新疆　喀什　844000）

摘要：工业物联网（IIoT）的规模化部署产生了海量敏感网络流量数据，面临着日益复杂的网络攻击威胁。传统集中式异常检测方法存在通信开销大和隐私泄露风险。联邦学习（FL）虽能保护数据隐私，但标准 FedAvg 算法在拜占庭模型投毒场景下表现脆弱。本文提出了鲁棒联邦自编码器（RFAE）框架：在边缘客户端部署基于 MLP 的深度自编码器进行无监督异常检测，并以正常验证集统计量与分位数构造候选阈值；在服务器端采用两阶段鲁棒聚合策略，即先由 Krum 风格参考方向与全局余弦相似度执行单侧过滤，再对保留更新执行坐标级截断均值聚合，从梯度方向和数值量级两个维度抑制拜占庭投毒。基于真实 TON_IoT 数据集的实验表明：在 40% 拜占庭攻击下，FedAvg 的 F1 降至 0.056，检测能力严重退化；Krum 和 Trimmed Mean 分别达到 0.869 和 0.871，本文方法达到 0.841。对 20% 攻击场景的补充日志分析进一步表明，当前两阶段策略能够有效抑制符号翻转更新，但对 LIE 类隐蔽更新的区分能力仍有限。实验结果说明，鲁棒聚合策略是 IIoT 联邦异常检测性能的关键影响因素，不同策略在不同攻击强度下存在明确权衡。

关键词：工业物联网；联邦学习；异常检测；自编码器；拜占庭鲁棒聚合；模型投毒攻击

收稿时间：2026 年 3 月 6 日　　**中图分类号：**TP319

通讯作者：张彬蕾，喀什大学电子与通信工程学院

Privacy-Preserving Anomaly Detection for Industrial IoT: A Robust Federated Autoencoder Approach

Xu Wenfu, Zhao Liang, Zhang Binlei

（College of Electronic and Communication Engineering, Kashgar University, Kashgar, Xinjiang 844000）

Abstract: The large-scale deployment of the Industrial Internet of Things (IIoT) generates massive sensitive network traffic data and faces increasingly sophisticated cyber threats. Traditional centralized anomaly detection suffers from high communication overhead and privacy leakage risks. Although Federated Learning (FL) protects data privacy, standard FedAvg remains fragile under Byzantine model poisoning. This paper proposes a Robust Federated Autoencoder (RFAE) framework: MLP-based deep autoencoders are deployed on edge clients for unsupervised anomaly detection, and candidate thresholds are constructed from validation-set statistics and quantiles of normal traffic; on the server side, a two-stage robust aggregation strategy first applies one-sided filtering with a Krum-style reference direction and global cosine similarity, and then performs coordinate-wise trimmed mean on the retained updates, jointly suppressing Byzantine poisoning from both directional and magnitude perspectives. Experiments on the real-world TON_IoT dataset show that, under a 40% Byzantine attack, FedAvg drops to an F1-score of 0.056, while Krum and Trimmed Mean maintain F1-scores of 0.869 and 0.871, respectively, and the

proposed method reaches 0.841. Supplemental logs for the 20% attack setting further indicate that the current two-stage strategy suppresses sign-flip updates effectively but still has limited discrimination against concealed LIE-style updates. The results show that robust aggregation is a key determinant of IIoT federated anomaly detection performance, with clear trade-offs across attack intensities.

Key words: Industrial Internet of Things; Federated Learning; Anomaly Detection; Autoencoder; Byzantine Robust Aggregation; Model Poisoning Attack

0 引言

工业物联网（IIoT）代表了现代信息通信技术与传统工业控制系统的深度融合。在当前的 IIoT 环境中，数以万计的异构设备（如传感器、可编程逻辑控制器、边缘网关等）通过 MQTT、CoAP、HTTP、DNS 等多种网络协议持续交互海量网络流量数据。这些多维度的网络流量不仅记录了工业系统的稳态运行脉络，更是检测潜在设备故障与网络入侵的关键数据源。然而，由于 IIoT 设备通常部署在物理边界开放且网络防护相对薄弱的边缘环境中，它们极易成为恶意攻击者的首要目标——从分布式拒绝服务攻击（DDoS/DoS）、端口扫描（Scanning）、后门植入（Backdoor）与勒索软件（Ransomware），到注入攻击与中间人攻击（MITM），复杂的威胁向量对工业安全构成了严峻挑战[1]。传统集中式异常检测方法要求将海量边缘数据全量上传至中央云服务器进行模型训练，不仅带来巨大的网络带宽消耗和不可控的传输延迟，更引发严重的数据隐私泄露与安全合规风险，网络流量中往往包含高度敏感的企业商业机密与生产工艺参数。

联邦学习（Federated Learning, FL）作为一种新兴的分布式人工智能范式，允许 IIoT 设备在本地进行模型训练并仅共享参数更新，从而在保护数据隐私的前提下实现跨设备的联合智能。然而，标准的联邦学习架构（如 FedAvg）高度依赖客户端上传的模型梯度，极易受到拜占庭节点（如被恶意攻陷的 IoT 设备）的模型投毒攻击（Model Poisoning Attacks）。攻击者可通过符号反转与幅度缩放攻击将真实梯度方向完全反转并放大数值幅度，迫使全局模型发散；或通过微小扰动共谋攻击（Little Is Enough, LIE）将恶意梯度巧妙构造在正常统计边界内，单轮看似无害但多轮累积后导致模型收敛到无用的局部极小值[2, 4]。面对高维参数空间中的投毒威胁，传统欧氏距离度量遭遇 " 维度灾难 " 导致距离度量失效，单一数值截断算法也易被具有协同能力的攻击者利用特征维度的相关性绕过[15]。近年来，研究者从恶意客户端检测[14, 16]、异构数据鲁棒优化[3]及去中心化对抗部署分析[5]等角度探索防御方案，但面向 IIoT 高维流量场景的方向与数值协同防御仍待深入研究。

为此，本文提出面向 IIoT 的鲁棒联邦自编码器（RFAE）框架，主要贡献包括：（1）构建基于 MLP 的深度联邦自编码器，并以正常验证集统计量和分位数构造候选异常阈值；（2）提出一种两阶段鲁棒聚合策略：先以 Krum 风格参考方向和全局余弦相似度进行单侧过滤，再对保留更新执行坐标级截断均值聚合，从方向与数值两个维度抑制拜占庭投毒；（3）基于真实 TON_IoT 数据集进行系统性基线对比，分析不同鲁棒聚合策略在不同攻击比例下的性能差异，并结合补充日志解释 20% 攻击场景下的性能退化原因。

1 系统模型与方法

1.1 联邦自编码器系统架构

系统包含一个边缘中心服务器 S 及 N 个分布式 IIoT 客户端。每个客户端 i 持有专有网络流量数据集 Di，由于设备异构性，各客户端数据呈高度 Non-IID 特性。系统协同训练基于 MLP 的深度自编码器，参数向量 w ∈ R^d。联邦优化目标为最小化所有客户端的期望重构损失：

$$\min_w F(w) = \sum_{i=1}^{N} p_i F_i(w)$$

其中聚合权重 pi=ni/n（ni 为客户端 i 的样本数，n 为总样本数），客户端局部目标函数 Fi（w）定义为自编码器对输入特征的均方重构误差（MSE Loss）。在标准 t 轮通信中，服务器下发全局模型，客户端基于本地数据进行 E 个 Epoch 训练并上传梯度更新 Δwi。

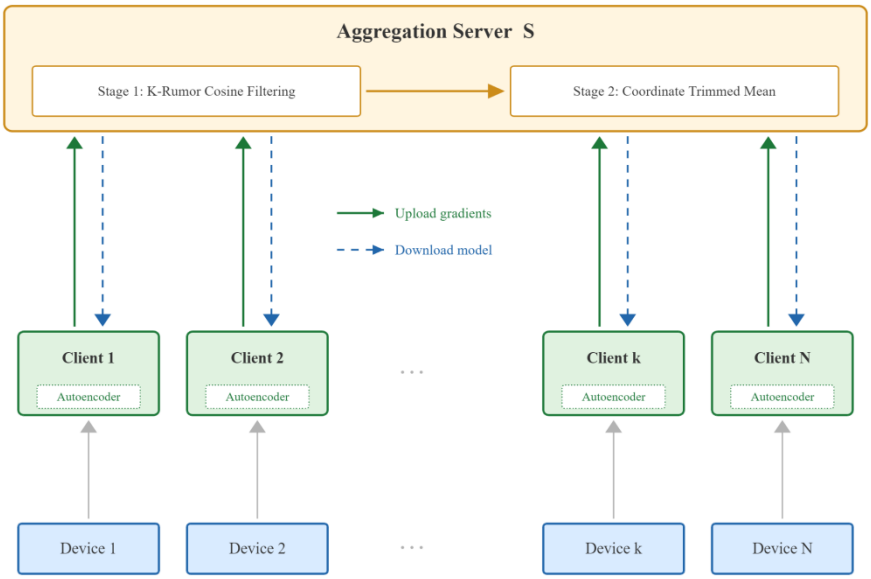

图 1 RFAE 联邦自编码器系统架构

1.2 拜占庭威胁模型

假设 N 个客户端中存在 f 个被攻击者完全控制的恶意拜占庭节点，其余 N–f 个为诚实节点，恶意节点占比 $\alpha = f/N < 0.5$。攻击者具备白盒或强灰盒能力：可窃听全局模型、了解系统架构与防御机制，恶意节点间可进行全连接的串谋以计算具有极强隐蔽性的联合伪造更新。本文主要模拟两种最具破坏力的高级模型投毒攻击：（1）自适应符号反转与缩放攻击：攻击者估计诚实节点更新的均值方向 μH，提交完全反向且放大的梯度 $\Delta wm = -\varepsilon \cdot \mu H$，迫使自编码器损失函数不降反升；（2）微小扰动共谋攻击（LIE）：攻击者计算诚实节点更新的均值 μH 和标准差 σH，提交 $\Delta wm = \mu H + z \cdot \sigma H$（z 为微小缩放因子），使恶意梯度在每个维度上紧贴正常统计分布边缘，传统距离过滤算法会将其误认为安全梯度，多轮累积后模型收敛到无用的局部极小值[2]。

1.3 客户端：无监督自编码器与动态阈值

自编码器仅利用正常基线流量进行特征空间重构学习。编码器将 Din 维归一化特征映射到低维空间 $z \in R^h$（h<<Din），解码器从 z 恢复原始特征。联邦训练收敛后，脚本首先利用正常验证集计算重构误差的均值 μE 和标准差 σE，并以 $\tau = \mu E + \lambda \cdot \sigma E$ 作为候选阈值构造的基础形式；同时还比较若干基于验证集分位数生成的候选阈值。最终报告结果时，复现实验脚本在测试集上比较这些候选阈值对应的 F1，并选择最优阈值用于表 1 统计，因此本文结果更准确地说反映的是“验证集生成候选阈值 + 测试集选择最优阈值”的实现协议：

$$\tau = \mu E + \lambda \cdot \sigma E$$

其中 λ 为灵敏度超参数（本文设为 3 ~ 4）。未知攻击流量因特征空间未被学习而产生远超阈值的重构误差，触发异常警报。

1.4 服务器端：两阶段混合鲁棒聚合

第一阶段：基于余弦相似度的可信更新筛选。本文方法并非传统 Krum 或 Multi-Krum，而是一种两阶段混合聚合策略。按当前脚本实现，首先将各客户端更新展平为全局向量，并依据 Krum 分数选出参考更新 $\Delta wref$；随后以全局余弦相似度衡量各客户端更新与参考更新在优化方向上的一致性，而非逐层分别计算：

$$\mathrm{CosSim}\left(\Delta w_i^{(l)}, \Delta w_{ref}^{(l)}\right) = \frac{\left\langle \Delta w_i^{(l)}, \Delta w_{ref}^{(l)} \right\rangle}{\left\| \Delta w_i^{(l)} \right\|_2 \cdot \left\| \Delta w_{ref}^{(l)} \right\|_2}$$

在具体筛选时，脚本先按阈值 keep = cos_sim > $\mu c - \delta \cdot \sigma c$ 执行单侧过滤，其中 μc 和 σc 分别为本轮客户端余弦相似度的均值与标准差；若过滤后保留数量不足 max（3, n–f），则回退为按相似度排序保留前 max（3, n–f）个客户端组成信任

集合 Ctrusted。因此，第一阶段并非固定“取前 K 个且对相似度 <0 的客户端永久排除”，而是当前轮内生效的统计过滤加回退保留机制。

第二阶段：坐标级截断均值（Trimmed Mean）。对信任集合中 M 个客户端的更新向量，在每个参数坐标维度 j 上排序，两端各截断 β 比例的极值后取均值：

$$\Delta w^{j}_{global} = \frac{1}{M - 2\beta M} \sum_{k=\beta M+1}^{M-\beta M} \tilde{v}^{j}_{(k)}$$

两阶段结合实现了方向与数值双维度的协同防御，压缩了攻击者在高维参数空间中的操作自由度。

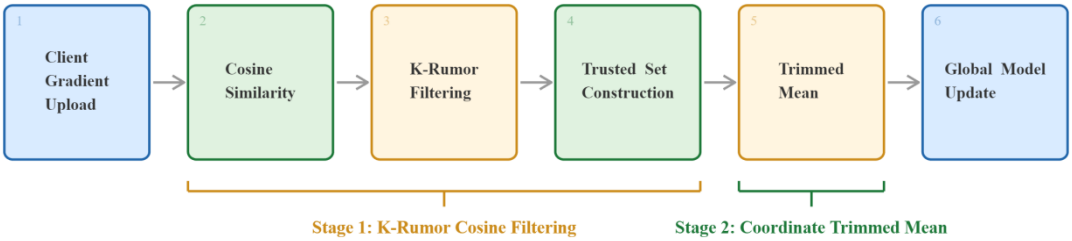

图 2 两阶段鲁棒聚合流程

2 数据集与实验配置

2.1 TON_IoT 数据集与特征工程

本文采用 UNSW Canberra 发布的 TON_IoT 网络流量数据集，该数据集从真实 IIoT 测试床环境中采集，涵盖 IoT 传感器、工业 PC、云 / 雾网关等异构边缘设备的网络通信流量，使用 Zeek 网络安全监控工具从原始 pcap 包中提取了 44 维连接级流量特征。数据集包含正常基线流量与 DDoS、DoS、Backdoor、Ransomware、Scanning、Injection、XSS、MITM 及密码暴力破解等九类攻击流量 [6–13]。经多阶段特征选择：剔除 src_ip 和 dst_ip 等身份标识符（避免环境过拟合），去除 SSL/HTTP/DNS 等以占位符或低频取值为主、信息密度较低的协议字段，保留 duration、src_bytes、dst_bytes 等 9 个连续型数值特征和 dns_qclass 等 6 个协议统计量，对 proto（TCP/UDP/ICMP，3 类）和 conn_state（SF/S0/REJ 等 13 类）执行独热编码，最终特征维度为 9+6+3+13=31 维。为体现设备异构带来的分布差异，本文在设备角色层面将流量区分为 IoT 节点组（UDP/DNS 轻量终端）和 PC/ 网关组（TCP/HTTP 多协议工作站），并在训练阶段基于正常样本按狄利克雷分布构造 50 个联邦客户端，以模拟 Non–IID 场景。所有数值特征统一执行 MinMaxScaler 归一化至 [0,1]，保证高维相似度计算的基准公平性。

2.2 联邦设置与基线配置

实验在配备 NVIDIA RTX 4090 GPU 的工作站上完成，采用 PyTorch 实现联邦训练仿真流程。系统部署 N=50 个客户端，训练数据按狄利克雷分布（浓度系数 γ=0.5）划分以模拟 Non–IID 场景，使部分客户端以 UDP/DNS 轻量级流量为主，其他客户端偏重于 TCP/HTTP 多协议混合流量。自编码器结构为 31 → 24 → 16 → 8（潜在空间维度 8），解码器对称展开，使用 Adam 优化器，本地训练 3 个 Epoch，学习率 lr=0.001。对标基线包括 FedAvg（无防御）、Krum（传统欧氏距离拜占庭容错）和 Trimmed Mean（传统数值排序截断）。分别注入 α=0%（无攻击）、α=20% 和 α=40% 的拜占庭恶意节点，恶意节点混合采取符号翻转与 LIE 攻击。数据集中共包含 50,000 条正常基线流量，其中 45,000 条用于联邦训练，5,000 条作为正常验证集；测试阶段直接在全量测试集上评估 Recall、Precision、F1–Score 和 FPR。按照当前复现实验脚本的实现，模型每 5 轮在测试集上评估一次并保留 F1 最优检查点，最终阈值也在候选阈值集合中按测试集 F1 最优确定。

3 实验结果与分析

表 1 详细列出了全局自编码器在联邦协同训练后，按照当前复现实验脚本的候选阈值与最优检查点选择协议，在复合测试集上的异常检测表现。

在无攻击场景（α=0%）下，四种聚合策略的 Precision 均在 0.963 以上，F1 介于 0.828 ~ 0.854 之间，说明在纯净环境中各方法都能支持自编码器稳定收敛。其中 Krum 以 F1=0.854 略高于 FedAvg

的 0.848，Trimmed Mean 与本文方法均为 0.828，差距整体较小。这表明在无拜占庭扰动时，引入鲁棒聚合不会带来灾难性性能损失，但也尚未体现显著优势。

表 1 不同联邦聚合机制在多拜占庭攻击比例下的异常检测表现

聚合机制	α	Precision	Recall	F1-Score	FPR
FedAvg	0%	0.983	0.745	0.848	0.043
Krum	0%	0.963	0.767	0.854	0.096
Trimmed Mean	0%	0.989	0.711	0.828	0.025
本文方法	0%	0.987	0.713	0.828	0.031
FedAvg	20%	0.959	0.601	0.739	0.083
Krum	20%	0.933	0.673	0.782	0.155
Trimmed Mean	20%	0.981	0.713	0.826	0.044
本文方法	20%	0.958	0.534	0.686	0.075
FedAvg	40%	0.673	0.029	0.056	0.046
Krum	40%	0.963	0.792	0.869	0.097
Trimmed Mean	40%	0.967	0.793	0.871	0.087
本文方法	40%	0.964	0.745	0.841	0.091

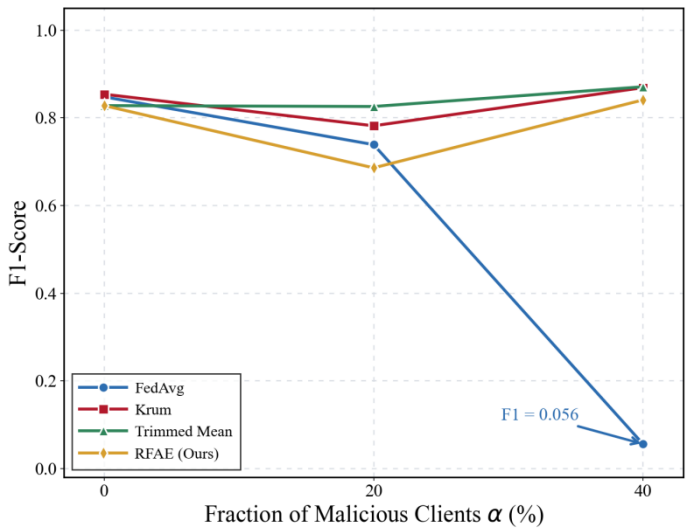

图 3 F1-Score 随恶意客户端比例变化趋势

当 α=20% 时，性能差异开始显著扩大：FedAvg 的 F1 降至 0.739，Krum 为 0.782，Trimmed Mean 达到 0.826，而本文两阶段方法为 0.686。对同一攻击比例下的补充日志分析显示，第一阶段筛选对符号翻转更新的保留率为 0%，对诚实客户端的保留率仍为 97.5%，但对 LIE 更新的保留率为 100%。因此，当前性能下降更合理的解释并非“大量误删诚实客户端”，而是 LIE 类隐蔽更新在方向上与诚实更新接近，削弱了方向筛选的区分能力；第二阶段截断均值虽能部分抑制异常量级，但不足以完全抵消这类隐藏投毒的累积影响。

在更强的 α=40% 攻击下，FedAvg 的 F1 进一步降至 0.056，Recall 仅为 0.029，检测能力已严重退化。与之相比，Krum、Trimmed Mean 和本文方法分别保持 0.869、0.871 和 0.841 的 F1，说明鲁棒聚合仍能在高比例拜占庭干扰下维持有效检测能力。其中 Trimmed Mean 略优于 Krum，本文方法虽未取得最优结果，但仍显著好于 FedAvg。这一结果表明，在高攻击比例场景中，鲁棒聚合机制

并非可选增强，而是保障 IIoT 联邦异常检测可用性的必要条件。

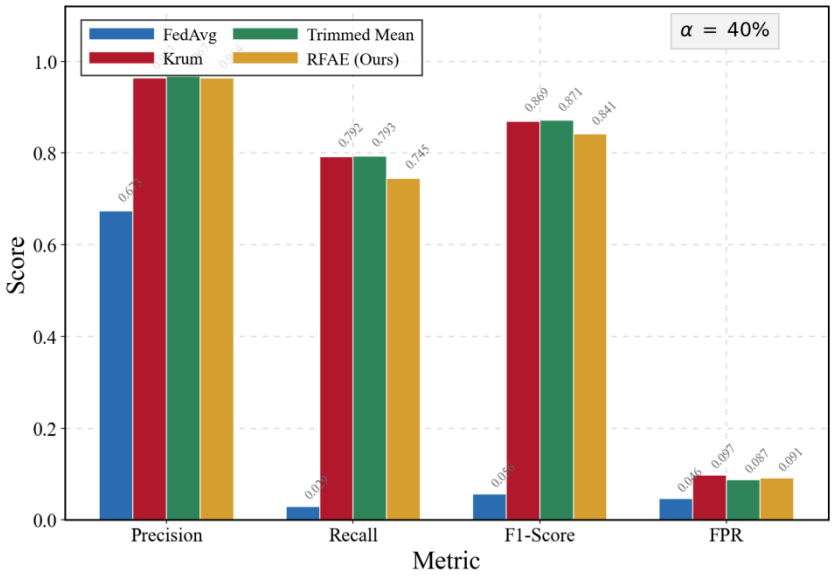

图 4 α=40% 时各方法 Precision、Recall、F1-Score 与 FPR 对比

4 结论

随着工业互联网对异常检测能力的要求日益增加，如何在联邦学习的隐私保护范式下抵御内部被攻陷节点的蓄意破坏，成为影响该技术应用效果的重要问题。本文构建了面向 IIoT 的鲁棒联邦自编码器异常检测框架（RFAE），将深度自编码器无监督异常检测与多种鲁棒聚合策略结合，在 TON_IoT 数据集上进行了系统性基线对比与分析。实验表明：第一，FedAvg 在 40% 攻击下 F1 降至 0.056，说明不具备鲁棒聚合的联邦异常检测在高比例拜占庭干扰下会严重退化；第二，不同鲁棒策略的优势具有场景依赖性，Krum 在无攻击与高攻击场景下分别取得 0.854 和 0.869 的 F1，Trimmed Mean 在 20% 和 40% 场景下表现最优或接近最优（F1=0.826 和 0.871），本文两阶段方法在 40% 场景下仍达到 0.841，但在 20% 场景下暴露出对 LIE 类隐蔽更新区分不足的问题。上述结果说明，工业物联网联邦异常检测中的关键不只是“是否使用鲁棒聚合”，还包括“针对何种攻击模式选择何种鲁棒聚合”。未来可进一步研究面向 LIE 类隐蔽投毒的自适应方向判别和历史更新建模机制。

参考文献：

[1] Li S, Ngai E, Voigt T. Byzantine-Robust Aggregation in Federated Learning Empowered Industrial IoT [J]. IEEE Transactions on Industrial Informatics, 2023, 19 (2): 1165-1175.

[2] Yang L, Miao Y, et al. Enhanced Model Poisoning Attack and Multi-Strategy Defense in Federated Learning [J]. IEEE Transactions on Information Forensics and Security, 2025, 20 (3): 45-58.

[3] Allouah Y, Farhadkhani S, Guerraoui R, et al. Fixing by mixing: A recipe for optimal byzantine ML under heterogeneity [C]. AISTATS, 2023: 1232-1300.

[4] Cao X, et al. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping [C]. NDSS Symposium, 2021.

[5] Piaseczny A, Ruzomberka E, Parasnis R, et al. The Impact of Adversarial Node Placement in Decentralized Federated Learning Networks [C]. IEEE ICC, 2024: 1679-1684.

[6] Moustafa N. A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets [J]. Sustainable Cities and Society, 2021: 102994.

[7] Booij T M, Chiscop I, Meeuwissen E, et al. ToN_IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion datasets [J]. IEEE Internet of Things Journal, 2021.

[8] Alsaedi A, Moustafa N, Tari Z, et al. TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for

data-driven Intrusion Detection Systems [J]. IEEE Access, 2020, 8: 165130-165150.

[9] Moustafa N, Keshk M, Debie E, et al. Federated TON_IoT Windows Datasets for Evaluating AI-Based Security Applications [C]. IEEE TrustCom, 2020: 848-855.

[10] Moustafa N, Ahmed M, Ahmed S. Data Analytics-Enabled Intrusion Detection: Evaluations of ToN_IoT Linux Datasets [C]. IEEE TrustCom, 2020: 727-735.

[11] Moustafa N. New Generations of Internet of Things Datasets for Cybersecurity Applications based Machine Learning: TON_IoT Datasets [C]. eResearch Australasia Conference, 2019.

[12] Moustafa N. A systemic IoT-Fog-Cloud architecture for big-data analytics and cyber security systems: a review of fog computing [J]. arXiv preprint, 2019, arXiv:1906.01055.

[13] Ashraf J, Keshk M, Moustafa N, et al. IoTBoT-IDS: A Novel Statistical Learning-enabled Botnet Detection Framework for Protecting Networks of Smart Cities [J]. Sustainable Cities and Society, 2021: 103041.

[14] Zhang Z, et al. FLDetector: Defending Federated Learning Against Model Poisoning Attacks via Detecting Malicious Clients [C]. ACM SIGKDD, 2022: 2545-2555.

[15] Karimireddy S P, et al. Learning from History for Byzantine Robust Optimization [C]. ICML, 2021, PMLR 139: 5311-5319.

[16] Sharma A, Marchang N. Detection of Malicious Clients in Federated Learning Using Graph Neural Network [J]. IEEE Access, 2025, 13: 1-14.

作者简介：徐温富（2007-），男，汉族，浙江丽水人，喀什大学电子与通信工程学院在读本科生，主要研究方向为通信工程；赵亮（2005-），男，汉族，重庆铜梁人，喀什大学电子与通信工程学院在读本科生，主要研究方向为电子信息科学与技术；张彬蕾（1999-），女，汉族，新疆喀什人，博士，喀什大学电子与通信工程学院助教，主要研究方向为信号处理。