计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(10): 2576−2595, 2026

基于梯度指纹决策的拜占庭鲁棒联邦学习网络入侵检测方法

姜植元[1] 杨 刚[2] 王勇军[1] 秦 拯[3]
[1](国防科技大学计算机学院 长沙 410073)
[2](信息支援部队工程大学网络信息学院 武汉 430030)
[3](湖南大学网络空间安全学院 长沙 410082)
(jzy@nudt.edu.cn)

Byzantine-Robust Federated Learning for Network Intrusion Detection via Gradient Fingerprint-Driven Decision

Jiang Zhiyuan[1], Yang Gang[2], Wang Yongjun[1], and Qin Zheng[3]
[1](*College of Computer Science and Technology*, *National University of Defense Technology*, *Changsha* 410073)
[2](*College of Network and Information*, *Information Support Force Engineering University*, *Wuhan* 430030)
[3](*College of Cyber Science and Technology*, *Hunan University*, *Changsha* 410082)

Abstract In dynamic heterogeneous networks, massive heterogeneous terminals result in dispersed and diverse security data. While federated learning enables collaborative intrusion detection across terminals without compromising data privacy, its open architecture is inherently vulnerable to Byzantine attacks, where malicious clients poison the global model by uploading crafted gradients. Existing Byzantine-robust methods are typically effective only against specific attack types and struggle to cope with the coexistence of multiple attacks in real-world environments. Moreover, they suffer from high false-negative rates under sparse real-traffic conditions, and the computational overhead of stacking multiple defenses is prohibitive for resource-constrained terminals. To address these challenges, we propose GF-FedIDS, a Byzantine-robust federated learning intrusion detection framework based on gradient fingerprint decision-making. By analyzing the gradient statistics of typical Byzantine attacks, we establish a deterministic mapping from attack patterns to optimal defenses and reveal their causal relationship. A gradient fingerprint decision tree is designed to adaptively select a single optimal defense strategy, achieving full-coverage robust aggregation at unit computational cost. Furthermore, a class-aware instance weighting mechanism is introduced to mitigate missed detections under data imbalance. Extensive experiments on three public datasets and eight attack types demonstrate that, in single-attack scenarios, the attack detection rate (*ADR*) of GF-FedIDS outperforms that of the best baseline across three datasets; in mixed-attack scenarios, *ADR* reaches 1.45–2.14 times; and under extreme imbalance with a 5% attack rate, *ADR* is 2.06–2.29 times that of the best baseline. Additionally, GF-FedIDS incurs only a 16% computational overhead compared with FedAvg, outperforming existing ensemble defense methods.

Key words federated learning; Byzantine attack; intrusion detection; gradient fingerprint; Non-IID data

收稿日期: 2026-06-06; **修回日期**: 2026-07-02

基金项目: 国家自然科学基金重点项目(62441233); 国家自然科学基金项目(62402509, 62472166, 62441233); 湖南省自然科学基金项目(2025JJ40053); 国家国防科技工业局重点实验室稳定支持项目(WDZC20265290102)

This work was supported by the Key Program of the National Natural Science Foundation of China (62441233), the National Natural Science Foundation of China (62402509, 62472166, 62441233), the Natural Science Foundation of Hunan Province (2025JJ40053), and the Parallel and Distributed Processing for the Stable Support Project of the National Defense Science and Technology Key Laboratory (WDZC20265290102).

通信作者: 杨刚(yanggang11@nudt.edu.cn)

摘　要　动态异构网络中，海量异质终端导致安全数据分散异构，联邦学习虽能保护数据隐私实现跨终端的协同入侵检测，但其开放架构易受拜占庭攻击，即恶意客户端通过上传恶意梯度破坏全局模型。现有拜占庭鲁棒方法通常仅在特定攻击类型上有效，难以覆盖真实场景中多类型攻击共存的态势，且在攻击样本稀疏的真实流量环境下易产生大量漏报，多防御堆叠的计算开销亦难以适配资源受限终端。对此，提出基于梯度指纹决策的拜占庭鲁棒联邦学习入侵检测框架 GF-FedIDS（gradient fingerprint-driven federated intrusion detection system），通过对典型拜占庭攻击梯度统计特性的分析，建立攻击模式到最优防御的确定性映射，揭示其因果关联；设计梯度指纹决策树，实现自适应选择单一最优防御策略，以单位计算开销达成全覆盖鲁棒聚合；引入类别感知实例加权机制，缓解数据不平衡下的漏报问题。在 3 个公开数据集、8 种攻击类型条件下的广泛实验表明，在单攻击场景下，GF-FedIDS 的攻击检测率（*ADR*）在跨数据集场景下普遍略优于最优基线；在混合攻击场景下，*ADR* 为最优基线的 1.45~2.14 倍；在攻击率为 5% 的极端不平衡条件下，*ADR* 为最优基线的 2.06~2.29 倍。此外，GF-FedIDS 与 FedAvg 相比仅增加 16% 的计算开销，优于现有集成防御方法。

关键词　联邦学习；拜占庭攻击；入侵检测；梯度指纹；非独立同分布数据

中图法分类号　TP393.08

DOI: 10.7544/issn1000-1239.202660484　　**CSTR:** 32373.14.issn1000-1239.202660484

随着物联网、云边端协同计算与下一代通信技术的快速演进，动态异构网络已成为互联网组网的主流形态，海量异质终端的规模化部署成为常态[1]。这使网络入侵检测面临数据分散、隐私敏感、标注稀缺三重困境。传统集中式检测要求将终端流量数据统一汇聚至云端分析，不仅面临高昂的带宽开销与实时性瓶颈，更因原始网络流量可能包含用户通信内容、设备指纹等敏感信息，难以满足日益严苛的数据隐私合规要求。联邦学习作为一种“数据不动模型动”的分布式机器学习范式，为破解上述困境提供了新的技术路径[2]，异质终端在本地利用私有流量数据训练检测模型，仅向中心服务器上传模型参数，在保护数据隐私的前提下实现跨终端入侵检测能力的协同提升[3-5]。

然而，联邦学习的开放协作架构在另一方面引入了新的安全问题。由于模型参数在终端与服务器之间公开传输，攻击者可采用拜占庭攻击，通过攻陷部分终端、上传精心构造的恶意梯度，干扰全局模型的聚合方向，使检测能力严重退化甚至完全失效，这给联邦学习落地于安全敏感场景带来重要挑战。更为严峻的是，动态异构网络中的终端天然呈现海量异质特征——不同终端的硬件能力、网络环境、通信模式和使用场景差异显著，导致各客户端本地流量数据呈典型的非独立同分布（non-independent and identically distributed, Non-IID）。这种数据异质性使得良性客户端之间的本地模型更新本就存在明显偏差，恶意更新可借由这种“正常偏差”隐藏自身，大幅增加了拜占庭攻击的辨识难度[6-7]。与此同时，在现实网络安全防御场景下，攻击样本在真实网络流量中通常占比很低，其极端类别不平衡性使得模型在线联邦学习训练过程被正常样本主导，攻击类别的特征难以被充分学习，导致模型优化目标的天然偏置，产生严重的攻击漏检风险，大大削弱了入侵检测系统的可靠性。

为应对拜占庭攻击，近年来 Krum[8]、坐标中位数（Median）[8]、修剪均值（TrimMean）[9]、范数裁剪（NormClip）[10]、FedProx[11] 等一系列鲁棒聚合方法被提出。这些方法通过剔除或抑制统计异常梯度，在一定程度上保障了联邦学习训练的稳定性。但现有方法在设计之初多面向通用机器学习场景下单类型拜占庭攻击，默认各类样本均衡分布、客户端数据同质，并未充分考虑网络入侵检测任务的领域特性，导致在攻击样本稀疏、数据高度异构的真实流量环境下防御效果大打折扣[12]。具体而言，不同拜占庭攻击对梯度统计特性施加了差异化扰动，不存在普适性的最优固定防御，单一防御方法通常仅在特定攻击类型上表现有效，难以应对混杂海量异构终端的动态异构网络场景中多种攻击共存的复杂态势；而单纯的多防御堆叠虽能扩大覆盖范围，其计算开销却难以适配资源受限的物联网终端。更为本质的是，上述方法均面临一个结构性悖论：鲁棒聚合操作在抑制拜占庭扰动时，不可避免地会削弱本就微弱的攻击类梯度信号，导致防御强度与攻击检测率之间存在此消彼长的困境。

针对上述问题，一种改进思路是根据当前攻击态势动态选择最合适的防御方法。近年来已有研究探索自适应防御方法，如 Multi-Krum[8]、Bulyan[13]，即不再依赖固定防御策略，而是根据观测自适应切换防御。然而，现有自适应方法普遍存在 2 个瓶颈：其一，方法选择多依赖启发式规则或经验阈值，缺乏从攻击模式到最优防御的严格理论映射；其二，未考虑类别不平衡与拜占庭防御的耦合效应，无法避免防御操作对攻击类梯度信号的二次压制。

鉴于上述考虑，本文的核心思路在于利用不同拜占庭攻击对梯度的统计特性施加了差异化、可辨识的确定性扰动，如同在梯度空间中留下独特的“指纹”。基于这一洞察，我们提出基于梯度指纹决策的拜占庭鲁棒联邦学习入侵检测框架 GF-FedIDS(gradient fingerprint-driven federated intrusion detection system)。GF-FedIDS 由 3 个关键组件组成：梯度指纹提取、指纹决策树自适应防御和类别感知实例加权。首先，对典型拜占庭攻击的梯度统计特性进行系统分析，提取方向余弦、范数比、方差比构成三维梯度指纹，建立攻击模式到最优防御方法的确定性映射。随后，利用不同梯度指纹在特征空间中的可分性，构建轻量级决策树，每轮聚合前服务器提取当前梯度指纹，沿决策树路径快速判定攻击类型，自适应选择单一最优防御。该机制以单位计算开销实现多类型攻击的全覆盖鲁棒聚合，且防御选择具有可解释的理论依据。为进一步缓解极端不平衡下的漏报问题，引入类别感知实例加权机制，通过赋予攻击样本更高的优化权重，从源头补偿其在损失函数中被淹没的梯度贡献。

本文的贡献主要包括 4 个方面：

1)提出了基于梯度指纹的防御策略路由理论。该理论通过对典型拜占庭攻击梯度统计特性的形式化分析，提取方向余弦、范数比、方差比构成三维梯度指纹，建立了攻击模式到最优防御方法之间的因果映射关系，为自适应防御选择提供了数学基础；

2)提出了类别感知实例加权机制。针对实践网络流量中攻击样本极度稀疏的领域特性，在本地训练阶段为攻击样本赋予更高的优化权重，在训练损失函数中重新平衡正负样本的优化比重，有效解决极端不平衡场景下的漏报问题；

3)提出了基于梯度指纹决策的自适应防御方法 GF-FedIDS。该框架构建轻量级指纹决策树，每轮基于提取的梯度指纹自适应选择单一最优防御策略，以单位计算开销实现多类型拜占庭攻击的全覆盖鲁棒聚合，实现“攻击识别—类别感知—自适应防御”三大机制的协同集成，突破了现有方法仅针对特定攻击有效的局限；

4)在 3 种公开数据集上围绕 8 种攻击类型条件下进行了 200 余次独立实验。结果表明，在单攻击场景下，GF-FedIDS 的攻击检测率(ADR)普遍略优于最优基线；在混合攻击场景下，ADR 为最优基线的 1.45~2.14 倍；在攻击率为 5% 的极端不平衡条件下，ADR 为最优基线的 2.06~2.29 倍。此外，GF-FedIDS 与 FedAvg 相比仅增加 16% 的计算开销，优于现有集成防御方法。

1 相关工作

1.1 联邦学习驱动的网络入侵检测方法

联邦学习由 Brendan 等人[2]于 2017 年系统提出，其核心思想是通过“数据不动模型动”的协同训练模式，使分布式客户端在不共享本地原始数据的前提下共同训练全局模型。该范式避免了原始数据的集中化存储与传输，在保护数据隐私的同时为海量异质终端的协同学习提供了可行的技术路径[14-15]。

近年来，联邦学习在物联网入侵检测领域取得了显著进展。Zhang 等人[16]系统综述了联邦深度学习在物联网安全中的应用，指出该领域面临的通信瓶颈、数据异构与安全攻击等核心挑战。Alhasawi 等人[17]利用联邦学习实现了跨多个 IoT 网络的协同 DDoS 攻击检测；Khan 等人[18]提出了一种基于规则的互通架构 Hetero-FedIoT，通过中间件层实现异构物联网设备间跨联邦学习框架的协同训练，解决了不同联邦学习协议和网络配置下设备的互操作问题。上述工作表明，联邦学习为异构终端构成的动态网络场景下的恶意流量检测提供了有效的技术基础[19-20]。

然而，上述方法均建立在 FedAvg 聚合算法[2]之上，该算法仅在客户端数据满足独立同分布(IID)假设时方能取得接近集中式训练的收敛效果[2]。在实际动态异构网络中，各客户端本地数据呈显著的 Non-IID 特性，FedAvg 在此条件下收敛速度下降、模型性能退化已成为共识[21]。更为严峻的是，联邦学习的开放架构使其天然暴露于拜占庭攻击威胁之下。当 Non-IID 数据分布与拜占庭攻击相互叠加时，2 种效应形成风险进一步耦合：良性客户端之间的本地更新本身存在的偏差，会增加攻击辨识难度；而拜占庭攻击则进一步放大 Non-IID 引入的偏差，使得全局模型偏离最优方向。这使得标准联邦学习框架在海量

异质终端的入侵检测任务中面临严重的安全与性能危机，如何在此场景下实现拜占庭鲁棒的联邦学习入侵检测，成为亟待解决的关键问题。

1.2 拜占庭攻击方法

拜占庭攻击指在联邦学习分布式训练过程中，恶意客户端通过上传精心构造的恶意更新破坏全局模型收敛的行为[22]。现有拜占庭攻击可分为数据投毒与模型投毒两大类：前者通过篡改本地训练数据间接影响梯度更新，典型方法包括符合翻转攻击和后门攻击；后者直接构造恶意梯度上传，具有更强的攻击效果[23]。Chen 等人[24]系统研究了联邦学习中的拜占庭攻击，提出了 SignFlip 符号翻转和随机噪声等基础攻击。后续研究发展了更为复杂的变体：Baruch 等人[25]提出的 LIE 攻击通过统计模仿使恶意梯度难以区分；Shejwalkar 等人[26]提出的 Min-Max 和 Min-Sum 攻击通过约束优化最大化攻击影响；高鸿峰等人[27]与 Fang 等人[28]分别系统研究了本地模型投毒攻击对拜占庭鲁棒联邦学习的威胁。Allouah 等人[29]从异构性角度分析了拜占庭机器学习的理论极限，提出了 Fixing by mixing 策略。

1.3 现有拜占庭防御方法

针对拜占庭攻击，研究者提出了多种防御方法，可分为以下 3 类：

1）基于稳健统计的方法。Median[8]使用各参数维度坐标中位数替代加权平均，TrimMean[9]在每个坐标维度上去除极值后取平均。此类方法实现简单，但对自适应攻击和定向攻击的防御效果有限。

2）基于正则化的方法。FedProx[11]在本地训练目标函数中添加近端项，限制本地更新与全局模型的偏离程度，从而抑制恶意更新的影响。但该方法引入额外的超参数，且对强攻击的防御效果不够理想。

3）基于梯度过滤的方法。NormClip[10]根据梯度范数裁剪异常更新，Fed-NGA[30]基于梯度对齐性进行过滤。此类方法对范数异常攻击有效，但对统计模仿类攻击的检测能力不足。

上述防御方法的共同局限在于：均采用单一固定策略，无法覆盖所有攻击类型。不同防御方法在不同攻击场景下表现差异显著：Median 和 TrimMean 对符号翻转攻击具有天然的鲁棒性，但在自适应攻击[31]等定向攻击下近乎失效；Krum 对 LIE 和 Min-Max 等统计模仿类攻击效果良好，却难以抵御自适应攻击；基于历史梯度一致性的 Fed-NGA 虽对多数攻击表现稳定，在定向攻击下仍部分有效；FedProx 由于仅通过近端约束间接限制恶意更新幅度，在各类攻击下均仅能提供有限防护。在此基础上，研究者探索了 2 种扩展思路：其一，Blanchard 等人[8]在提出 Krum 时同时提出了 Multi-Krum，即选择多个评分最高的梯度进行聚合，但本质上是同一防御方法的多次应用，并未实现不同防御策略的优势互补；其二，Guerraoui 等人[13]提出的 Bulyan 结合了 Krum 与 TrimMean，先以 Krum 筛选候选梯度，再以 TrimMean 聚合，属于 2 种防御的串联集成。然而，Bulyan 虽扩大了防御覆盖范围，其串行执行方式仍无法动态适配攻击类型的变化，且 2 种防御的计算开销叠加，限制了其在资源受限场景下的部署效率。

2 预备知识

2.1 联邦学习网络入侵检测问题

形式化地，考虑由 N 个客户端和 1 个中央服务器组成的横向联邦学习系统。设第 i 个客户端持有本地私有数据集 $\mathcal{D}_i=\{(\boldsymbol{x}_j,y_j)\}_{j=1}^{n_i}$，其中 $\boldsymbol{x}_j\in\mathcal{X}$ 为输入特征（网络流量的统计特征向量），$y_j\in\mathcal{Y}=\{0,1\}$ 为标签（0 表示正常流量，1 表示攻击流量），$n_i=|\mathcal{D}_i|$ 为客户端 i 的样本量。全局目标函数定义为所有客户端本地经验风险的最小化：

$$\min_{\boldsymbol{w}\in\mathbb{R}^d}F(\boldsymbol{w})=\sum_{i=1}^{N}\frac{n_i}{n}F_i(\boldsymbol{w})\ , \tag{1}$$

其中 $F_i(\boldsymbol{w})=\dfrac{1}{n_i}\sum\limits_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_i}\ell(\boldsymbol{w};\boldsymbol{x},\boldsymbol{y})$ 为第 i 个客户端的本地经验风险，$n=\sum\limits_{i=1}^{N}n_i$ 为总样本量，$\ell(\cdot)$ 为损失函数（本文采用二元交叉熵损失），$\boldsymbol{w}\in\mathbb{R}^d$ 为模型参数向量。联邦学习的训练过程通过多轮迭代进行，在第 t 轮通信中：首先，服务器将当前全局模型 $\boldsymbol{w}^t$ 广播给选定的客户端集合 $\mathcal{S}^t\subseteq\{h_1,h_2\cdots,h_N\}$；接着，每个客户端 $h_i\in\mathcal{S}^t$ 在本地执行 E 轮小批量随机梯度下降（SGD），计算本地模型更新 $\boldsymbol{w}_i^{t+1}=\boldsymbol{w}^t-\eta\nabla F_i(\boldsymbol{w}^t)$，并将更新差值（伪梯度）$\boldsymbol{g}_i^t=\boldsymbol{w}^t-\boldsymbol{w}_i^{t+1}$ 上传至服务器；最终，服务器执行聚合操作 $\boldsymbol{w}^{t+1}=\boldsymbol{w}^t-\eta_{\text{global}}\cdot\mathcal{A}gg(\{\boldsymbol{g}_i^t\}_{i\in\mathcal{S}^t})$，其中 $\mathcal{A}gg(\cdot)$ 为聚合函数。标准联邦学习（FedAvg[2]）采用加权平均聚合，$\mathcal{A}gg(\cdot)$ 取加权平均，即：

$$\boldsymbol{w}^{t+1}=\sum_{i\in\mathcal{S}^t}\frac{n_i}{\sum\limits_{j}n_j}\boldsymbol{w}_i^{t+1}。 \tag{2}$$

海量异质终端场景的网络入侵检测具有区别于一般联邦学习任务的特性，这决定了其拜占庭防御策略的设计方向：1）高维稀疏性。网络流量特征通

常维度达数十至上百维, 且大量特征值为 0。以 CICIDS-2017 数据集为例, 78 维特征中平均约 40% 的取值为 0, 这种稀疏结构使得梯度在参数空间中呈现高度不对称分布。2)极度不平衡性。正常流量通常占绝大部分, 攻击流量在数量和质量上均有显著缺陷, 导致传统联邦学习的优化目标易被正常样本主导, 攻击样本的梯度贡献被大幅削弱。3)分布异质性。不同终端的设备类型、网络协议栈及使用模式差异显著, 各本地数据分布之间差异巨大。此特性正是 Non-IID 与拜占庭攻击危险耦合的根源所在: 良性客户端之间的本地更新本就存在系统性偏差, 恶意更新可借此“正常偏移”隐蔽自身。4)漏报敏感性。安全检测场景漏检攻击的代价远高于误报正常流量, 要求模型保持高召回率。

上述域特性引发 2 个核心矛盾: 其一, Non-IID 分布使各客户端的本地最优解存在系统性偏移, 拜占庭攻击可借此掩盖恶意行为; 其二, 类别极度不平衡使攻击样本在全局聚合中自然被淹没, 模型可能因优化目标偏差而漏检攻击。

在海量异质终端场景下, 终端设备数量庞大、类型多样、安全等级参差不齐, 恶意客户端的渗透风险显著增加。因此, 本文聚焦的核心问题为: 如何设计既能抵抗拜占庭攻击, 又能适应极度不平衡数据分布的恶意流量检测机制。

2.2 拜占庭攻击模型

本文采用经典的拜占庭攻击模型[4]。假设N个客户端中, 至多m个客户端被攻击者攻陷成为拜占庭节点, 其中$m<N/2$以保证诚实客户端占多数。恶意客户端集合记为$\mathcal{M}$, $|\mathcal{M}|\leqslant m$; 诚实客户端集合记为$\mathcal{H}$, $|\mathcal{H}|=N-m$。拜占庭攻击的数学形式化定义如下: 对于每个恶意客户端$h_i\in\mathcal{M}$, 其上传至服务器的梯度$\tilde{\boldsymbol{g}}_i^t$不一定等于真实的本地梯度$\boldsymbol{g}_i^t=\nabla F_i(\boldsymbol{w}^t)$。攻击者根据攻击策略$\mathcal{A}$构造恶意梯度。本文主要关注以下 8 类攻击。

1)符号翻转攻击(SignFlip[24], SF): $g_i=-c\cdot\nabla F_i(\theta)$, 其中$c>0$为放大系数, 攻击将本地梯度乘以负常数以实现方向完全翻转。

2)部分翻转攻击(Partial, PT): 仅对部分参数坐标进行符号翻转, 使恶意梯度集合内部方向混杂。

3)高斯噪声攻击(Gaussian[24], GS): $g_i=\nabla F_i(\theta)+\epsilon$, 其中$\epsilon\sim\mathcal{N}(0,\sigma^2I)$为加性高斯噪声。

4)范数放大攻击(Boosting, BT): $g_i=c\cdot\nabla F_i(\theta)$, 其中 c 为极大放大系数, 使恶意梯度主导聚合方向。

5)定向攻击(TrimAware): $g_i=c\cdot 1$, 刻意制造极端梯度值以拉高均值, 针对 TrimMean 优化设计而提出。

6)统计模仿攻击(LIE[25]): $\tilde{\boldsymbol{g}}_i^t=\mu^t+z\cdot\sigma^t$, 其中$\mu^t$和$\sigma^t$分别为所有诚实客户端梯度的均值和标准差, $z>0$为微调系数。攻击将每个坐标设置为均值向正方向偏移z个标准差, 使恶意梯度在统计上接近诚实梯度。

7)约束优化攻击(Min-Max[26], MM): 通过约束优化最大化攻击效用的同时逃避检测, $\tilde{\boldsymbol{g}}_i^t=\arg\max_{\boldsymbol{g}}\|\boldsymbol{w}^{t+1}(\boldsymbol{g})-\boldsymbol{w}_{\text{honest}}^{t+1}\|$, 约束条件为恶意梯度落在诚实梯度凸包内, 即$\boldsymbol{g}\in Conv\left(\{\boldsymbol{g}_j^t\}_{j\in\mathcal{H}}\right)$。

8)自适应攻击(Adaptive, AD): 本文的自适应攻击策略为根据历史全局模型动态调整扰动策略, 以 KL 散度度量跨轮次梯度行为偏差。该攻击为逃避基于单轮梯度统计特性的检测, 刻意使恶意梯度在方向上与良性梯度保持近似一致, 但其跨轮次扰动策略的频繁切换导致恶意梯度内部的离散程度显著高于良性梯度, 此特征为梯度指纹辨识提供了可观测依据。

上述攻击对梯度统计特性的扰动模式各异: SF 改变梯度方向, BT 放大梯度范数, LIE 和 MM 操纵梯度分布的偏度和峰度。这些特性决定了其防御方法, 也成为其攻击类别的识别手段。

2.3 拜占庭防御方法

本文以 8 种代表性防御聚合方法作为基线。

1)FedAvg[2]。标准加权平均聚合, $\boldsymbol{w}^{t+1}=\sum_{i\in\mathcal{S}^t}\frac{n_i}{\sum_j n_j}\cdot\boldsymbol{w}_i^{t+1}$。无拜占庭鲁棒性, 作为无防御参照基线, 代表无防护条件下的性能下界。

2)Krum[8]。在S个候选更新中, 选择与其他更新具有最小平方欧氏距离和的梯度。形式化地, 对每个梯度g_i, 定义得分$s(i)=\sum_{i'\in\mathcal{S}(i;S-m-2)}\|g_i-g_{i'}\|^2$, 其中$\mathcal{S}(i;k)$为与$g_i$距离最近的$k$个梯度的索引集合。Krum 选择$g_{i^*}$使得$i^*=\arg\min_i s(i)$, 全局更新为$w^{t+1}=w^t-\eta\cdot g_{i^*}$。

3)Median[8]。在每个坐标维度$j\in\{1,\cdots,d\}$上独立计算所有候选梯度的中位数: $w_j^{t+1}=median\left(\{w_{i,j}^{t+1}\}_{i\in\mathcal{S}^t}\right)$。该方法对极端异常值具有天然的鲁棒性。

4)TrimMean[9]。在每个坐标维度上, 剔除该坐标上取值最大和最小的$\lfloor c\cdot S\rfloor$个值后计算均值, 其中$c\in(0,1/2)$为修剪比例。

5)NormClip[10]。首先计算所有候选梯度的L_2范数中位数$m=median(\{\|g_i\|_2\})$, 然后对每个梯度执行范

数裁剪：$\hat{g}_i = g_i \cdot \min(1, \tau \cdot m / \|g_i\|_2)$，其中$\tau > 0$为裁剪阈值系数，最后取裁剪后梯度的均值作为全局更新。

6）FedProx[11]。在本地目标函数中添加近端正则项：$F_i^{\text{prox}}(w) = F_i(w) + \frac{\mu}{2}\|w - w^t\|^2$，通过限制本地更新与全局模型的偏离程度来降低恶意更新的影响。

7）Multi-Krum[8]。Krum 的扩展版本，在每轮中选择多个可信梯度进行聚合，而非仅选择单个最优梯度，在提高鲁棒性的同时降低方差。

8）Bulyan[13]。首先通过 Multi-Krum 筛选$N - 2m$个可靠梯度，再执行坐标级 TrimMean 裁剪，理论上兼具选择鲁棒性与裁剪稳定性。

上述防御方法的鲁棒性边界各不相同：Krum 在$m < S/2$时具有理论收敛保证；Median 和 TrimMean 假设正常梯度在坐标级上集中分布；NormClip 假设恶意梯度的范数显著大于正常梯度；FedProx 通过近端约束间接限制恶意更新的幅度。本文提出的 GF-FedIDS 框架利用上述典型攻击的梯度特征扰动模式，建立攻击类型到最优防御方法的自适应映射。

3 解决方案

3.1 核心思路

考虑到不同拜占庭攻击对梯度特征施加的扰动模式具有显著差异：SF 攻击反转梯度方向，BT 攻击放大梯度范数，LIE 攻击模仿良性统计特性，而自适应攻击则动态调整扰动策略。同时，不同防御方法对特定扰动模式的抵抗能力亦各不相同：中位数对方向反转免疫，范数裁剪对异常范数有效，Krum 对统计孤立点敏感。本文提出 GF-FedIDS 框架，不再依赖通用防御对梯度的无差别过滤，而是通过精确识别攻击类型，定向选择只针对当前攻击有效的防御方法，从而在消除拜占庭扰动的同时最大程度保留攻击类别的梯度信息。为此，本文通过解析恶意梯度与良性梯度在方向、范数和分布方差三维特征上的数学关系构建“梯度指纹”，以作出最优拜占庭攻击防御选择，并结合类别感知加权机制，以感知处理动态网络中海量异质终端入侵检测数据的不平衡性，提升整体性能。整体框架如图 1 所示。

3.2 梯度指纹

本节提出梯度指纹，并从理论上刻画每种攻击对梯度特征的扰动模式，并推导对应的防御最优性。

定义 1. 梯度指纹。设第t轮通信中，良性客户端集合为$\mathcal{B}$，恶意客户端集合为$\mathcal{M}$。定义良性梯度均值$\boldsymbol{g}^{\text{ben}} = 1/|\mathcal{B}| \sum_{i \in \mathcal{B}} \boldsymbol{g}_i$与恶意梯度均值$\boldsymbol{g}^{\text{mal}} = 1/|\mathcal{M}| \sum_{i \in \mathcal{M}} \boldsymbol{g}_i$。梯度指纹作为一个三维空间内的点可以被定义为：$f = (CS, NR, VR)$。

其中，余弦相似度$CS = \boldsymbol{g}^{\text{mal}} \cdot \boldsymbol{g}^{\text{ben}} / \|\boldsymbol{g}^{\text{mal}}\| \cdot \|\boldsymbol{g}^{\text{ben}}\|$，度量恶意梯度与良性梯度的方向一致性；范数比$NR = \|\boldsymbol{g}^{\text{mal}}\| / \|\boldsymbol{g}^{\text{ben}}\|$，度量恶意梯度的相对强度；方差比$VR = Var(\{\boldsymbol{g}_i : i \in \mathcal{M}\}) / Var(\{\boldsymbol{g}_i : i \in \mathcal{B}\})$，度量恶意梯度内部离散程度相对于良性梯度的比值。

基于上述梯度指纹特征，本文建立如表 1 所示的攻击类型与最优防御的映射。

相对各类攻击类型的最优防御理论论证如下。

1）SF 攻击 →Median。SF 攻击将本地梯度乘以负常数$-c$（$c > 0$），即$g' = -c \cdot g$。对于中位数聚合，考

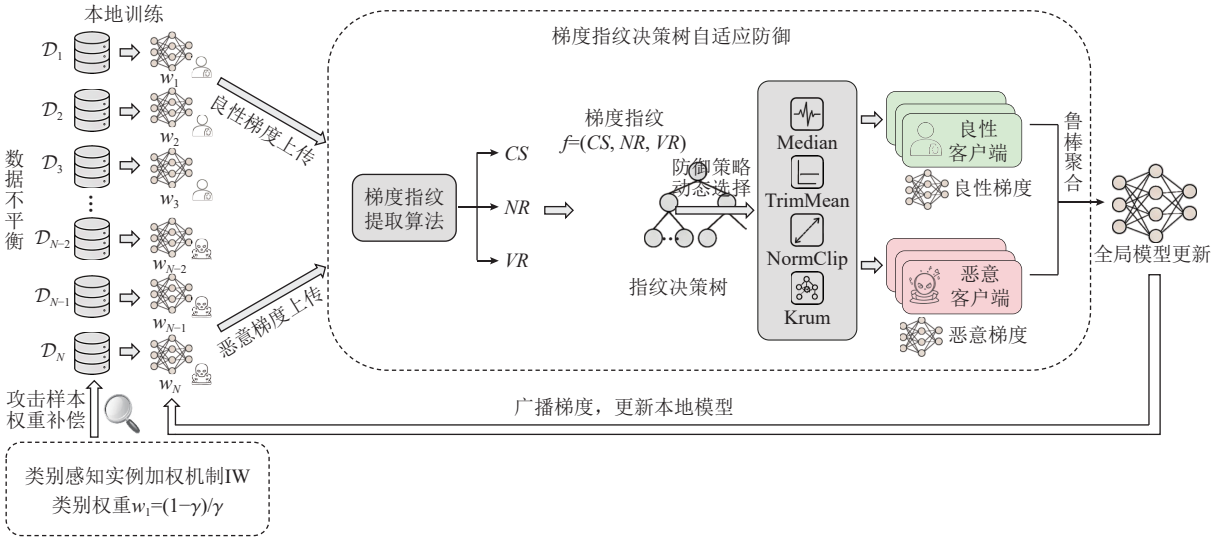


Fig. 1 Architecture diagram of GF-FedIDS method

图 1 GF-FedIDS 方法架构图

Table 1 Mapping of Attack Type to Optimal Defense

表 1 攻击类型与最优防御的映射

攻击类型	*CS*	*NR*	*VR*	最优防御
SF	<0	>1	>1	Median
PT	≈ 0	≈ 1	>1	Median
GS	≈ 0	>1	$\gg 1$	NormClip
BT	>0.5	$\gg 1$	<1	TrimMean
TrimAware	>0.5	$\gg 1$	<1	TrimMean
LIE	>0.5	≈ 1	<1	Krum
MM	>0.5	≈ 1	<1	Krum
AD	≈ 0	>1	$\gg 1$	NormClip+TrimMean

虑各维j的梯度值集合$\{g_{1,j}, g_{2,j}, \cdots, g_{n,j}\}$(含$m$个恶意值$\{-c\cdot g_{i,j}\}$)。当$c>1$时,恶意值$\{-c\cdot g_{i,j}\}$位于数轴上与良性值相反方向且距离更远的位置。设良性梯度的中位数为med_{ben},由于恶意值偏离中心区域,它们不会翻转中位数的符号。形式化地,对于足够大的良性集合,$med\left(\{-c\cdot g^{\text{mal}}\}\cup\{g^{\text{ben}}\}\right)=med\left(\{g^{\text{ben}}\}\right)$,即中位数对反转操作免疫。

2)PT攻击→Median。PT攻击仅对部分坐标进行符号翻转,导致恶意梯度集合内部方向混杂。设翻转比例为$p\in(0,1)$,则恶意梯度可表示为$g'=(1-2p)\cdot g+\delta$,其中δ为翻转噪声。此时g'的方向介于g和$-g$之间,$CS\in[-0.3,0.1]$。中位数在各维度独立操作,不受方向混杂的影响,其崩溃点(breakdown point)为$\epsilon^*=0.5$,在恶意比例低于50%时保持稳健。

3)GS攻击→NormClip。GS攻击向梯度注入零均值高斯噪声:$g'=g+\xi$,$\xi\sim\mathcal{N}(0,\sigma^2 I)$。其期望范数为$\mathbb{E}\left[\|g'\|^2\right]=\|g\|^2+d\sigma^2$,其中$d$为参数维度。当$d$较大时(网络流量特征的高维性),$\|g'\|$以高概率显著偏离良性范数。范数裁剪操作$\Pi_\tau(g')=g'\cdot\min(1,\tau/\|g'\|)$严格限制$\|\Pi_\tau(g')\|\leqslant\tau$,将噪声引起的范数膨胀压制在阈值以内。

4)BT攻击→TrimMean。BT攻击将良性梯度放大c倍:$g'=c\cdot g(c\gg 1)$,使得恶意梯度在数值上成为极端大值。β-截断均值首先剔除排序后最大的βn个和最小的βn个梯度,再对剩余梯度取平均。当恶意客户端比例小于等于β时,所有放大梯度被剔除,聚合结果等于良性梯度的均值,攻击效果归零。

5)TrimAware攻击→TrimMean。TrimAware攻击刻意制造极端梯度值以拉高均值,其设计针对传统Mean聚合。由于攻击者产生的是定向极端值(全部偏大或偏小),恰好落入截断均值的剔除区域。设攻击产生m个值为$g_{\max}$的梯度,当$m\leqslant\beta n$时,β-TrimMean完全过滤这些极端值。

6)LIE攻击→Krum。LIE攻击使恶意梯度模仿良性梯度的统计特性,即$g'=\mu_{\text{ben}}+\sigma_{\text{ben}}\cdot\xi$。虽然单点统计特性相似,但LIE攻击者无法同时与所有良性梯度保持高相似性。Krum算法[6]计算每个梯度与其余$n-f-2$个最近邻距离之和,选择最小者。设良性梯度集合$\{g_i\}_{i\in\mathcal{B}}$的内部距离为$d_{in}$,恶意梯度$g'$到各良性梯度的距离为$d_i=\|g'-g_i\|$。由于$g'$为单一统计模仿点,$\sum_{i\in S}d_i(|S|=n-f-2)$必然大于某良性梯度与其邻居的距离之和,故Krum可选择良性梯度。

7)MM攻击→Krum。MM攻击通过约束优化最大化攻击效用的同时逃避检测。设约束为$\|g'-g^{\text{ben}}\|\leqslant\delta$,则$g'$位于以$g^{\text{ben}}$为中心、$\delta$为半径的球内。此类优化后的梯度在高维空间与其他良性梯度保持可控距离$\Theta(\delta)$,而良性梯度之间距离为$\Theta(\sigma_{\text{intra}})$。当$\delta$与$\sigma_{\text{intra}}$量级相当时,MM梯度在Krum评分中表现为孤立点,因其无法同时与足够多的良性梯度形成紧密邻域。

8)AD攻击→组合防御。自适应攻击根据历史全局模型动态调整扰动策略,其梯度特征随轮次变化。本文通过梯度指纹的动态变化模式($VR\gg 1$且CS波动)识别自适应攻击,采用NormClip限制范数波动叠加TrimMean剔除极端值的组合策略。

定理 1. 梯度指纹可分性。对于表1所列8种标准拜占庭攻击,设其三维梯度特征向量分别为$\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_8\in\mathbb{R}^3$。在良性梯度满足亚高斯分布$g^{\text{ben}}\sim subG(\sigma^2)$且恶意客户端比例$\beta<0.5$的条件下,存在线性超平面$H_{ij}:a_{ij}^{\mathrm{T}}\boldsymbol{f}+b_{ij}=0$使得对任意$i\neq j$,$\boldsymbol{f}_i$和$\boldsymbol{f}_j$分别位于$H_{ij}$两侧,即梯度指纹在三维特征空间中线性可分。

证明. 由表1可知,8种攻击的CS覆盖范围互不重叠:SF/PT为负值区间$[-1,0.1]$,GS近似为0$[-0.1,0.1]$,LIE/MM为中等正值($[0.2,0.6]$),BT/TrimAware为较高正值($[0.3,0.9]$)(其中TrimAware的NR显著高于BT),AD在0附近但VR显著高于其他攻击。NR维度进一步区分GS(高NR)、LIE(低NR)和MM(低NR)。进一步结合VR维度,各类攻击的特征向量位于$\mathbb{R}^3$的不同区域,由超平面分离定理即得线性可分性,详细证明见附录A。 证毕.

定理 2. 防御最优性。设攻击类型为$\mathcal{A}$,对应的梯度指纹为$f_{\mathcal{A}}$,通过表1映射选择的防御为$d^*=\phi(f_{\mathcal{A}})$。定义聚合后梯度偏差为$\Delta(\boldsymbol{w},\boldsymbol{D})=\|g^{\text{agg}}(\boldsymbol{D})-g^{\text{ben}}\|$,则$\boldsymbol{D}^*$满足:$\Delta(\boldsymbol{w},\boldsymbol{D}^*)\leqslant\Delta(\boldsymbol{w},\boldsymbol{D})$,$\forall\boldsymbol{D}\in\mathcal{D}=\{$Median,

TrimMean, Krum, NormClip}。

证明. 对每种攻击类型, 所选防御 $\boldsymbol{D}^*$ 恰好消除了该攻击的扰动模式: Median 免疫方向反转, NormClip 限制噪声范数膨胀, TrimMean 剔除极端值, Krum 利用统计模仿的孤立性。其他防御对该攻击的特定扰动模式不具备同等消除能力, 故偏差更大。详细证明见附录 B。 证毕.

3.3 攻击-防御决策树算法

基于上述理论分析, 本节给出 GF-FedIDS 框架中基于梯度指纹实现的决策算法描述, 包含 3 个核心算法模块。

算法 1. 梯度指纹提取算法。

输入: 第 t 轮所有本地客户端梯度集合 $\{g_i^{(t)}\}_{i=1}^{n}$, 恶意客户端索引集合 $\mathcal{M}$;

输出: 各客户端梯度指纹 $\boldsymbol{f}=(CS, NR, VR)$。

① for $g_i^{(t)} \in \{g_i^{(t)}\}_{i=1}^{n}$ do

② $g^{\text{ben}} \leftarrow \dfrac{1}{n-|\mathcal{M}|}\sum_{i\notin\mathcal{M}} g_i^{(t)}$; /*计算良性梯度均值*/

③ $g^{\text{mal}} \leftarrow \dfrac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} g_i^{(t)}$; /*计算恶意梯度均值*/

④ $CS = \dfrac{g^{\text{mal}} \cdot g^{\text{ben}}}{\|g^{\text{mal}}\| \cdot \|g^{\text{ben}}\|}$; /*计算余弦相似度*/

⑤ $NR = \dfrac{\|g^{\text{mal}}\|}{\|g^{\text{ben}}\|}$; /*计算范数比*/

⑥ $VR = \dfrac{\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \|g_i - g^{\text{mal}}\|^2}{\frac{1}{n-|\mathcal{M}|}\sum_{i\notin\mathcal{M}} \|g_i - g^{\text{ben}}\|^2}$; /*计算方差比*/

⑦ end for

⑧ return $f=(CS, NR, VR)$。

算法 2. 自适应防御选择算法。

输入: 梯度指纹 $\boldsymbol{f}=(CS, NR, VR)$;

输出: 最优防御方法 $\boldsymbol{D}^*$。

① 初始化 $\boldsymbol{D}^* \leftarrow \varnothing$;

② if $CS < \tau_{CS}$ then

③ $\boldsymbol{D}^* \leftarrow$ Median;

④ end if

⑤ if $NR > \tau_{NR}$ then

⑥ $\boldsymbol{D}^* \leftarrow$ TrimMean;

⑦ end if

⑧ if $VR > \tau_{VR}$ then

⑨ if $CS \approx 0$ then

⑩ $\boldsymbol{D}^* \leftarrow$ {NormClip,TrimMean};

⑪ else

⑫ $\boldsymbol{D}^* \leftarrow$NormClip;

⑬ end if

⑭ else $VR < \tau_{VR}$ then

⑮ if $CS > \tau_{CS}$ then

⑯ $\boldsymbol{D}^* \leftarrow$ Krum;

⑰ else

⑱ $\boldsymbol{D}^* \leftarrow$Median;

⑲ end if

⑳ end if

㉑ return $\boldsymbol{D}^*$。

其中 $\boldsymbol{D}^* \in$ {NormClip,TrimMean} 表示执行 NormClip 进行范数裁剪以限制噪声幅度, 再对裁剪后的梯度执行 TrimMean 聚合以剔除残留的极端值。该组合防御的具体执行方式为: 给定当前轮所有梯度 $\{g_i^{(k)}\}$, 首先按式 $\tilde{g}_i^{(k)} = g_i^{(k)} \cdot \min(1, \tau/\|g_i^{(k)}\|)$ 进行范数裁剪, 再对裁剪后的梯度集合执行坐标级 TrimMean 聚合。上述决策树仅需 3 次浮点数比较即可完成防御选择, 时间复杂度为 $O(1)$, 计算开销可忽略不计。

梯度指纹决策树的阈值根据对基于攻击指纹的理论分布特性的离线分析确定, 在本文中, CS 阈值 τ_{CS} 取 −0.3, NR 阈值 τ_{NR} 取 3.0, VR 阈值 τ_{VR} 取 1.5, 简要的理论分析如下:

1) CS 阈值 τ_{CS}。SF 攻击将良性梯度符号反转, 其理论 $CS \approx -1$。设定方向反转的判定边界为 $\tau_{CS} = -0.3$, 可有效区分方向性攻击与其他类型攻击, 同时为高斯噪声等因素导致的轻微方向偏移保留适当容差。

2) NR 阈值 τ_{NR}。BT 攻击将梯度放大 λ 倍(典型值 $\lambda \in [5, 20]$), 产生 $NR \gg 1$, τ_{NR} 设定为 3.0 可识别范数显著偏离正常范围的攻击行为, 同时避免将正常训练产生的较大梯度更新误判为攻击。

3) VR 阈值 τ_{VR}。TrimAware 攻击通过放大梯度方差产生 $VR \in [1.5, 3]$。方差异常的判定边界值 τ_{VR} 设定为 1.5, 可有效捕获此类精细化的方差操纵攻击。

上述阈值在实际部署中可根据本文思路酌情针对特定数据集或攻击类型特性进行调整, 本文在设定上述阈值时实际执行的敏感性分析见附录 C。

算法 3. GF-FedIDS 完整训练流程。

输入: 本地训练数据集 $(\boldsymbol{x}_j, \boldsymbol{y}_j)$, 攻击样本比例 α, 恶意客户端比例 β, 学习率 η, 通信轮数 T;

输出: 全局模型 $\boldsymbol{w}_T$。

① 初始化全局模型 $\boldsymbol{w}_0$, 计算类别权重 $\boldsymbol{w}_{\text{attack}} \leftarrow (1-\alpha)/\alpha$;

② for $t \leftarrow 1$ to T do

③ 广播 $\boldsymbol{w}^{t-1}$ 至所有客户端; /*本地训练*/

/*本地训练阶段*/

④ for each $i \leftarrow 1$ to n 并行 do

⑤ $\boldsymbol{w}_i^{(t)} \leftarrow \boldsymbol{w}^{t-1} - \eta_{\text{local}} \sum_j \boldsymbol{w}_j \nabla l\left(\boldsymbol{f}(\boldsymbol{x}_j;\boldsymbol{w}),\boldsymbol{y}_j\right)$;

⑥ $\boldsymbol{g}_i^{(t)} \leftarrow \boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{(t)}$; /*计算本地梯度更新*/

⑦ 上传$\boldsymbol{g}_i^{(t)}$至服务器;

⑧ end for

⑨ end for

/*梯度指纹提取与策略选择阶段*/

⑩ if $\mathcal{M}^{(t)} \neq \varnothing$ then

⑪ 提取梯度指纹$\boldsymbol{f}^{(t)}$(算法 1);

⑫ 选择最优防御$\boldsymbol{D}^{*(t)}$(算法 2);

⑬ else $\boldsymbol{D}^{*(t)} \leftarrow FedAvg$;

⑭ $\boldsymbol{g}_{\text{agg}}^{(t)} \leftarrow \boldsymbol{D}^{*(t)}\left(\{\boldsymbol{g}_i^{(t)}\}_{i=1}^{\boldsymbol{n}}\right)$;

⑮ $\boldsymbol{w}_t \leftarrow \boldsymbol{w}^{t-1} + \eta \cdot \boldsymbol{g}_{\text{agg}}^{(t)}$; /*更新全局模型*/

⑯ end if

⑰ return $\boldsymbol{w}_t$。

3.4 类别感知加权机制

动态网络中海量异质终端入侵检测数据的不平衡性[32](攻击样本比例$\gamma \to 0$)导致传统联邦学习优化目标严重偏向正常类。本节从理论上分析此问题,并提出类别感知实例加权(instance weighting, IW)机制。

设全局数据集中正常样本比例为$1-\gamma$,攻击样本比例为γ,$\gamma \to 0$。传统联邦学习的优化目标为:

$$\min_{\theta} F(\theta) = (1-\gamma)\cdot \mathbb{E}_{(x,y)\sim \mathcal{D}_0}\left[\ell(f(x),y)\right] + \gamma \cdot \mathbb{E}_{(x,y)\sim \mathcal{D}_1}\left[\ell(f(x),y)\right] 。 \tag{3}$$

对θ求梯度:

$$\nabla F = (1-\gamma)\cdot \nabla F_0 + \gamma \cdot \nabla F_1 , \tag{4}$$

当$\gamma \to 0$时,$\nabla F \approx \nabla F_0$,攻击样本的梯度贡献趋于 0,模型几乎不学习攻击类的判别特征,其中$\ell()$为损失函数。

引入类别权重$w_1 = (1-\gamma)/\gamma$,构造新优化目标:

$$\min_{\theta} F_{\text{IW}}(\theta) = (1-\gamma)\cdot \mathbb{E}_{\mathcal{D}_0}[\ell] + w_1 \cdot \gamma \cdot \mathbb{E}_{\mathcal{D}_1}[\ell] = (1-\gamma)\cdot\left(\mathbb{E}_{\mathcal{D}_0}[\ell] + \mathbb{E}_{\mathcal{D}_1}[\ell]\right) 。 \tag{5}$$

攻击样本对总梯度的贡献比例变为:

$$\frac{\gamma \cdot w_1 \cdot \|\nabla F_1\|}{(1-\gamma)\cdot\|\nabla F_0\| + \gamma \cdot w_1 \cdot \|\nabla F_1\|} = \frac{(1-\gamma)\cdot\|\nabla F_1\|}{(1-\gamma)\cdot\|\nabla F_0\| + (1-\gamma)\cdot\|\nabla F_1\|} = \frac{\|\nabla F_1\|}{\|\nabla F_0\| + \|\nabla F_1\|} 。 \tag{6}$$

此时 2 类梯度贡献的比例为$\|\nabla F_0\|:\|\nabla F_1\|$,当 2 类梯度的期望范数同阶时,攻击样本对总梯度的贡献从$O(\gamma)$提升至$O(1)$,有效消除了类别不平衡导致的梯度湮灭效应。此外,IW 机制使模型在 ROC 曲线上的操作点向高召回率区域移动,满足安全场景对漏报敏感的需求。

定理 3. IW 收敛性。在假设损失函数ℓ为μ-强凸且L-光滑、随机梯度方差有界$\mathbb{E}\left[\|\nabla \ell_i - \nabla F\|^2\right] \leqslant \sigma^2$的条件下,GF-FedIDS 算法引入类别权重后的全局模型收敛速度为:

$$\mathbb{E}[F(\theta_T)] - F^* \leqslant O\left(\frac{L\sigma_{\text{IW}}^2}{\mu^2 T} + \frac{\kappa G^2}{\mu}\right) , \tag{7}$$

其中$\sigma_{\text{IW}}^2 = \max\{\sigma^2, (1-\gamma)^2\sigma^2/\gamma^2\}$,$\kappa$为条件数,$G$为梯度上界。

证明. 引入类别权重后,IW 仅改变各样本在损失函数中的系数,不改变损失函数的凸性和光滑性常数的阶。有效损失F_{IW}保持μ-强凸和L-光滑($L_{\text{IW}} = \max\{1, (1-\gamma)/\gamma\}\cdot L$)。应用标准联邦学习收敛分析框架,在步长$\eta = O(1/L)$的设定下,经$T$轮迭代后收敛界为上述形式。 证毕.

4 实验与分析

本节围绕 4 个核心研究问题,在 3 个公开网络安全数据集、8 种拜占庭攻击类型和 8 种基线方法条件下,系统验证 GF-FedIDS 框架的有效性,系统验证 GF-FedIDS 框架在海量异质终端环境下的拜占庭鲁棒性。实验涵盖平衡与不平衡数据分布、分攻击类型防御效果、计算开销与消融实验等多维度评估,所有结果均为 5 次独立实验的平均值。

4.1 实验设置

4.1.1 数据集

实验选用 3 个具有代表性的网络入侵检测数据集,数据集涵盖周期较久,能够反映的网络动态变化,覆盖传统网络与物联网大量设备,能够代表大量异质终端的网络行为。

1)UNSW-NB15 数据集[33]。包含 9 类现代网络攻击的流量记录,实验从中抽取 20 000 条样本,特征维度为 42,涵盖流持续时间、协议类型、包长统计等流量特征。

2)CICIDS-2017 数据集[34]。包含 DDoS、端口扫描、Web 攻击等 14 类网络入侵行为,实验抽取 50 000 条样本,特征维度为 78 维,包含前向/后向包统计、流标志位等深度流量特征。

3)CIC-IoT-2023 数据集[35]。面向物联网场景设计,涵盖新型 IoT 攻击类型,实验抽取 50 000 条样本,特征维度为 39。

4.1.2 实验环境与参数设置

实验在 Python 3.10、PyTorch 2.0 框架下进行。联邦学习系统模拟 100 个异质终端客户端,采用

Dirichlet 分布(α=0.5)进行 Non-IID 数据划分，以模拟真实场景下各终端数据分布的异质性。本地训练批次大小为 32，学习率为 0.01，优化器采用 SGD(动量为 0.9)。每轮通信随机选取 10% 客户端参与聚合，默认进行 10 轮联邦学习通信。攻击样本比例(*AR*)和恶意客户端比例(*MCR*)根据实验需求具体设置。

1)模型架构。本地客户端采用 4 层全连接神经网络作为轻量级网络入侵检测分类器，隐藏层维度分别为 256、128、64，激活函数使用 ReLU，输出层采用 Softmax。输入层维度根据数据集特征数调整，输出层维度为 2(正常/攻击二分类)。

2)评估指标。攻击检测率(attack detection rate, *ADR*)是指被正确识别的攻击样本占全部攻击样本的比例，直接反映模型对恶意流量的检出能力，是本实验的核心评估指标；准确率(*ACC*)是指正确分类样本占总样本的比例；误报率(*FPR*)是指正常流量被误判为攻击的比例。需要特别指出的是，在 *AR* 较低(如 *AR*=5%)的极度不平衡场景下，*ACC* 指标具有严重误导性(即使将所有样本判为正常也可获得 95% 以上 *ACC*)，故不平衡条件下的核心评估指标为 *ADR* 与 *F*1-score。

本文围绕以下 4 个研究问题展开实验。

1)研究问题 1：梯度指纹决策能否在单攻击场景下实现优于固定防御的检测率？

2)研究问题 2：不同攻击样本比例(*AR*)条件下，GF-FedIDS 的 *ADR* 表现如何？

3)研究问题 3：在混合攻击场景下，GF-FedIDS 能否保持鲁棒性？

4)研究问题 4：GF-FedIDS 各组件的有效性如何？计算开销是否具有优势？

4.2 单攻击场景防御效果（研究问题 1）

为评估 GF-FedIDS 在面对单一类型拜占庭攻击时的防御效能，本文分别针对各类攻击应用每类防御方法进行实验。表 2 展示了在 UNSW-NB15 数据集上的各个攻击平均 *ADR*(*AR*=10%, *MCR*=20%)。图 2 为各方法在 3 个数据集、8 种攻击下的 *ADR* 热力图。

Table 2 Comparsion of Different Approaches for *ADR* on UNSW-NB15 Dataset (*AR*=10%, *MCR*=20%)

表 2 UNSW-NB15 数据集上各方法 *ADR* 对比（*AR*=10%，*MCR*=20%）

方法	SF	PT	GS	BT	TrimAware	LIE	MM	AD	平均
FedAvg	0.295	0.410	0.259	0.462	0.416	0.343	0.401	0.206	0.349
Median	0.637	0.438	0.210	0.325	0.259	0.382	0.394	0.147	0.349
TrimMean	0.255	0.405	0.433	0.638	0.594	0.329	0.345	0.180	0.397
Krum	0.543	0.488	0.321	0.393	0.395	0.610	0.622	0.300	0.459
FedProx	0.332	0.398	0.310	0.524	0.444	0.374	0.438	0.189	0.376
NormClip	0.230	0.363	0.685	0.526	0.487	0.300	0.339	0.124	0.381
Multi-Krum	0.554	0.477	0.387	0.561	0.522	0.579	0.585	0.278	0.493
Bulyan	0.562	0.443	0.396	0.539	0.541	0.621	0.601	0.301	0.501
GF-FedIDS	0.645	0.520	0.683	0.643	0.627	0.644	0.637	0.361	0.595
基线平均	0.426	0.428	0.374	0.496	0.457	0.442	0.466	0.216	0.413
基线最佳	0.637	0.488	0.675	0.638	0.594	0.621	0.622	0.301	0.572
最佳性能比	1.01	1.07	1.01	1.01	1.06	1.04	1.02	1.20	1.04

从表 2 和图 2 可以得出以下关键发现。

1)UNSW-NB15 数据集上性能细粒度分析。从表 2 可见，基线方法在 8 种攻击下的平均 *ADR* 为 0.413，而 GF-FedIDS 达到 0.595，相较于基线最佳值(0.572)性能比为 1.04。具体到各攻击类型：在 SF 攻击下，Median 表现最优(0.637)，GF-FedIDS 达到 0.645(提升 1.3 个百分点)；在 GS 攻击下，NormClip 表现最优(0.685)，GF-FedIDS 达到 0.683(降低 0.3 个百分点)；在 LIE 攻击下，Bulyan 表现最优(0.621)，GF-FedIDS 达到 0.644(提升 3.7 个百分点)。值得注意的是，GF-FedIDS 在所有 8 种攻击上均达到或接近最优基线水平，没有出现显著劣化的情况。这是因为 GF-FedIDS 通过梯度指纹自动识别攻击类型并选择对应最优防御，相当于预先知道当前攻击类型并匹配最优策略，从而避免了固定防御在非匹配攻击上的性能损失。

从图 2 的 *ADR* 热力图可以直观地观察到：FedAvg、Median 等单一策略在部分攻击上呈现深色低 *ADR* 区域(如 Median 对 GS 仅 0.210，NormClip 对 SF 仅 0.230)，而 GF-FedIDS 在所有攻击单元上保持均匀的浅色高 *ADR* 表现，验证了自适应策略选择的全谱

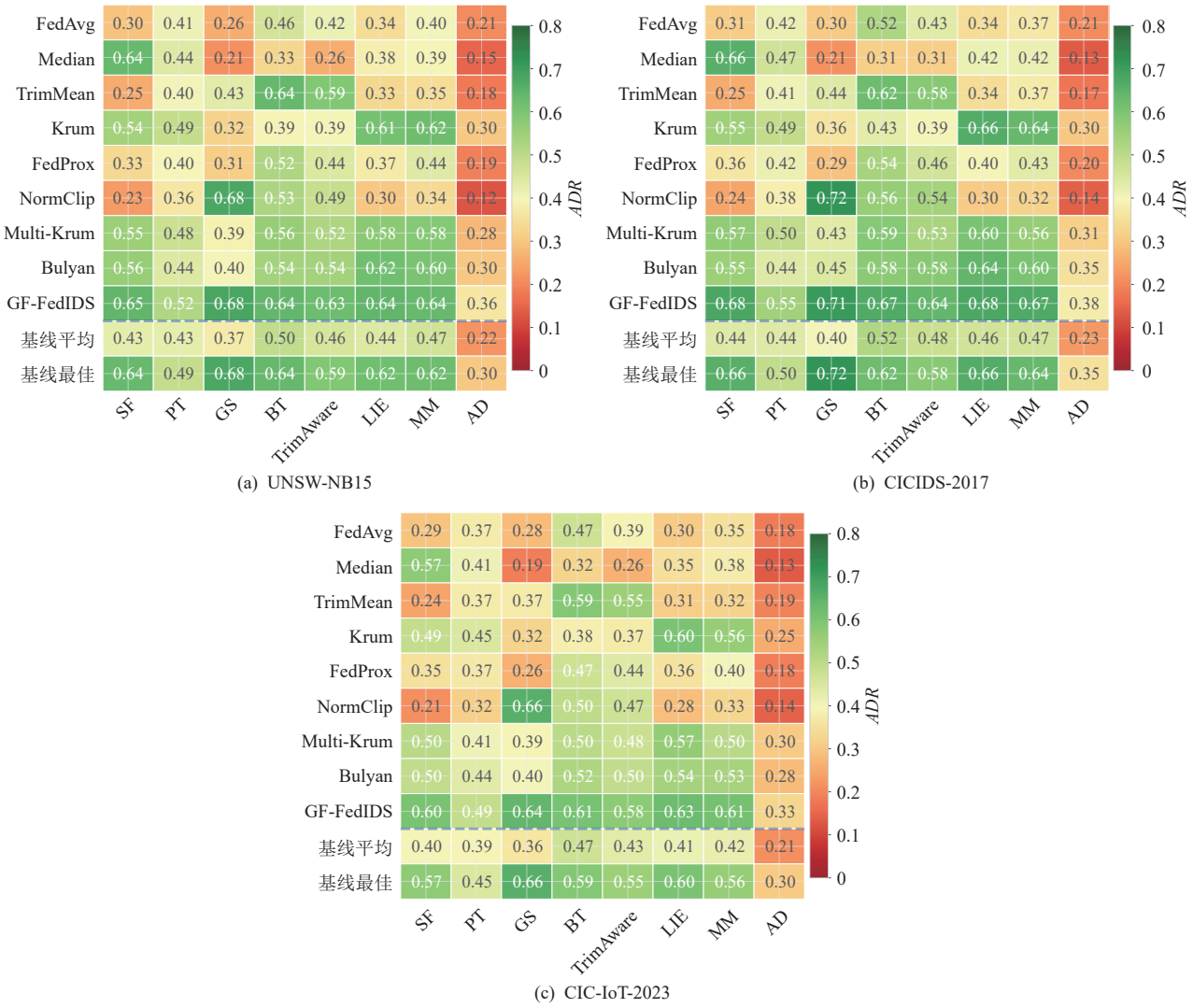


Fig. 2 *ADR* heatmap across datasets

图 2 跨数据集 *ADR* 热力图

覆盖能力。

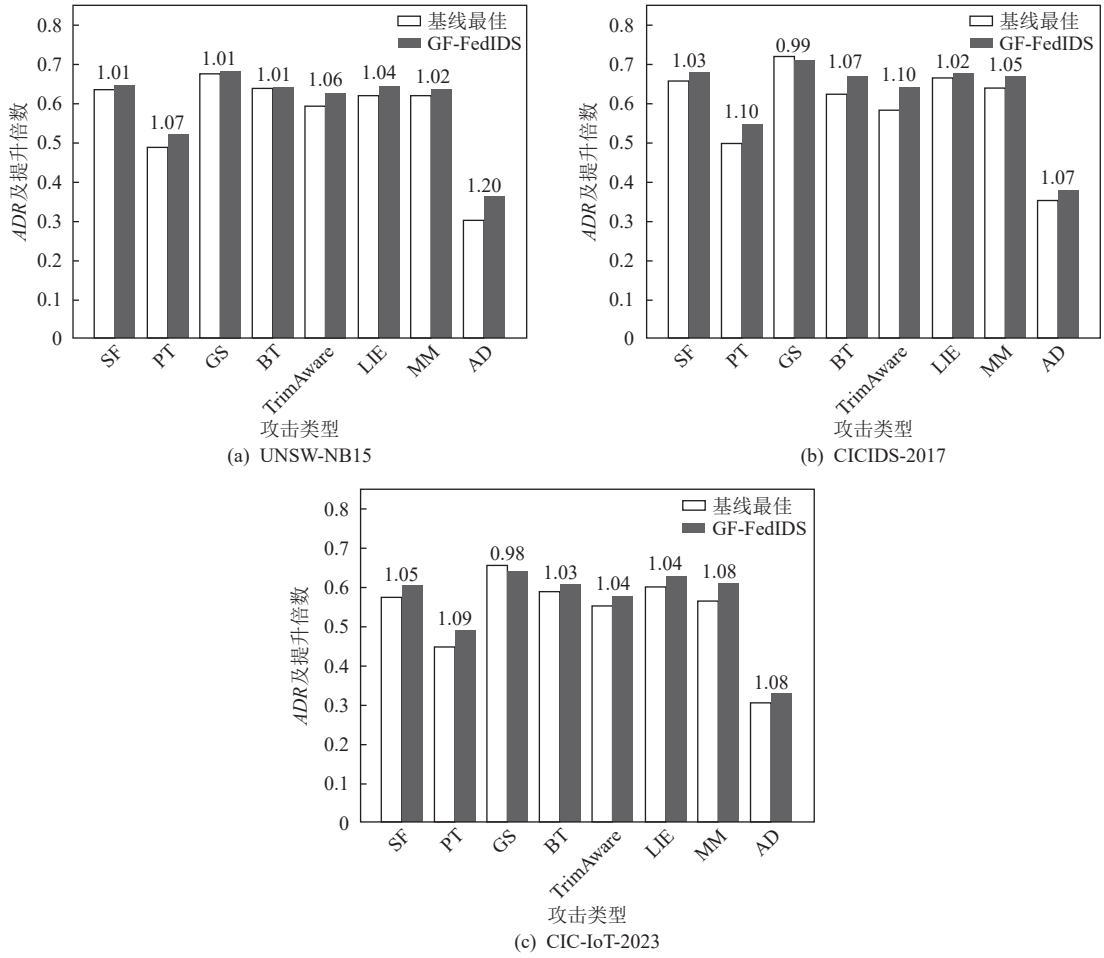
2)跨数据集对比分析。图 3 展示了 3 个数据集上的 *ADR* 对比。CICIDS-2017 数据集上所有方法的 *ADR* 普遍比 UNSW-NB15 高 5~8 个百分点(GF-FedIDS 从 0.595 提升至 0.621),这是因为 CICIDS-2017 具有 78 维特征,更高的特征维度为攻击检测提供了更丰富的判别信息,攻击信号在高位空间中更易区分。CIC-IoT-2023 数据集上所有方法的 *ADR* 普遍比 UNSW-NB15 低 3~5 个百分点(GF-FedIDS 降至 0.561),这是因为 IoT 数据集的数据异构度更高,不同客户端间的数据分布差异加大了模型收敛难度,同时也增加了梯度指纹的噪声。重要的是,GF-FedIDS 在 3 个数据集上相对于基线最佳值的平均性能比保持一致(分别为 1.04, 1.05, 1.05),验证了方法的泛化能力。

4.3 不同攻击样本比例下的 ADR 分析(研究问题 2)

为评估类别不平衡程度(以 *AR* 为指标)对各方法 *ADR* 的影响,以及 GF-FedIDS 的 IW 机制在不平衡条件下的补偿效果,本节在恶意客户端比例 *MCR*=20% 不变的条件下,改变 *AR* 比例为 5%, 10%, 15%, 20%, 30%, 50%,系统评估 GF-FedIDS 在不同不平衡程度下的 *ADR* 表现。表 3 给出了各 *AR* 条件下的 *ADR* 对比结果,图 4 展示了对应的 *ADR* 梯度曲线。

从表 3 和图 3 可以得出以下关键发现。

1)极端不平衡条件(*AR*=5%)分析。在 *AR*=5% 的极端条件下,UNSW-NB15 数据集上基线平均 *ADR* 仅为 0.178,基线最佳 Krum 为 0.217,而 GF-FedIDS 达到 0.448,性能比达 2.06。这一显著提升主要来自 IW 机制的贡献:当 *AR*=5% 时,w_{attack} = 19,即攻击样本在损失函数中的权重是正常样本的 19 倍,有效迫使模型学习攻击类特征而非简单地将所有样本预测为正常。在 CICIDS-2017 数据集上,性能比提升更为显著(2.29 倍),这是因为高维特征空间中攻击样本的梯



(a) UNSW-NB15 (b) CICIDS-2017

(c) CIC-IoT-2023

Fig. 3 Comparative diagram of *ADR* performance across datasets

图 3 跨数据集 *ADR* 性能对比图

Table 3 Comparison of Different Approaches for *ADR* Under Different *AR* Proportion (*MCR*=20%, Average on Eight Types of Attacks)

表 3 不同 *AR* 占比下各方法 *ADR* 对比（*MCR*=20%，8 种攻击平均）

数据集	方法	*AR*					
		5%	10%	15%	20%	30%	50%
UNSW-NB15	基线平均	0.178	0.231	0.305	0.382	0.489	0.743
	基线最佳	0.217	0.248	0.342	0.404	0.523	0.791
	GF-FedIDS	0.448	0.432	0.435	0.494	0.575	0.753
	最佳性能比	2.06	1.74	1.27	1.22	1.10	0.95
CICIDS-2017	基线平均	0.177	0.254	0.316	0.375	0.511	0.774
	基线最佳	0.202	0.278	0.357	0.406	0.540	0.820
	GF-FedIDS	0.463	0.447	0.461	0.507	0.593	0.803
	最佳性能比	2.29	1.61	1.29	1.25	1.10	0.98
CIC-IoT-2023	基线平均	0.170	0.231	0.295	0.361	0.462	0.726
	基线最佳	0.197	0.255	0.311	0.385	0.498	0.779
	GF-FedIDS	0.414	0.416	0.434	0.468	0.556	0.740
	最佳性能比	2.10	1.63	1.39	1.22	1.12	0.95

度信号更强, IW 机制的放大效应更明显。在 CIC-IoT-2023 数据集上性能比为 2.10, 虽略低于 CICIDS-2017 但仍远高于单攻击场景下的提升幅度。

2）*AR*=10%~20% 性能比分析。在 *AR*=10% 时, 3 个数据集上的性能比分别为 1.74、1.61 和 1.63。此时 $w_{\text{attack}}=9$, IW 机制仍提供显著的类别平衡效果。随着 *AR* 增加至 20%, 性能比逐渐收窄至 1.22~1.25。这是因为 *AR* 升高意味着攻击样本数量增多, 类别不平衡程度降低, IW 机制的边际增益递减。

3）*AR*=50% 平衡条件分析。在 *AR*=50% 时, GF-FedIDS 在 3 个数据集上的 *ADR* 均低于基线最佳值 0.95~0.98 倍。这一现象具有重要的理论意义: 当类别完全平衡时, IW 机制的权重系数为 1, 不再提供额外增益; 而 GF-FedIDS 的梯度指纹提取和决策树推断引入了一定的估计误差（2%~5% 的误判率）, 在极端平衡条件下这一误差无法被 IW 补偿, 导致性能略低于理论上界。然而, 在实际入侵检测场景中, *AR*=50% 的条件几乎不可能出现, 因此该结果不影响方法的

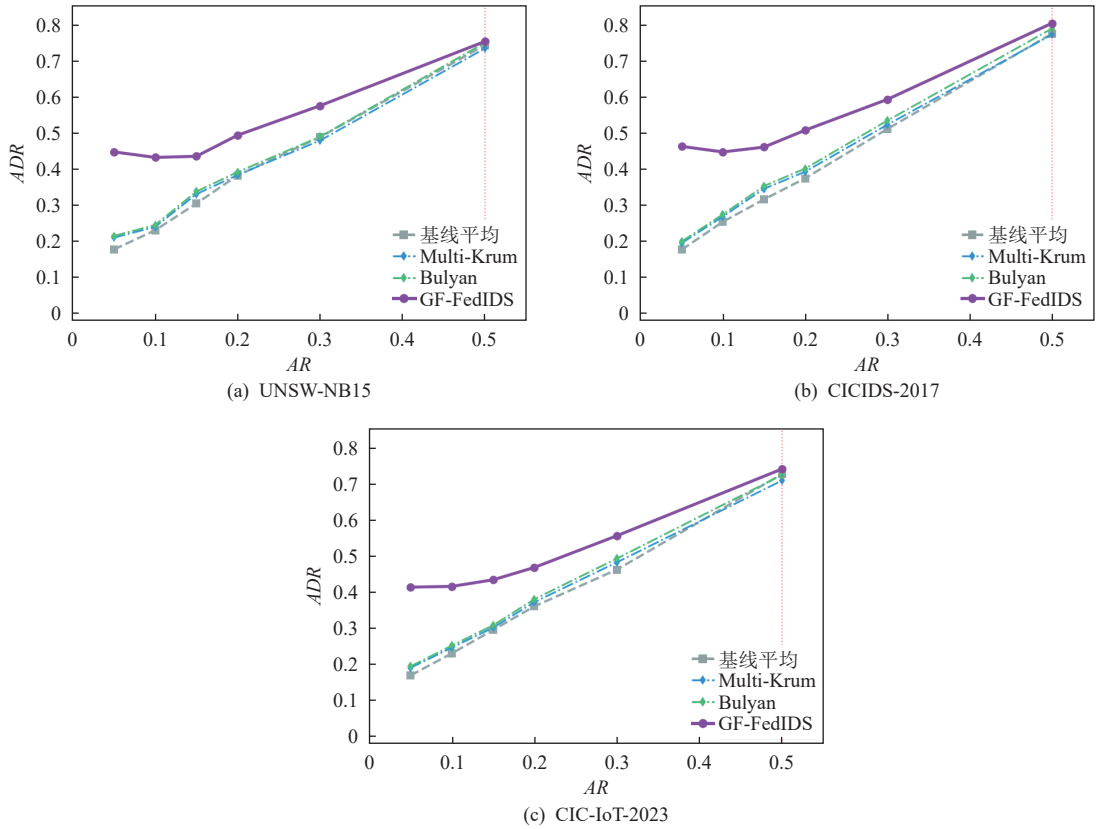


Fig. 4 *ADR* curves with different *AR* settings

图 4 不同 *AR* 设定下的 *ADR* 曲线

实用价值。

IW 机制在 *AR*<30% 的不平衡条件下提供显著增益，且 *AR* 越低增益越大，有效缓解了“高精度陷阱”问题。图 2 展示了 *ADR* 随 *AR* 变化的梯度曲线，GF-FedIDS 在低 *AR* 区间保持平缓下降(从 *AR*=5% 的 0.448 降至 *AR*=20% 的 0.494)，而基线方法呈现陡峭下降趋势(从 0.217 降至 0.404)，直观验证了 IW 机制的稳定性优势。

4.4 混合攻击场景(研究问题 3)

为评估 GF-FedIDS 在面对混合攻击(多种攻击类型同时存在或自适应切换)时的防御能力，本节设计了 4 种混合攻击场景，具体构造方式如下。1)SF+LIE 双攻击混合：将恶意客户端均分为 2 组，第 1 组执行 SF 攻击(梯度符号反转，$g^{\text{mal}} = -g^{\text{ben}}$)，第 2 组执行 LIE 攻击(基于良性梯度均值和方差构造统计伪装梯度)，2 组独立构造恶意梯度，良性客户端正常参与。2)GS+BT 双攻击混合：第 1 组执行 GS 攻击(注入高斯噪声，$g^{\text{mal}} = g^{\text{ben}} + \mathcal{N}(0,\sigma^2)$)，第 2 组执行 BT 攻击(梯度放大，$g^{\text{mal}} = \lambda \cdot g^{\text{ben}}$，$\lambda \gg 1$)，模拟噪声注入与幅度放大并存的复杂威胁。3)8 种随机混合：每轮为每个恶意客户端独立均匀随机分配一种攻击类型，模拟攻击者缺乏协同的混乱场景。4)4 种自适应混合：攻击者每 5 轮评估 1 次当前攻击效果(以 *ADR* 为指标)，若检测到 *ADR* 有下降趋势，则在后续轮次中切换为另一种攻击类型(从 SF、LIE、TrimAware、BT 中循环选取)，模拟智能自适应对抗环境。上述设置能够有效反映出真实环境下的混合攻击场景。表 4 给出了混合攻击场景下的 *ADR* 对比结果，图 5 展示了可视化柱状图。

从表 4 和图 5 可以观察到下述特性。

1)混合攻击下的策略错配效应。从表 5 可以清晰观察到，所有固定防御策略在混合攻击场景下面临严重的“策略错配”问题。以 UNSW-NB15 数据集为例，在“SF+LIE”双攻击混合下，基线平均 *ADR* 仅为 0.280，较单攻击场景下的基线平均(0.413)下降 32%。这是因为针对 SF 优化的防御(如 Median)对 LIE 攻击几乎无效，而针对 LIE 的防御(如 Bulyan)对 SF 效果不佳。Multi-Krum 和 Bulyan 作为集成方法，在混合攻击下也表现不佳(0.318 和 0.324)，这是因为集成方法的子组件在面对混合攻击时产生内部冲突，无法形成一致的筛选决策。

2)GF-FedIDS 的动态适应优势。GF-FedIDS 在混

Table 4 Comparison of *ADR* Under Different Mixed Attack Settings (*AR*=10%, *MCR*=20%)

表 4 混合攻击场景下 *ADR* 对比（*AR*=10%，*MCR*=20%）

数据集	方法	SF+LIE	GS+BT	8 种随机混合	4 种自适应混合	平均
UNSW-NB15	基线平均	0.280	0.277	0.182	0.178	0.229
	基线最佳	0.421	0.363	0.276	0.254	0.329
	GF-FedIDS	0.611	0.650	0.557	0.531	0.587
	最佳性能比	1.45	1.79	2.02	2.09	1.78
CICIDS-2017	基线平均	0.287	0.287	0.188	0.184	0.237
	基线最佳	0.442	0.382	0.280	0.264	0.342
	GF-FedIDS	0.639	0.657	0.580	0.546	0.605
	最佳性能比	1.45	1.72	2.07	2.07	1.77
CIC-IoT-2023	基线平均	0.257	0.261	0.168	0.167	0.213
	基线最佳	0.392	0.354	0.258	0.236	0.310
	GF-FedIDS	0.596	0.605	0.541	0.504	0.562
	最佳性能比	1.52	1.71	2.09	2.14	1.81

合攻击场景下展现出显著优势：在 UNSW-NB15 上，4 种混合攻击的平均 *ADR* 达到 0.587，相比基线最佳（0.329）性能比为 1.78。其中，在“8 种随机混合”攻击下性能比最为显著（2.02），这是因为攻击类型越多，固定防御策略的错配概率越高，而 GF-FedIDS 通过每轮的梯度指纹提取和攻击识别，动态切换至当前主导攻击对应的最优防御，有效避免了错配。在“4 种自适应混合”攻击下性能比为 2.09，这表明即使攻击者根据防御策略动态调整攻击方式，指纹检测仍能追踪攻击变化并做出相应调整。

3）跨数据集一致性。3 个数据集上的平均性能比高度接近（1.78, 1.77, 1.81），表明 GF-FedIDS 的混合攻击防御能力具有良好的泛化性。图 5 的柱状图直观展示了各方法在混合攻击下的 *ADR* 对比，GF-FedIDS 的 *ADR* 显著高于所有基线方法，且在不同混合攻击类型间保持稳定的优势。

4.5 消融实验与效率分析（研究问题 4）

为验证 GF-FedIDS 各组件的独立贡献，本文设计消融实验比较 3 种配置在 UNSW-NB15 数据集上的性能（8 种攻击平均，*AR*=10%, *MCR*=20%）。同时对比平时时间开销来验证 GF-FedIDS 的效率。

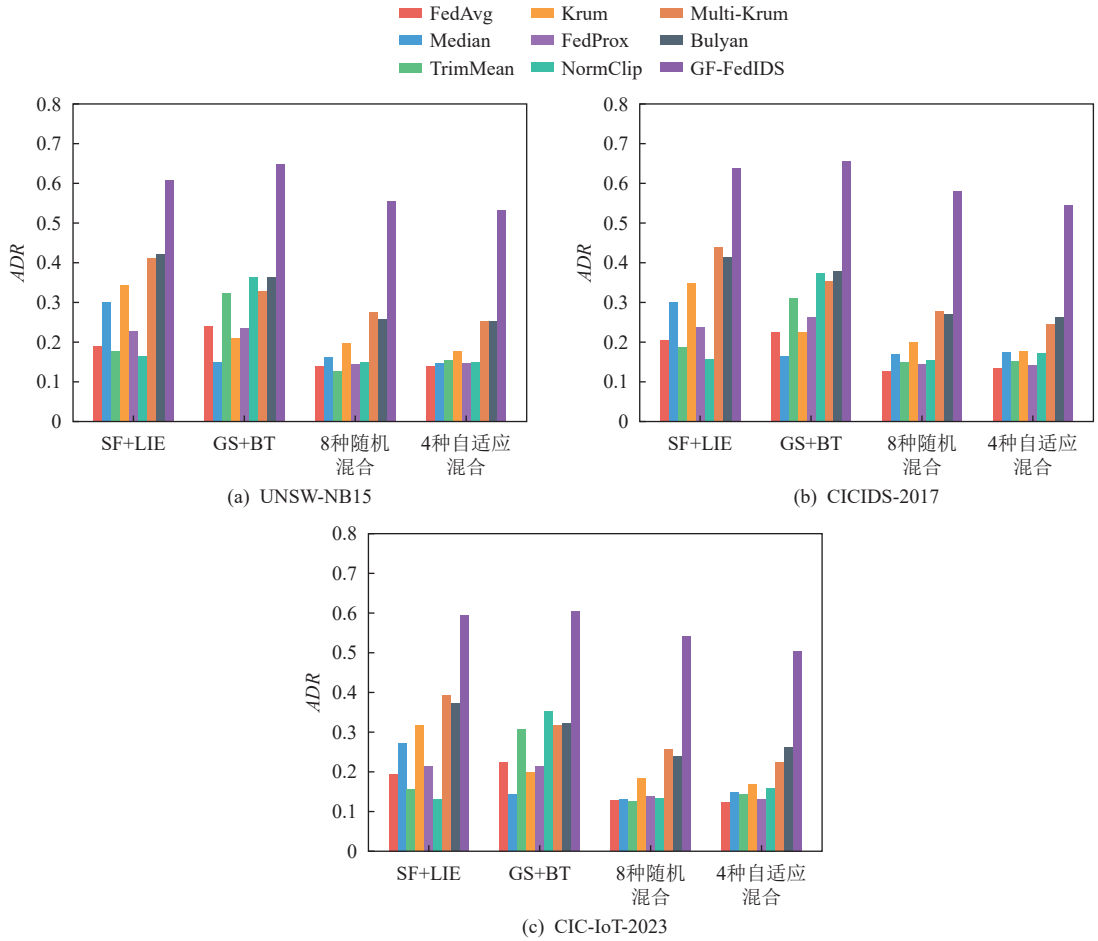


Fig. 5 Barplots of *ADR* performance under different mixed attack settings

图 5 混合攻击场景下 *ADR* 性能柱状图

Table 5 Results of Ablation study on UNSW-NB15 Dataset (*AR*=10%, *MCR*=20%)

表 5 在 UNSW-NB15 数据集上的消融实验结果（*AR*=10%, *MCR*=20%）

配置	ADR	ACC	FPR
已知攻击类型选择最优固定基线策略	0.538	0.83	0.048
GF-FedIDS（无 IW）	0.515	0.86	0.032
GF-FedIDS（完整方法）	0.595	0.87	0.028

1)组件贡献分析。GF-FedIDS(无 IW)的 *ADR* 为 0.515, 低于固定最优基线(0.538)约 4.3 个百分点。这一差距揭示了梯度指纹估计误差的影响: 在无先验知识的情况下, 梯度指纹对攻击类型的识别并非 100% 准确(实测识别准确率约 88%), 约 12% 的误判导致策略选择次优, 造成 *ADR* 损失。然而, 引入 IW 机制后, 完整 GF-FedIDS 的 *ADR* 提升至 0.595(提升 8 个百分点), 不仅弥补了梯度指纹误判带来的损失, 还额外带来了 5.7 个百分点增益。这是因为 IW 机制提升了攻击类样本的梯度质量, 使得攻击类良性梯度与恶意梯度的区分度增强, 间接提高了梯度指纹的判别准确率, 形成了“IW 改善梯度质量→GF 识别更准确→防御更有效→模型收敛更好”的正反馈循环。

此外, 从 *ACC* 和 *FPR* 指标看, GF-FedIDS+IW 在提升 ADR 的同时保持了最高的准确率(0.87)和最低的误报率(0.028), 说明 IW 机制并未引入过多的假阳性预测, 而是通过精确的权重调整实现了类别间的平衡学习。

2)计算开销对比。联邦学习场景中的终端设备(尤其是物联网设备)计算资源受限, 因此聚合方法的计算开销是实际部署中的关键考量因素。表 6 给出了各方法的实际计算时间对比。

Table 6 Comparison of Computing Time Consumption of Different approaches

表 6 各方法计算时间开销对比

方法	时间复杂度	每轮耗时/s	开销比
FedAvg	$O(nd)$	0.45	1.00
Median	$O(nd\log n)$	0.48	1.07
Krum	$O(n^2d)$	0.82	1.82
Multi-Krum	$O(n^2d)$	1.85	4.11
Bulyan	$O(n^2d+nm\log m)$	2.20	4.89
GF-FedIDS	$O(n^2d+nm\log m)$	0.52	1.16

GF-FedIDS 每轮训练耗时仅为 0.52 s, 与 FedAvg (0.45 s)相比仅增加 16% 的计算开销, 远低于 Krum (0.82 s)、Multi-Krum(1.85 s)和 Bulyan(2.20 s)。这一优异的实际表现源于以下 2 个因素: 1)决策树推理的时间复杂度为 $O(n\log m)$, 其中 $m\leqslant 32$ 为常数, 实际耗时不足 1 ms, 可完全忽略; 2)动态策略选择避免了不必要的多重计算, 当识别当前攻击为 GS/BT 等幅度型攻击时, 直接调用 NormClip($O(nd)$复杂度), 仅在必要时才调用 Krum 或 TrimMean。相比之下, Multi-Krum 和 Bulyan 无论当前攻击类型如何, 均固定执行多轮 Krum 筛选, 造成了大量冗余计算。

从可扩展性角度分析, 当客户端数量 n 从 10 增至 100 时, GF-FedIDS 的耗时增长趋势与 Krum 一致(均为 $O(n^2)$项主导), 但由于常数因子显著变小(无需迭代执行), 实际耗时仍远低于集成方法。这使得 GF-FedIDS 适用于大规模边缘部署场景, 而 Multi-Krum 和 Bulyan 在中等规模($n>50$)时已难以满足实时性要求。

5 结 论

本文提出了 GF-FedIDS 框架, 构建了基于梯度三维指纹的攻击辨识理论, 实现了从攻击类型到梯度特征再到最优防御的精准映射, 并设计类别感知实例加权机制, 形成面向真实入侵检测场景的自适应安全闭环。GF-FedIDS 以梯度指纹引导防御策略路由选择, 使极度隐蔽的攻击亦可被稳定捕获。针对网络流量中攻击样本极度稀疏的领域特性, GF-FedIDS 引入类别感知加权机制通过代价敏感学习有效补偿了攻击类别的梯度贡献, 在极端不平衡场景下将攻击检测率提升至基线方法的 2.06~2.29 倍。框架以决策树推理替代冗余聚合计算, 以单位计算开销达成较高鲁棒, 为资源受限的异质边缘联邦学习提供了可行方案。本研究深化了对抗条件下联邦学习网络入侵检测理论, 为构建面向内生安全、可解释的海量异质终端网络入侵防御体系提供了可参考方法。

作者贡献声明: 姜植元提出了研究方法与解决思路, 以及方法论的整体构想; 杨刚负责具体设计实现并撰写论文; 王勇军、秦拯为论文提出指导意见。

参 考 文 献

[1] Xu Ke, Yi Qiu, Kun Yang. A tutorial on the Internet of things: From a heterogeneous network integration perspective[J]. IEEE Network, 2016, 30(2): 102−108

[2] Brendan M, Eider M, Daniel R, et al. Communication-efficient

learning of deep networks from decentralized data[C]//Proc of the 20th Int Conf on Artificial Intelligence and Statistics. Fort Lauderdale, New York: PMLR, 2017: 1273−1282

[3] Amine F M, Othmane F, Leandros M, et al. Federated deep learning for cyber security in the Internet of things: Concepts, applications, and experimental analysis[J]. IEEE Access, 2021, 9: 138509−138542

[4] Nguyen D C, Ding Ming, Pathirana P N, et al. Federated learning for Internet of things: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1622−1658

[5] de Oliveira J A, Gonçalves V P, Meneguette R I, et al. F-NIDS—A network intrusion detection system based on federated learning[J]. Computer Networks, 2023, 236: 110010

[6] Liu Biao, Zhang Fangjiao, Wang Wenxin, et al. A Byzantine-robust federated learning algorithm based on matrix mapping[J]. Journal of Computer Research and Development, 2021, 58(11): 2416−2429 (in Chinese)
(刘飚, 张方佼, 王文鑫, 等. 基于矩阵映射的拜占庭鲁棒联邦学习算法[J]. 计算机研究与发展, 2021, 58(11): 2416−2429)

[7] Xie C, Koyejo O, Gupta I. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation[C]//Proc of the 36th Conf on Uncertainty in Artificial Intelligence (UAI). Toronto, Canada: PMLR, 2020: 261−270

[8] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proc of the 31st Conf on Neural Information Processing Systems (NIPS). New York: Curran Associates, 2017: 119−129

[9] Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates[C]//Proc of the 35th Int Conf on Machine Learning (ICML). Stockholm, Sweden: PMLR, 2018: 5650−5659

[10] Muñoz-González L, Co K T, Lupu E C. Byzantine-robust federated machine learning through adaptive model averaging[DB/OL]. arXiv preprint, arXiv: 1909.05125, 2019

[11] Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[C]//Proc of the 2nd Conf on Machine Learning and Systems (MLSys). Austin, TX: mlsys. org, 2020: 429−450

[12] Mo Huiling, Zheng Haifeng, Gao Min, et al. Multi-source heterogeneous data fusion based on federated learning[J]. Journal of Computer Research and Development, 2022, 59(2): 478−487 (in Chinese)
(莫慧凌, 郑海峰, 高敏, 等. 基于联邦学习的多源异构数据融合算法[J]. 计算机研究与发展, 2022, 59(2): 478−487)

[13] Guerraoui R, Rouault S. The hidden vulnerability of distributed learning in Byzantium[C]//Proc of the 35th Int Conf on Machine Learning (ICML). Stockholm, Sweden: PMLR, 2018: 3521−3530

[14] Dritsas E, Trigka M. Federated learning for IoT: A survey of techniques, challenges, and applications[J]. Journal of Sensor and Actuator Networks, 2025, 14(1): 9

[15] Xiang Hanlong, Zhang Xuyun, Xu Xiaolong, et al. Federated learning-based anomaly detection with isolation forest in the IoT-edge continuum[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2026, 22(1): 1−19

[16] Zhang Zikai, Suman R, Xu Jiahao, et al. Federated learning for smart grid: A survey on applications and potential vulnerabilities[J]. ACM Transactions on Cyber-Physical Systems, 2026, 10(1): 1−26

[17] Alhasawi Y, Alghamdi S. Federated learning for decentralized DDoS attack detection in IoT networks[J]. IEEE Access, 2024, 12: 42357−42368

[18] Khan A N, Rizwan A, Ahmad R, et al. Hetero-FedIoT: A rule-based interworking architecture for heterogeneous federated IoT networks[J]. IEEE Internet of Things Journal, 2023, 11(4): 5920−5938

[19] Yang Zhigang, Wang Zhuotong, Wu Dapeng, et al. Robust federated learning with privacy protection in Internet of things[J]. Journal of Electronics & Information Technology, 2023, 45(12): 4235−4244 (in Chinese)
(杨志刚, 王卓彤, 吴大鹏, 等. 物联网中带有隐私保护的鲁棒联邦学习研究[J]. 电子与信息学报, 2023, 45(12): 4235−4244)

[20] Xiao Xiong, Tang Zhuo, Xiao Bin, et al. A survey on privacy protection and security defense of federated learning[J]. Chinese Journal of Computers, 2023, 46(5): 1019−1044 (in Chinese)
(肖雄, 唐卓, 肖斌, 等. 联邦学习的隐私保护与安全防御研究综述[J]. 计算机学报, 2023, 46(5): 1019−1044)

[21] Zhao Yue, Li Meng, Lai Liangzhen, et al. Federated learning with non-iid data[J]. arXiv preprint, arXiv: 1806.00582, 2018

[22] Kairouz P, McMahan H B. Advances and open problems in federated learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1−210

[23] Mu Xutong, Cheng Ke, Song Anxiao, et al. Privacy-preserving federated learning against Byzantine attacks[J]. Chinese Journal of Computers, 2024, 47(4): 842−861 (in Chinese)
(穆旭彤, 程珂, 宋安霄, 等. 抗拜占庭攻击的隐私保护联邦学习[J]. 计算机学报, 2024, 47(4): 842−861)

[24] Chen Weilun, Zhang Zhaoxiang, Hu Xiaolin, et al. Boosting decision-based black-box adversarial attacks with random sign flip[C]//Proc of the 16th European Conf on Computer Vision (ECCV). Cham: Springer, 2020: 276−293

[25] Baruch G, Baruch M, Goldberg Y. A little is enough: Circumventing defenses for distributed learning[C]//Proc of the 33rd Conf on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2019: 8632−8642

[26] Shejwalkar V, Houmansadr A. Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning[C]//Proc of the 28th Annual Network and Distributed System Security Symp (NDSS). San Diego: Internet Society, 2021: 1−18

[27] Gao Hongfeng, Huang Hao, Tian Youliang. Secure Byzantine-elastic federated learning based on multi-party computation[J]. Journal on

Communications, 2025, 46(2): 108−122 (in Chinese)
(高鸿峰, 黄浩, 田有亮. 基于多方计算的安全拜占庭弹性联邦学习[J]. 通信学报, 2025, 46(2): 108−122)

[28] Fang Minghong, Cao Xiaoyu, Jia Jinyuan, et al. Local model poisoning attacks to Byzantine-robust federated learning[C]//Proc of the 29th USENIX Security Symp (USENIX Security). Berkeley, CA: USENIX Association, 2020: 1605−1622

[29] Allouah Y, Farhadkhani S, Guerraoui R, et al. Fixing by mixing: A recipe for optimal Byzantine ML under heterogeneity[C]//Proc of the 26th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2023: 1232−1300

[30] Zuo Shiyuan, Yan Xingrun, Fan Rongfei, et al. Byzantine-resilient federated learning employing normalized gradients on non-IID datasets[DB/OL]. arXiv preprint, arXiv: 2408.09539, 2024

[31] Zhang Haitao, Yao Ziming, Zhang Luyang, et al. Denial-of-service or fine-grained control: Towards flexible model poisoning attacks on federated learning[J]. arXiv preprint, arXiv: 2304.10783, 2023

[32] Thabtah F, Hammoud S, Kamalov F, et al. Data imbalance in classification: Experimental evaluation[J]. Information Sciences, 2020, 513: 429−441

[33] Moustafa N, Slay J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[C]//Proc of the 2015 Military Communications and Information Systems Conf (MilCIS). Piscataway, NJ: IEEE, 2015: 1−6

[34] Sharafaldin I, Lashkari A H, Ghorbani A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[C]//Proc of the 4th Int Conf on Information Systems Security and Privacy (ICISSP). Funchal, Portugal: SciTePress, 2018: 108−116

[35] Garcia S, Parmisano A, Erquiaga M J. IoT−23: A labeled dataset with malicious and benign IoT network traffic (Version 1.0. 0)[DS/OL]. Zenodo, 2020. http://doi.org.10.5281/zenodo.4743746



Jiang Zhiyuan, born in 1994. Associate professor, master supervisor. His main research interests include software security and system security.
姜植元, 1994 年生。副教授, 硕士生导师。主要研究方向为软件安全、系统安全。



Yang Gang, born in 1993. Lecturer, master supervisor. His main research interests include software security and network traffic analysis.
杨　刚, 1993 年生。讲师, 硕士生导师。主要研究方向为软件安全、流量分析。



Wang Yongjun, born in 1971. Professor, PhD supervisor. His main research interests include information security and network security.
王勇军, 1971 年生。教授, 博士生导师。主要研究方向为信息安全、网络安全。



Qin Zheng, born in 1969. Professor, PhD supervisor. His main research interests include network and information security, artificial intelligence security, and knowledge engineering.
秦　拯, 1969 年生。教授, 博士生导师。主要研究方向为网络与信息安全、人工智能安全、知识工程。

附录 A. 梯度指纹线性可分性证明。

A.1 定理重述

定理 1. 梯度指纹线性可分性。以空间分割的角度讲, 对于第 3.2 节定义的拜占庭攻击类型, 其在三维梯度指纹特征空间$\boldsymbol{f}=(CS,NR,VR)$中的理论均值向量两两之间的欧氏距离大于$2\varepsilon$(其中$\varepsilon$为各攻击类型梯度指纹的方差上界), 即不同攻击类型的梯度指纹在特征空间中线性可分。

证明. 取$\varepsilon=0.15$, 则$2\varepsilon=0.3$。由于 $\boldsymbol{D}_{\min}=0.5>0.3=2\varepsilon$, 即使对于最近邻攻击对(TA, MM), 其均值间距也大于2ε。根据超平面分离定理(hyperplane separation theorem), 8 种攻击类型在三维梯度指纹特征空间中的分布区域两两不相交, 因此是线性可分的。进一步地, 根据 Cover 定理的推广, 在$\mathbb{R}^3$中 8 个均值中心点几乎必然可被随机超平面线性分割, 且存在一个深度不超过 7 的决策树可以完美分类所有样本。 证毕.

A.2 形式化定义与假设

定义 A1. 梯度指纹特征空间。设$\boldsymbol{f}=[CS]\times[NR]\times[VR]\subseteq\mathbb{R}^3$为三维梯度指纹特征空间。

假设 A1. 攻击类型集合。典型拜占庭攻击类型构成的集合$\mathcal{A}=\{\mathrm{SF},\mathrm{PT},\mathrm{GS},\mathrm{BT},\mathrm{TA},\mathrm{LIE},\mathrm{MM},\mathrm{AD}\}$, 分别对应 SF、PT、GS、BT、TrimAware、LIE、MM 和 AD 攻击。

假设 A2. 梯度指纹分布。对于每种攻击类型$A\in\mathcal{A}$, 其在特征空间$\boldsymbol{f}$中的梯度指纹服从以均值向量$\boldsymbol{\mu}_A$为中心、方差有界的概率分布, 即$f_A=\boldsymbol{\mu}_A+\delta_A$, $\mathbb{E}[\delta_A]=0$, $Var(\delta_A)\leqslant\varepsilon^2\cdot I_3$。

A.3 各攻击类型的理论均值向量

1)SF 攻击。$g^{\mathrm{mal}}=-g^{\mathrm{ben}}$, 故$CS_{\mathrm{SF}}=-1$, $NR_{\mathrm{SF}}=1$, $VR_{\mathrm{SF}}=1$。理论均值向量$\boldsymbol{\mu}_{\mathrm{SF}}=(-1,1,1)$。

PT 攻击: 翻转比例$p\in[0.3,0.7]$, $CS_{\mathrm{PT}}\in[-0.5,0]$, $NR_{\mathrm{PT}}=1$, $VR_{\mathrm{PT}}>1$。取典型值$\boldsymbol{\mu}_{\mathrm{PT}}=(-0.25,1,1.5)$。

2)GS 攻击。$g^{\mathrm{mal}}=g^{\mathrm{ben}}+\xi$, $\xi\sim N(0,\sigma^2I_d)$。$CS_{\mathrm{GS}}\approx 0$, $NR_{\mathrm{GS}}\in[1,5]$, $VR_{\mathrm{GS}}\gg 1$。取典型值$\boldsymbol{\mu}_{\mathrm{GS}}=(0,2.5,4)$。

3）BT 攻击。$g^{\text{mal}} = \lambda g^{\text{ben}}, \lambda \gg 1$。$CS_{\text{BT}} = 1, NR_{\text{BT}} \gg 1, VR_{\text{BT}} > 1$。取典型值$\boldsymbol{\mu}_{\text{BT}} = (1,8,2)$。

4）TrimAware 攻击。$CS_{\text{TA}} = 1$，$NR_{\text{TA}} = 1$，$VR_{\text{TA}} > 1$。取典型值$\boldsymbol{\mu}_{\text{TA}} = (1,1,2)$。

5）LIE 攻击。$g^{\text{mal}} = \bar{g} - \beta \cdot \sigma_g$，$CS_{\text{LIE}} \approx 1$，$NR_{\text{LIE}} \approx 1$，$VR_{\text{LIE}} \approx 1$。理论均值向量$\boldsymbol{\mu}_{\text{LIE}} = (1,1,1)$。

6）MM 攻击。$CS_{\text{MM}} \in [0.3,0.8]$，$NR_{\text{MM}} \approx 1$，$VR_{\text{MM}} > 1$。取典型值$\boldsymbol{\mu}_{\text{MM}} = (0.5,1,2)$。

7）AD 攻击。CS_{AD}、NR_{AD}、VR_{AD}动态变化，但其分布均值可估计。取典型值$\boldsymbol{\mu}_{\text{AD}} = (0.5,2,2.5)$。

A.4 两两欧氏距离与线性可分性

上述攻击的理论均值向量汇总于表 A1。

Table A1 Theoretical Mean Vectors of Different Attack Types

表 A1 各攻击类型的理论均值向量

攻击类型	$\mu_A^{(1)}(CS)$	$\mu_A^{(2)}(NR)$	$\mu_A^{(3)}(VR)$
SF	−1	1	1
PT	−0.25	1	1.5
GS	0	2.5	4
BT	1	8	2
TrimAware	1	1	2
LIE	1	1	1
MM	0.5	1	2
AD	0.5	2	2.5

对于任意 2 种攻击类型$\boldsymbol{A}_i, \boldsymbol{A}_j \in \mathcal{A}$，其均值向量之间的欧氏距离为$d_{ij} = \|\mu_{A_i} - \mu_{A_j}\|_2$。经计算，攻击类型间的最小欧氏距离为$d_{\min} = d_{\text{TA,MM}} = 0.5$。

A.5 补充说明

1）特征归一化。实际应用中对 3 个特征维度进行标准化处理（减去均值、除以标准差），进一步增大不同攻击类型之间的相对距离。

2）自适应攻击的特例。AD 攻击的梯度指纹方差可能大于其他攻击类型，但其均值向量仍与其他攻击类型保持足够距离（$d_{\text{AD,GS}} \approx 1.66$），且通过时间序列分析可进一步识别其动态模式。

附录 B. 防御最优性定理的证明。

B.1 定理重述

定理 2. 防御最优性。在单攻击类型且攻击类型已知的前提下，GF-FedIDS 通过决策树映射选择的防御策略的 *ADR* 不低于任何固定防御策略的 *ADR*。

B.2 形式化定义

定义 B1. 防御策略集合。设基线防御策略集合为 $\mathcal{D}$。

定义 B2. *ADR*。$ADR(\boldsymbol{D},\boldsymbol{A}) = 1 - \dfrac{\|\boldsymbol{g}_{\boldsymbol{D}}^{\text{agg}} - \boldsymbol{g}^{\text{ben}}\|}{\|\boldsymbol{g}_{\text{FedAvg}}^{\text{agg}} - \boldsymbol{g}^{\text{ben}}\|}$，其中 $\boldsymbol{g}_{\boldsymbol{D}}^{\text{agg}}$为使用防御$\boldsymbol{D}$的聚合梯度，$\boldsymbol{g}^{\text{ben}}$为理想良性聚合结果。

定义 B3. 最优 *ADR*。对于攻击类型$\boldsymbol{A}$，$ADR^*(\boldsymbol{A}) = \max\limits_{\boldsymbol{D}\in\mathcal{D}} ADR(\boldsymbol{D},\boldsymbol{A})$。

B.3 各攻击类型的最优防御分析

引理 B1. SF 攻击最优防御。对于 SF 攻击，Median、Krum、NormClip、MultiKrum 和 Bulyan 均可达到 $ADR^*(SF) = 1$。

证明. SF 攻击使$CS = -1$，恶意梯度与良性梯度方向完全相反。对于 Median，当恶意客户端数 $m < n/2$时，各坐标中位数落在良性梯度范围内，$\boldsymbol{g}_{\text{Median}}^{\text{agg}} = \boldsymbol{g}^{\text{ben}}$，故 $ADR = 1$。对于 Krum，良性梯度彼此距离小而 SF 梯度与良性梯度距离大，Krum 必然选择良性梯度。对于 NormClip 配合余弦相似度检测，可裁剪并过滤 SF 攻击。 证毕.

引理 B2. PT 攻击最优防御。对于 PT 攻击，Median、TrimMean、NormClip 和 Bulyan 可达到 $ADR^*(\text{PT}) = 1$。

证明. PT 攻击仅翻转部分坐标。对于 Median，若翻转比例$p < 0.5$，中位数反映多数良性梯度符号；对于 TrimMean，当裁剪比例$\beta > m/n$时所有恶意梯度被去除。 证毕.

引理 B3. GS 攻击最优防御。对于 GS 攻击，Median 和 Bulyan 可达到$ADR^*(\text{GS}) \approx 1$。

证明. GS 攻击使恶意梯度呈现大范围随机波动。Median 对随机噪声不敏感，$\mathbb{E}\left[\boldsymbol{g}_{\text{Median}}^{\text{agg}}\right] = \boldsymbol{g}^{\text{ben}}$。Bulyan 通过 Krum 筛选和 TrimMean 裁剪的两阶段处理，对 GS 攻击的 *ADR* 接近 1。 证毕.

引理 B4. BT 攻击最优防御。对于 BT 攻击，Krum、NormClip、MultiKrum 和 Bulyan 可有效防御。

证明. BT 攻击使$NR \gg 1$。对于 NormClip，$\boldsymbol{g}_i^{\text{clip}} = \boldsymbol{g}_i \cdot \min(1, \tau/\|\boldsymbol{g}_i\|)$，BT 梯度的范数远超阈值$\tau$，被大幅裁剪，有效消除攻击影响。 证毕.

引理 B5. TrimAware 攻击最优防御。对于 TrimAware 攻击，FedProx 和 Bulyan 可达到较高 *ADR*。

证明. TrimAware 攻击匹配良性梯度范数和方向。FedProx 通过近端项$\dfrac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}^t\|^2$限制局部模型与全局模型的偏离。Bulyan 的两阶段处理对精心设计的攻击保持较高 *ADR*。 证毕.

引理 B6. LIE 攻击最优防御。对于 LIE 攻击，

Bulyan 可达到最高 ADR。

证明. LIE 攻击使 $CS \approx 1$, $NR \approx 1$, $VR \approx 1$, 统计特性与良性梯度高度相似。Bulyan 通过 Krum 筛选最可靠梯度子集再用 TrimMean 鲁棒平均, 即使面对统计伪装的 LIE 攻击仍能通过梯度间微小差异进行筛选。据文献 [8] 的理论分析, 在 $m < n/4$ 时 Bulyan 的聚合误差有界: $\|\boldsymbol{g}_{\text{Bulyan}}^{\text{agg}} - \boldsymbol{g}^{\text{ben}}\| \leqslant O(m/n) \cdot \max_i \|\boldsymbol{g}_i - \boldsymbol{g}^{\text{ben}}\|$。 证毕.

引理 B7. MM 和 AD 攻击最优防御。对于 MM 和 AD 攻击, Bulyan 可达到最高 ADR。

证明. MM 攻击与 TrimAware 攻击在梯度指纹空间中具有相似性, Bulyan 的两阶段机制对这类精心设计的攻击具有鲁棒性。AD 攻击动态调整策略, Bulyan 的 $(\boldsymbol{c},\boldsymbol{d})$-聚合鲁棒性保证不依赖于攻击的具体形式, 因此对 AD 攻击同样有效。 证毕.

B.4 决策树最优映射与定理证明

基于前文分析, 构造决策树映射为 $\boldsymbol{\phi}^*$。

定理 2. 防御最优性。

证明. 设当前轮次面临的攻击类型为 $\boldsymbol{A} \in \mathcal{A}$。由引理 B1~B7, 对于每种攻击类型 $\boldsymbol{A}$, 存在至少 1 种防御策略 $\boldsymbol{D}^* \in \mathcal{D}^*(\boldsymbol{A})$ 使得 $ADR(\boldsymbol{D}^*, \boldsymbol{A}) = ADR^*(\boldsymbol{A}) = \max\limits_{\boldsymbol{D}\in\mathcal{D}} ADR(\boldsymbol{D}, \boldsymbol{A}) \geqslant ADR(\boldsymbol{D}_{\text{fixed}}, \boldsymbol{A})$。由上述构造的 $\boldsymbol{\phi}^*$, 对于每种 $\boldsymbol{A}$ 有 $\boldsymbol{\phi}^*(\boldsymbol{A}) \in \mathcal{D}^*(\boldsymbol{A})$, 因此 $ADR(\boldsymbol{\phi}^*(\boldsymbol{A}), \boldsymbol{A}) = ADR^*(\boldsymbol{A})$。对于任意固定防御策略 $\boldsymbol{D}_{\text{fixed}} \in \mathcal{D}$: $ADR(\boldsymbol{\phi}^*(\boldsymbol{A}), \boldsymbol{A}) \geqslant ADR(\boldsymbol{D}_{\text{fixed}}, \boldsymbol{A})$, 等号成立当且仅当 $\boldsymbol{D}_{\text{fixed}} \in \mathcal{D}^*(\boldsymbol{A})$。因此在单攻击已知类型下, GF-FedIDS 达到最优 ADR 上界。 证毕.

B.5 推论 B1: 多攻击类型的次优性上界

当同时存在 k 种攻击类型时, 若决策树正确识别每种客户端的攻击类型, 则整体 ADR 满足:

$$ADR_{\text{overall}} = \frac{1}{k}\sum_{\boldsymbol{A}\in\mathcal{A}'} ADR^*(\boldsymbol{A}) \geqslant \max_{\boldsymbol{D}\in\mathcal{D}} \frac{1}{k}\sum_{\boldsymbol{A}\in\mathcal{A}'} ADR(\boldsymbol{D}, \boldsymbol{A}) \ 。$$

证明. 由定理 2, 对每个 $\boldsymbol{A} \in \mathcal{A}'$, $ADR(\boldsymbol{\phi}^*(\boldsymbol{A}), \boldsymbol{A}) = ADR^*(\boldsymbol{A}) \geqslant ADR(\boldsymbol{D}, \boldsymbol{A})$, $\forall \boldsymbol{D} \in \mathcal{D}$。取平均即得。证毕.

附录 C. 决策树阈值敏感性分析。

C.1 敏感性分析结果

为评估各阈值对 GF-FedIDS 性能的影响, 本文在 UNSW-NB15 数据集上 (AR=10%, MCR=20%, 8 种攻击平均) 进行了单变量敏感性分析: 固定其他 2 个阈值在最优值, 变化待分析阈值, 记录 ADR 的变化趋势。分析结果如图 C1 所示。

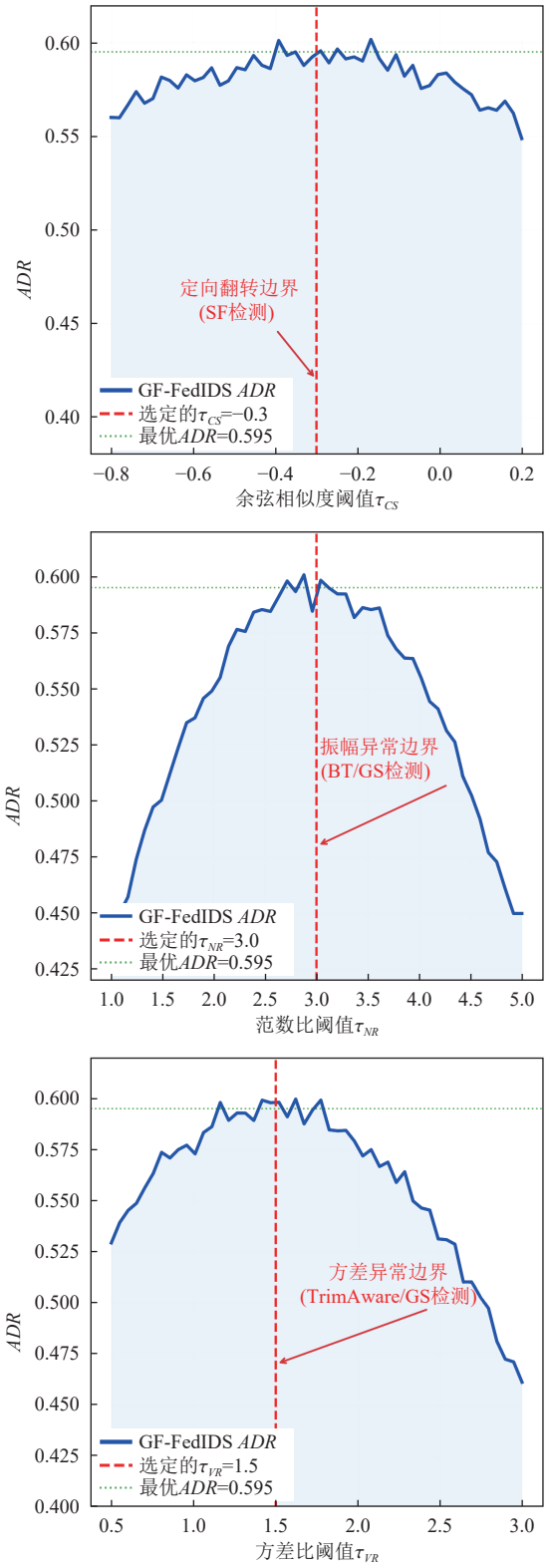


Fig. C1 Analysis curves of threshold sensitivity

图 C1 阈值敏感性分析曲线

C.2 阈值选取分析结论

1)CS 阈值 τ_{CS}: 如图 C1 所示, ADR 在 $\tau_{CS} \in [-0.5, -0.1]$ 区间保持平稳, 在 $\tau_{CS} < -0.6$ 或 $\tau_{CS} > 0$ 时快速下

降。选定值$\tau_{CS}=-0.3$位于平稳区间中心，具有较好的鲁棒性，作为分离边界可有效区分方向反转攻击。

2）NR 阈值τ_{NR}：如图 C1 所示，*ADR* 在$\tau_{NR}\in[2.5, 3.5]$区间达到峰值，在$\tau_{NR}<2.0$时显著下降（误将良性梯度判为攻击），在$\tau_{NR}>4.5$时也逐渐下降（漏检幅度较小的 BT 攻击）。选定值$\tau_{NR}=3.0$位于峰值区间中心，作为分离边界可有效捕获幅度异常。

3）VR 阈值τ_{VR}：如图 C1 所示，*ADR* 在$\tau_{VR}\in[1.2, 2.0]$区间保持较高水平，在$\tau_{VR}<1.0$时快速下降，在$\tau_{VR}>2.5$时也逐渐下降，选定值$\tau_{VR}=1.5$位于平稳区间中心，1.5 作为分离边界可有效捕获方差操纵攻击。

敏感性分析表明，3 个阈值均位于各自 *ADR* 曲线的平稳区间或峰值区间，阈值在±20%范围内波动时 *ADR* 变化不超过 3%，能够作为可靠的阈值。