**Computer Science and Application** 计算机科学与应用, 2026, 16(2), 261-276
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162057

Hans汉斯

# 基于联邦学习的隐私保护和抗投毒攻击方法研究

**张铁雪，王　佳***

新疆大学计算机科学与技术学院，新疆　乌鲁木齐

收稿日期：2026年1月3日；录用日期：2026年2月3日；发布日期：2026年2月11日

## 摘　要

**联邦学习中的参数传输和模型训练，使其面临着投毒攻击和隐私泄露的双重威胁。现有结合隐私保护和抗投毒攻击的联邦学习研究中，通常先对客户端梯度进行加密或扰动再在密文域中执行投毒攻击检测操作，容易模糊或消除恶意梯度所具有的差异性特征，导致检测算法难以准确区分不同类型的投毒攻击。本文提出基于联邦学习的隐私保护和抗投毒攻击方法研究中，采用基于明文的梯度历史信息对客户端类型进行识别，再对筛选出的正常客户端进行隐私保护和安全聚合操作，从而在保障数据机密性的同时提升检测的有效性。考虑符号翻转、噪声注入和标签翻转攻击的普遍性，以及不同投毒攻击在目标、强度和行为模式上的显著差异，引入模型局部梯度的长短历史信息，通过比较不同梯度历史之间的异常差异性实现对多类投毒攻击的有效检测。同时设计周期性投毒攻击检测策略，实现客户端的隐私保护。此外，考虑多链聚合中固定的链数量及链内客户端，设计自适应多链安全聚合方法，增强对客户端集合动态变化的适应性，从而在隐私保护的同时提升聚合效率。实验结果表明，所提算法在MNIST和Fashion-MNIST以及CIFAR10数据集上的平均准确率达到85.18%，较基准算法平均提升约6%，具有良好的多类投毒攻击检测能力，能有效提升模型性能并满足复杂攻击场景下的防御需求。**

## 关键词

**联邦学习，投毒攻击，隐私保护，梯度历史，攻击检测**

# Research on Privacy Protection and Anti-Poisoning Attack Methods Based on Federated Learning

**Tiexue Zhang, Jia Wang***

School of Computer Science and Technology, Xinjiang University, Urumqi Xinjiang

Received: January 3, 2026; accepted: February 3, 2026; published: February 11, 2026

*通讯作者。

文章引用: 张铁雪, 王佳. 基于联邦学习的隐私保护和抗投毒攻击方法研究[J]. 计算机科学与应用, 2026, 16(2): 261-276. DOI: 10.12677/csa.2026.162057

## Abstract

**Parameter transmission and model training in federated learning face the dual threats of poisoning attacks and privacy leakage. Existing privacy-preserving and anti-poisoning federated learning studies typically encrypt or perturb client gradients prior to ciphertext-domain attack detection, which obscures distinctive malicious gradient characteristics and hinders accurate poisoning attack distinction. A method is proposed that identifies client types via plaintext-based gradient history and applies privacy protection and secure aggregation exclusively to filtered legitimate clients, thereby enhancing detection effectiveness while maintaining data confidentiality. Considering the prevalence of sign flipping, noise injection, and label flipping attacks, along with significant differences in objectives, intensity, and behavioral patterns, short- and long-term local gradient history is incorporated. Analyzing anomalous differences across gradient histories enables effective detection of multiple poisoning attack types. A periodic poisoning attack detection strategy is designed to ensure client privacy. Furthermore, addressing the fixed chain count and clients in multi-chain aggregation, an adaptive multi-chain secure aggregation method is developed to enhance adaptability to dynamic client set changes, improving aggregation efficiency while preserving privacy. Experimental results on the MNIST, Fashion-MNIST, and CIFAR10 datasets demonstrate that the proposed algorithm achieves an average accuracy of 85.18%, an improvement of approximately 6% over baseline algorithms. The method exhibits robust detection capabilities for multiple poisoning attacks, effectively enhancing model performance and meeting defense requirements in complex attack scenarios.**

## Keywords

**Federated Learning, Poisoning Attack, Privacy Protection, Gradient History, Attack Detection**



## 1. 引言

随着人工智能、大数据、物联网等前沿技术的快速发展，电力、医疗等行业的大数据分析处理成为各领域研究热点。胸部 X 光片是结核病的基础诊断手段，卷积神经网络(Convolutional Neural Network, CNN)通过对大量胸部 X 光片进行特征提取、模型训练等，能够更精准地识别病灶的微小特征，对结核病的早期发现具有重要意义[1]。通常集中式的数据分析方法将各类行业数据上传到云服务器或数据处理中心，而电力、医疗等行业数据通常包含较多敏感信息，如医疗数据中患者的身份信息、患病信息、社会关系等。敏感数据一旦被窃取或操控，将会带来严重的安全威胁。美国联合健康集团首次确认，2024 年前 9 个月中超 1 亿人的个人信息和医疗数据被盗，预计损失已增加至 24.5 亿美元[2]。鉴于联邦学习在数据隐私保护、分布式计算方面的优越性[3]，被广泛应用于电力、医疗等行业的大数据分析处理，通过只传递参数来更新全局模型，在一定程度上解决了客户端数据的隐私泄露问题。然而客户端发送的模型参数信息依然可推理出其本地原始数据的相关信息[4]，联邦学习中的隐私保护还需进一步深度研究。同时由于联邦学习中客户端独立于服务器进行训练以及服务器无法访问客户端数据的特点，使得联邦学习容易受到客户端发起的投毒攻击[5]，极大降低全局模型的可用性甚至破坏全局模型。

随着网络攻击类型的不断变化、升级，实际行业领域中联邦学习框架的部署环境多为非完全可信环

境，隐私保护必不可少。现有隐私保护的联邦学习方法通常采用差分隐私、同态加密或安全多方计算等技术对模型梯度进行加密或扰动以防止“诚实但好奇”的服务器窃取客户端数据。同时非完全可信的联邦学习环境中，恶意客户端可以通过恶意修改梯度参数来发起投毒攻击，由于数据经过加密或扰动，服务器对原始梯度的可观测性显著降低，这使得依赖梯度信息进行异常检测的投毒防御方法难以发挥作用。同时不同类型的投毒攻击在攻击目标、强度、行为模式上差异显著。如标签翻转攻击可能导致模型在特定类别上出现错误分类，噪声注入则是破坏全局模型收敛。不同攻击者的模型更新往往呈现出多样化的差异性，需要从多个维度进行联合建模和识别。现有检测方法多使用单一方法进行识别，难以在统一框架下同时识别多种攻击行为。综上所述，隐私保护机制限制了模型更新的可区分性，使得投毒攻击检测方法难以识别；同时多类投毒攻击的混合存在要求检测方法具备更强的泛化能力和鲁棒性。如何在非完全可信的联邦学习环境中，综合考虑隐私保护与抗投毒攻击，对多类型混合投毒攻击环境中的投毒攻击进行有效识别面临极大挑战。

面向非完全可信的联邦学习环境，针对多类型混合投毒攻击的有效识别和隐私保护问题，本文提出PP-MPAD (Privacy-Preserving Multi-Type Poisoning Attack Detection，隐私保护且抗投毒攻击的联邦学习方法)。考虑现有方法中加密梯度无法准确识别投毒攻击类型，PP-MPAD 基于明文的客户端梯度历史信息，通过比较不同梯度历史间的异常差异对客户端类型进行精准识别，以抵抗多类型投毒攻击。考虑服务器可通过客户端上传的明文梯度长短历史信息间接推断出客户端原始梯度，PP-MPAD 提出周期性投毒攻击检测策略以有效减少隐私泄露的可能性。同时考虑基于链式的安全多方计算方法的简单部署和无损计算，提出自适应多链安全聚合方法。本文的主要贡献如下：

(1) 针对不同投毒攻击在模型更新中呈现出多样化的异常特征，提出基于梯度历史的多类型投毒攻击检测算法以提升混合攻击环境下检测的鲁棒性。算法引入局部梯度的长短历史信息，通过比较不同梯度历史之间的余弦相似度、四分位距等差异性，提出基于梯度短历史的检测、基于梯度幅值的四分位距统计检测和基于多数投票的检测，分别实现对符号翻转攻击、噪声注入攻击以及标签翻转攻击的有效检测，提高了全局模型的性能。

(2) 针对服务器通过明文梯度历史间接推断出客户端原始梯度以及梯度直接传输引起的隐私泄露问题，设计周期性投毒攻击检测策略和自适应多链安全聚合方法。其中周期性投毒攻击检测方法仅在检测回合识别客户端类型，使服务器只能获取部分梯度之和，防止原始梯度的间接泄露。自适应多链安全聚合方法则根据实变的客户端集合将正常客户端动态分配在多条链中以在隐私保护的同时提升聚合效率。

## 2. 相关工作

大数据、机器学习等技术的快速发展，使得联邦学习在通信瓶颈、系统与数据异质性、隐私保护以及投毒攻击等方面成为研究热点[6]。考虑实际机器学习应用场景(如医疗系统中的图像识别)中的数据高敏感性以及模型训练过程中的数据可靠性，隐私保护和投毒攻击成为联邦学习框架中各类机器学习应用不可忽略的问题。考虑现有联邦学习中隐私保护和投毒攻击的相关文献，可将其分为抗投毒攻击的联邦学习、隐私保护的联邦学习和隐私保护及抗投毒攻击的联邦学习。

(1) 抗投毒攻击的联邦学习

现有联邦学习中的抗投毒攻击研究主要通过鲁棒聚合[7]-[9]、相似度分析[10]-[12]以及异常检测[13][14]等抵御各类投毒攻击以保障机器学习应用的全阶段数据可靠性。基于鲁棒聚合的方法旨在通过改进联邦学习中的聚合算法来缓解投毒攻击的负面影响。Benjamin 等人通过一致性比率诱导符号选举模块来确定每个参数梯度的共识方向，利用鲁棒坐标级聚合策略仅聚合与选举符号方向一致的方差缩减稀疏梯

度，从而有效抵御联邦学习中的投毒攻击[9]。基于相似度分析的方法通过比较模型梯度间的相似性来识别并过滤恶意数据。Yaldiz 等人提出基于余弦相似性的方法，通过计算全局模型与客户端权重间的余弦相似性来识别并过滤低于平均值的恶意客户端以抵御投毒攻击[12]。基于异常检测的方法旨在通过挖掘梯度更新中的异常行为特征检测并去除可疑客户端。蒋伟进等人提出基于可解释贡献异常检测与动态裁剪的抗投毒攻击方法，通过结合 SHAP 值和局部异常因子来量化模型参数对预测行为的贡献，从而识别恶意客户端以保留对全局模型有益的参数[14]。上述方法聚焦于恶意投毒识别，难以区分具体的投毒攻击类型，无法应对多类型混合投毒攻击。Gupta 等人提出基于梯度历史的客户端类型检测算法，引入客户端梯度的长短历史信息来计算梯度中值，从而识别不同梯度历史的差异特征以检测多种类型的投毒攻击[15]。现实模型应用中，数据的非独立同分布特性导致基于梯度中值和聚类的方法不能准确识别各类投毒攻击。

(2) 隐私保护的联邦学习

现有联邦学习中的隐私保护研究主要采用同态加密[16]-[18]、差分隐私[19]-[21]以及安全多方计算[22]-[24]等技术防止模型梯度泄露，以保障客户端数据的机密性。基于同态加密的方法旨在允许服务器在密文上进行计算，从而在不泄露原始数据的前提下完成模型训练。He 等人对 Paillier 算法进行改进，通过预计算随机中间值来减少加密时间从而在保证安全性的前提下简化加密流程，提升协议效率[16]。该技术已被广泛应用于医疗等高度敏感领域，以保护患者的数据隐私[17]。基于差分隐私的方法旨在梯度传输中加入校准噪声，模糊化个体数据贡献以实现隐私保护。Wu 等人提出自适应的差分隐私联邦学习方法，融合自适应的梯度下降算法和高斯机制，在实现差分隐私的同时提高模型的泛化能力[19]。Li 等人提出融合聚类采样与记忆机制的差分隐私方法，通过聚类采样缓解数据异质性的影响，并利用记忆机制保留历史模型信息，减弱噪声对全局模型的干扰[21]。基于安全多方计算的方法通过设计协同计算协议，使得服务器无需收集客户端梯度完成模型训练以实现隐私保护。Li 等人提出基于链式安全多方计算的隐私保护联邦学习框架，将客户端按链式结构组织，由链尾客户端上传聚合结果，服务器去除掩码后完成全局模型更新，在实现隐私保护的同时保障模型准确性[22]。上述方法主要对模型梯度数据进行隐私保护，无法有效应对因其他辅助信息(如梯度历史、数据分布等)产生的间接性隐私泄露风险。

(3) 隐私保护及抗投毒攻击的联邦学习

现有联邦学习中综合隐私保护与抗投毒攻击的研究除对梯度信息进行隐私保护外，还对加密或扰动后的数据进行相似度分析或异常检测，以保障梯度信息的机密性与全局模型的可靠性，包括有同态加密与抗投毒结合[25]-[27]、差分隐私与抗投毒结合[28] [29]以及安全多方计算与抗投毒结合[30]-[32]等。Chen 等人采用同态加密保护客户端数据，并在密文域中计算余弦相似度以调整聚合权重，从而削弱恶意客户端的影响[25]。Liu 等人则在密文域中利用皮尔逊相关系数比较客户端梯度间的相似性，以检测潜在投毒行为[26]。Le 等人通过自适应差分隐私方法减轻噪声对模型性能的影响，并在噪声扰动的梯度上进行 DBSCAN 聚类来实现恶意客户端识别[28]。高鸿峰等人提出基于多方计算的安全拜占庭弹性联邦学习方案，使用加性秘密共享技术保护梯度隐私，并在密态数据上基于余弦相似度来构建有毒数据筛选机制实现对投毒攻击的防御[31]。上述方法聚焦于采用单一方法对加密或扰动后的数据计算，无法有效应对多类型混合投毒攻击。

综上所述，现有联邦学习中隐私保护和抗投毒攻击的相关文献要么仅关注隐私保护，要么仅关注抗投毒攻击，无法同时兼顾隐私保护与抗投毒攻击对模型应用的安全威胁。尽管部分文献中同时考虑隐私保护和抗投毒攻击，其均在隐私保护基础上采用单一方法用于多类型投毒攻击。然而混合投毒攻击环境下，单一相似度等方法无法准确检测相应的投毒类型。如何同时考虑隐私保护与抗投毒攻击，对多类型混合投毒攻击环境中的投毒攻击进行有效识别是当前亟需解决的问题。本文提出新型的隐私保护且抗投

毒攻击联邦学习方法 PP-MPAD，旨在保障隐私的同时实现对多类投毒攻击的鲁棒防御。

## 3. 系统模型

考虑非完全可信环境(诚实但好奇的服务器与恶意客户端共存)的联邦学习，本文提出 PP-MPAD 以解决多类投毒攻击的鲁棒防御和隐私保护。服务器负责聚合与分发全局模型、计算全局梯度短历史及执行客户端类型检测。同时诚实但好奇的服务器作为半信任的第三方，虽然遵循聚合协议，但仍然会主动实施攻击以窃取或泄露客户端的隐私信息。客户端负责训练本地模型及计算局部梯度长短历史，包括有正常客户端与恶意客户端。其中正常客户端诚实参与训练；恶意客户端则可能发起投毒攻击，主要包括非定向攻击与定向攻击。鉴于非定向攻击中符号翻转和噪声注入操作的简便及普遍性，定向攻击中标签翻转的隐蔽性与目标性，本文重点考虑上述三类投毒攻击。假设每个恶意客户端仅能操控自身数据或模型，无法干预其他客户端或服务器聚合过程，且在每一通信回合中均持续实施攻击行为。本文中通信回合是指客户端和服务器之间完成一次完整的参数交换与聚合的过程，包括模型分发、本地训练、上传更新和模型聚合。设 N 为参与联邦学习模型训练过程的客户端总数，其中包含 m 个恶意客户端，K = N − m 个正常客户端，并假设恶意客户端的比例始终小于 50%。

本文所提 PP-MPAD 的系统模型如图 1 所示。为防止服务器通过客户端上传的梯度长短历史信息间接推断出客户端原始梯度，PP-MPAD 按照梯度短历史滑动窗口将通信回合划分为检测回合和标准通信回合，从而使得服务器只能获取一部分梯度之和来保护客户端梯度隐私。其中梯度短历史是指最近若干通信回合中梯度的滑动平均值，若梯度短历史设置为最近三个通信回合中梯度的平均值，那么梯度短历史的滑动窗口即为三。梯度长历史指历史通信回合中梯度的累加值，即当前通信回合前的所有通信回合中的梯度之和。本文根据梯度短历史的滑动窗口来设置当前通信回合是否为检测回合，如当滑动窗口为三时，设置每三个标准通信回合后进行一次检测回合。本文中定义检测回合是联邦学习过程中周期性触发的、具备投毒攻击识别功能的特殊通信回合。其不仅包含标准通信回合的模型更新与聚合流程，还引入了客户端梯度历史的上传与服务器端的恶意客户端检测两个关键步骤。标准通信回合为检测回合之外的其他通信回合。考虑现有隐私保护且抗投毒攻击的联邦学习方案中，先对客户端梯度进行加密或扰动再在密文域中执行投毒攻击检测操作，容易模糊或消除恶意梯度所具有的差异性特征，导致检测算法难以准确区分不同类型的投毒攻击。本文采用基于明文的梯度历史信息对客户端类型进行识别，再对筛选出的正常客户端进行隐私保护和安全聚合操作，从而在保障数据机密性的同时提升检测的有效性。

鉴于本文的隐私保护和多类型抗投毒攻击研究集中在检测回合，图 1 给出典型的检测回合示例。在抗投毒攻击阶段，服务器向所有客户端分发当前的全局模型参数。同时 N 个客户端在完成本地训练后，同步更新并上传其梯度长短历史信息。服务器采用多类型投毒攻击检测算法进行客户端类型识别。依据 Gupta 等人“由显著到隐蔽”的检测原则，PP-MPAD 按符号翻转攻击、噪声注入攻击、标签翻转攻击先后顺序识别不同攻击[15]。通过多类型抗投毒攻击检测后，系统将识别出的 m 个恶意客户端予以排除，从而获得 K 个可信的正常客户端集合。在标准通信回合中，系统将跳过客户端类型识别步骤，直接沿用上一检测回合所确定的正常客户端集合进行模型聚合。在模型聚合阶段，采用自适应多链安全聚合方法来保障正常客户端的数据隐私，主要将正常客户端动态划分至多条链路中，并在各链路上并行执行基于链式的安全多方计算[22]。该方法在每一通信回合中根据正常客户端数量的平方根来设置链路的数量，并通过发牌算法将客户端随机分配至各条链路中以确保各链路中客户端数量相当。在每条链路中，服务器将随机向量掩码发送给起始客户端，客户端通过掩码累加机制实现本地梯度的加密聚合，链尾客户端将聚合结果返回至服务器。每条链路中所分配客户端在链路中先后顺序的设置方式与文献[22]相同。服务器将每条链路的聚合结果减去随机向量即可获取每条链的最终聚合结果，再将各链的最终聚合结果进行加

权平均计算即可更新全局模型，同时保存全局梯度的短历史信息。上述过程持续迭代，直至达到预设的通信回合数或模型收敛条件。

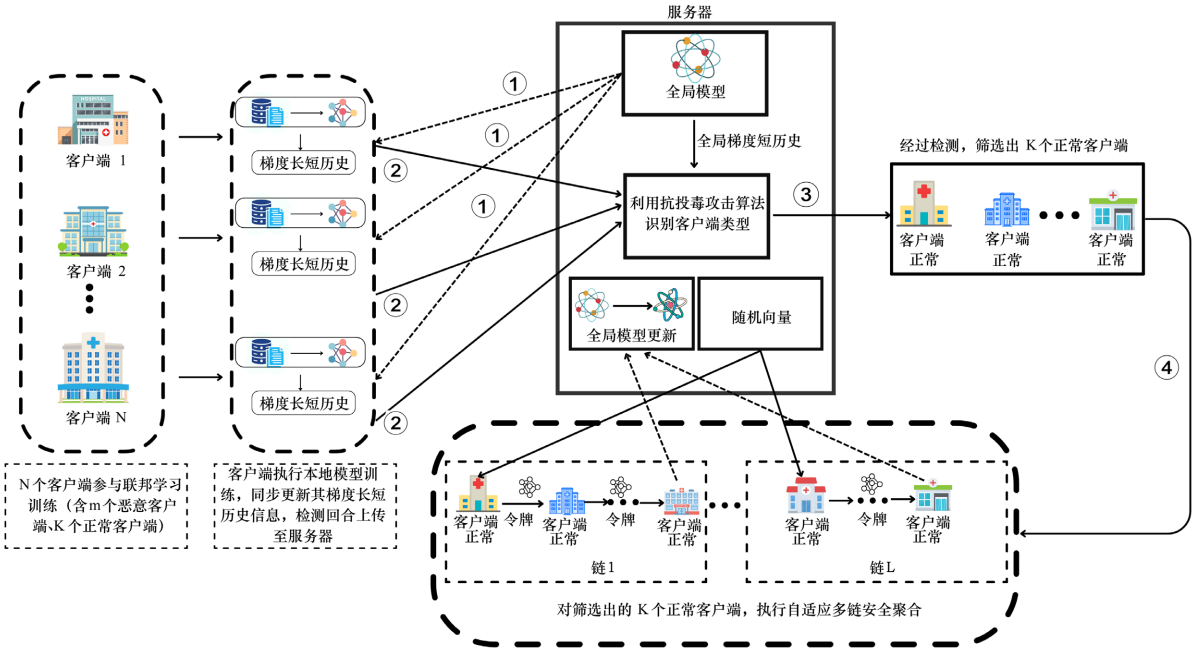



**Figure 1.** System model of PP-MPAD
**图 1.** PP-MPAD 的系统模型

## 4. PP-MPAD 方法

针对现有联邦学习同时考虑隐私保护与投毒攻击检测的相关研究中，因加密后恶意梯度差异特征模糊化所导致的投毒攻击检测不准确问题，本文提出 PP-MPAD 策略，基于明文的梯度历史信息对客户端类型进行识别，再对筛选出的正常客户端进行隐私保护和安全聚合操作，以兼顾隐私保护与投毒攻击检测性能。考虑服务器可通过客户端上传的明文梯度长短历史信息间接推断出客户端原始梯度的隐私泄露风险，考虑服务器可通过客户端上传的明文梯度长短历史信息间接推断出客户端原始梯度的隐私泄露风险，PP-MPAD 提出周期性投毒攻击检测策略，仅在检测回合识别客户端类型，使服务器只能获取部分梯度之和，即使通过差分梯度长短历史也无法恢复局部梯度的具体数值，从而有效减少隐私泄露的可能性。考虑基于链式的安全多方计算方法的简单部署和无损计算，结合现有多链安全聚合方法中固定链数和链内客户端对动态场景的不适用性，提出自适应多链安全聚合方法，根据实际参与客户端数量动态调整链的数量，从而使服务器在客户端集合发生变化时仍可通过多链聚合提升聚合效率。PP-MPAD 提出在投毒攻击检测过程中，依据 Gupta 等人“由显著到隐蔽”的检测原则[15]，采用全局梯度短历史与客户端梯度短历史的余弦相似度识别符号翻转攻击，利用基于四分位距(IQR, Interquartile Range)的统计方法检测梯度幅值异常的噪声注入攻击，在准确识别并剔除非定向攻击后，通过基于多数投票的标签翻转攻击检测方法识别定向攻击者。

### 4.1. PP-MPAD 中的抗投毒攻击

现有多类型投毒攻击研究中，多采用单一方法进行投毒攻击识别。由于不同投毒攻击在目标、强度和行为模式上差异显著，现有方法难以在统一框架下同时准确识别多种攻击行为。PP-MPAD 采用不同方

法按符号翻转攻击、噪声注入攻击、标签翻转攻击先后顺序识别不同攻击。

#### 4.1.1. 符号翻转攻击

现有投毒攻击研究中，多采用余弦相似度、欧式距离等对梯度进行计算判断是否存在投毒攻击，缺乏对隐私保护的考虑。鉴于符号翻转攻击对全局模型聚合的重要影响，MUD-HoG [15]以客户端梯度短历史的中值为参照，通过余弦相似度判别攻击者，但是数据非独立同分布特性使得梯度短历史中值易受干扰，导致误判。PP-MPAD 提出基于梯度短历史的检测方法，主要采用余弦相似度对全局梯度短历史与客户端梯度短历史进行计算，从而避免客户端梯度短历史中值的误差导致误判。全局梯度短历史即若干通信回合内全局梯度的平均值，以平滑单次更新的随机波动并增强对短期异常的敏感性。若函数 $d_{cos}(\cdot)$ 计算余弦相似度，$\nabla_i^{sHoG}$ 表示客户端 $c_i$ 的梯度短历史，$\nabla_{global}^{HoG}$ 表示全局梯度短历史，其中，$\nabla_i^{sHoG}$ 与 $\nabla_{global}^{HoG}$ 在通信回合 $t$ 处的计算方式如下式(1)和(2)。

$$\nabla_i^{sHoG} = \frac{1}{l}\sum_{t=t-l}^{t-1}\nabla_{t,i} \tag{1}$$

$$\nabla_{global}^{HoG} = \frac{1}{l}\sum_{t=t-l}^{t-1}\nabla_t \tag{2}$$

其中 $l$ 是滑动窗口大小，$\nabla_{t,i}$ 表示客户端 $c_i$ 在通信回合 $t$ 处的梯度，$\nabla_t$ 表示在通信回合 $t$ 处的全局梯度。若客户端 $c_i$ 满足式(3)时，则将其标记为符号翻转攻击者：

$$d_{cos}\left(\nabla_{global}^{HoG}, \nabla_i^{sHoG}\right) < 0 \tag{3}$$

由于全局梯度能够更稳定地反映整体更新方向，通过计算客户端梯度短历史与全局梯度短历史的余弦相似度，即可在复杂攻击场景下有效识别符号翻转攻击者，从而提升检测的准确性。

#### 4.1.2. 噪声注入攻击

噪声注入攻击通过向梯度中添加大幅度随机噪声干扰模型收敛，其攻击特征表现为梯度幅值异常偏大。基于恶意客户端的比例始终小于 50%的假设，现有噪声注入攻击检测中根据局部梯度短历史对客户端进行聚类，并采用欧式距离对小类别(客户端数量较少)的客户端梯度短历史和大类别(客户端数量较多)的客户端梯度短历史中值进行计算来判断噪声注入攻击，但是梯度短历史中值不具稳健性，容易导致误判。考虑梯度幅值异常，PP-MPAD 提出基于梯度幅值的四分位距统计检测方法，即正常客户端的梯度幅值服从一定的统计分布，而噪声注入攻击者的梯度幅值显著偏离上述分布。

前述已检测出符号翻转攻击的客户端被移除后，其余客户端计算局部梯度短历史的 L2 范数(即梯度幅值)，将所有计算值按照升序排序，并计算其四分位距值。四分位距是样本数据的第三四分位数($Q3$)与第一四分位数($Q1$)之差 $IQR = Q3 - Q1$，通常在异常值识别中，将低于 $Q1 - 1.5 \cdot IQR$ 或高于 $Q3 + 1.5 \cdot IQR$ 的样本视为异常。由于噪声注入攻击的典型特点为梯度幅值异常增大，此处采用上界作为判别阈值。若 $\left\|\nabla_i^{sHoG}\right\|_2$ 表示客户端 $c_i$ 的梯度短历史的 L2 范数，则客户端 $c_i$ 满足式(4)时，则将其标记为噪声注入攻击者。

$$\left\|\nabla_i^{sHoG}\right\|_2 > Q3 + 1.5 \cdot IQR \tag{4}$$

#### 4.1.3. 标签翻转攻击

现有标签翻转攻击的防御方法中，通常将梯度历史空间中的较小簇识别为攻击者。然而非独立同分布数据场景下，正常客户端的梯度本身就存在显著差异，导致聚类结果不稳定，严重影响检测的可靠性。考虑标签翻转攻击具有长期性与全局性，结合客户端梯度长历史，提出基于多数投票的标签翻转攻击检测方法。标签翻转攻击的核心目标是篡改特定类别的决策边界，这种扰动在模型的深层网络参数中(尤其

是最后两层)体现得最为显著。本文提取客户端梯度长历史的最后两层参数用于标签翻转攻击检测。借鉴联邦平均算法的思路，各客户端 $c_i$ 梯度长历史与其他所有客户端梯度长历史的余弦相似度之和作为每个客户端 $c_i$ 的权重，所有客户端权重与梯度长历史加权平均后得到全局梯度参考值以标识稳健的参考基准，识别并排除恶意偏差。基于多数投票共识算法，计算各客户端梯度长历史与全局梯度参考值间的余弦相似度，将少数余弦相似度符号不一致的客户端标记为标签翻转攻击者并将其移除客户端集合。考虑标签翻转攻击客户端与全局梯度参考值间的余弦相似度远小于正常客户端与全局梯度参考值间的余弦相似度，故通过对计算出的余弦相似度进行升序排序，获取最大间隙来识别相对隐蔽的标签翻转攻击威胁。若 $H\left(\nabla_i^{lHoG}\right)$ 表示客户端 $c_i$ 的梯度长历史的最后两层参数，$\nabla_i^{lHoG}$ 表示客户端 $c_i$ 的梯度长历史，$\nabla_{base}^{lHoG}$ 表示对应的全局梯度参考值。其中 $\nabla_i^{lHoG}$ 在通信回合 $t$ 处的计算方式如下式(5)。

$$\nabla_i^{lHoG}=\sum_{t=1}^{t-1}\nabla_{t,i} \tag{5}$$

若剩余客户端 $c_i$ 与全局梯度参考值的余弦相似度满足式(6)，则将其标记为标签翻转攻击者：

$$d_{cos}\left(H\left(\nabla_i^{lHoG}\right),\nabla_{base}^{lHoG}\right)<d_{\phi} \tag{6}$$

其中判别阈值 $d_{\phi}$ 为最大间隙相关的两个余弦相似度的中点值。

### 4.2. PP-MPAD 中的隐私保护方法

现有隐私保护方法研究中，基于同态加密的方法因加解密过程复杂，难以在实际系统中高效部署；基于差分隐私的方法虽能缓解隐私泄露，其引入的噪声显著影响模型性能。考虑基于链式的安全多方计算方法的简单部署和无损计算[22]，PP-MPAD 提出自适应多链安全聚合方法改进现有多链安全聚合方法中的固定链数量和链内客户端，增强对动态场景的适用性。

PP-MPAD 中的隐私保护方法主要根据实际参与客户端数量来动态调整链的数量，从而使服务器在客户端集合发生变化时仍可通过多链聚合提升聚合效率。为最大化客户端并行度及各链中的计算效率，按照客户端数量的平方根调整各回合的链数。在确定链数 $L$ 后，服务器按洗牌算法将所有正常客户端随机划分至各链路中进行单链聚合以确保每条链路的客户端数量相当。假设第 $r$ 个通信回合中识别出的正常客户端数量为 $n$，当客户端数量较少小于阈值 $q$ 时，仅分配一条链进行聚合即可，则当前回合中链数量 $L_r$ 的计算方式如下式(7)所示：

$$L_r=\begin{cases}1, & \text{if } n\le q\\ \max\left(2,\left\lfloor\sqrt{n}\right\rfloor\right), & \text{if } n>q\end{cases} \tag{7}$$

阈值 $q$ 用于区分单链与多链。当客户端数量较少(如仅有 3 个客户端)时采用单链以避免开销，当客户端数量较多时，系统至少分配 2 条链，并最多不超过 $\sqrt{n}$ 条链，以保证链长适中并提升聚合效率。通过上述方法自适应得到各通信回合的链数量和链内客户端后，每条链中的客户端进行链内聚合。服务器利用伪随机生成器生成一个与梯度维度一致的随机向量作为初始掩码，并将其封装为令牌发送至各条链的起始客户端。起始客户端在接收令牌后，将该令牌与自身的本地梯度更新相加，形成加密后的梯度更新，并生成新的令牌传递给下一个客户端。链上的其余客户端依次将自身梯度累加至接收到的令牌中，从而构建出链式累加结构。最终，每条链的最后一个客户端将累加结果返回服务器。所有链并行执行上述过程，其中每条链路中所分配客户端的先后顺序的设置方法与文献[22]相同。服务器将每条链路的聚合结果减去随机向量即可获取每条链的最终聚合结果，再将各链的最终聚合结果进行加权平均计算即可更新全局模型，通过自适应多链安全聚合方法，服务器可在不获取任何客户端原始梯度的前提下，通过掩码累

加机制恢复所有正常客户端的梯度总和，并据此更新全局模型。考虑实际网络环境中客户端连接的不稳定性，客户端掉线可能发生于多链构建或聚合计算的任意阶段。为避免掉线客户端对整条链路产生影响，根据 TCP 三次握手机制，超时应答的掉线客户端数据不再收集。从收敛性角度分析，节点的随机掉线等效于对客户端集合的随机采样。根据随机梯度下降理论，若掉线事件独立于数据分布，剩余存活节点的聚合梯度仍为全局梯度的无偏估计[33]。尽管参与节点的减少会轻微增加梯度估计的方差并对收敛速率产生线性影响，但不会改变算法的敛散性，从而保证了模型在动态网络下的鲁棒性。相比传统方法，自适应多链结构不仅显著提升了聚合效率，还有效克服了传统多链聚合方法难以适应动态客户端集合的局限性。

## 5. 实验结果与分析

本文实验环境为 Python 3.8，深度学习框架为 PyTorch 1.12.1 和 PaddlePaddle 2.3.2，硬件配置为 6 核 CPU、16 GB 系统内存以及 12 GB 的 NVIDIA GPU 显存，操作系统为 Ubuntu 20.04。为验证 PP-MPAD 算法的有效性，本文在 MNIST，Fashion-MNIST 以及 CIAFR10 数据集上开展实验。

同时为确保实验的公平性，在 MNIST 和 Fashion-MNIST 数据集中进行实验时，其实验设置(包括数据集划分、模型结构、超参数设置、恶意客户端比例等)均与基准算法 MUD-HoG [15]保持一致。在攻击方式设置中，本文考虑定向攻击类型为多标签翻转攻击，攻击者将数据集中源标签“1”，“2”以及“3”翻转为目标标签“7”。实验中的非定向攻击设置保持一致，恶意客户端比例从 12.5%逐步增加至 47.5%。在 CIAFR10 数据集中进行实验时，使用 3 层卷积和 3 层全连接的卷积网络结构，为了确保收敛，设置通信回合数为 85。

本文采用准确率(Accuracy)、精度(Precision)和召回率(Recall)三个指标对算法进行综合评价。Accuracy、Precision 以及 Recall 的计算分别如式(8)~(10)所示。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

其中，$TP$ 表示真正类数量，$TN$ 表示真负类数量，$FP$ 表示假正类数量，$FN$ 表示假负类数量。

通过比较全局模型的准确率评估算法的抗投毒能力，准确率越高，说明算法的鲁棒性越强。对于定向攻击的防御效果，则通过目标标签的精度和源标签的召回率进行衡量，二者数值越高，说明算法在抵御定向攻击方面的能力越强。

### 5.1. 消融实验

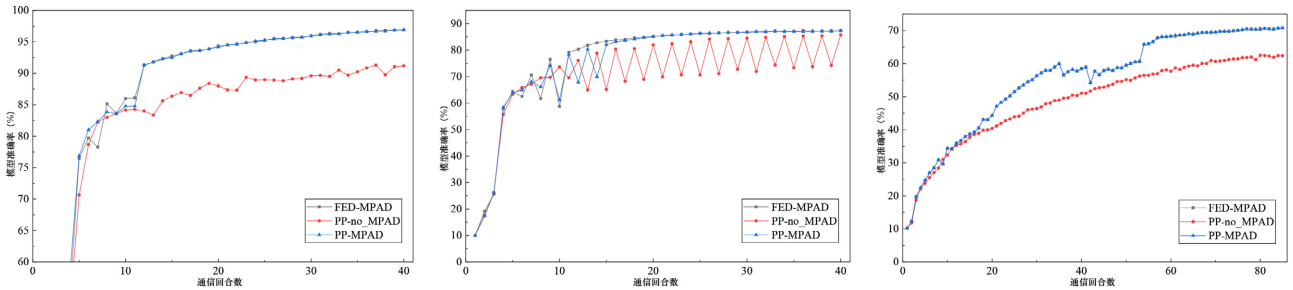
为验证 PP-MPAD 算法在多类投毒攻击检测方面的有效性与隐私保护方法在模型效用方面的优势，本文针对投毒攻击检测模块和隐私保护模块开展了消融实验。各变体方法如表 1 所示，其中“√”表示该方法能检测对应的攻击类型。如变体方法 FED-MPAD 表示可检测多类投毒攻击，但不采用自适应多链安全聚合。

当恶意客户端占比为 42.5%时进行实验，图 2 展示了 PP-MPAD 及其表 1 中 PP-no_MPAD、FED-MPAD 的准确率变化曲线，旨在验证攻击检测模块的必要性与自适应多链安全聚合方法的优越性。从图中可以清晰地看出，在三个数据集上，缺乏攻击检测机制的 PP-no_MPAD 模型准确率最差，其中在

Fashion-MNIST 数据集上其准确率曲线甚至呈现剧烈振荡且无法有效收敛，这证实了投毒攻击会严重降低模型性能，缺乏防御机制将导致训练模型不可靠。与此同时，PP-MPAD 的性能始终近似于缺乏隐私保护机制的 FED-MPAD，这充分表明本文采用的自适应多链安全聚合方法在实现隐私保护的同时，几乎未引入额外的性能损失。

**Table 1.** Experimental results of PP-MPAD variants ignoring different components
**表 1.** PP-MPAD 算法省略不同模块的变体方法实验结果

| 变体方法 | 符号翻转攻击 | 噪声注入攻击 | 定向攻击 | 自适应多链安全聚合 |
|---|---|---|---|---|
| PP-MPAD | √ | √ | √ | √ |
| PP_no_MPAD | | | | √ |
| FED-MPAD | √ | √ | √ | |




(a) MINIST 数据集　(b) Fashion-MINIST 数据集　(c) CIFAR10 数据集

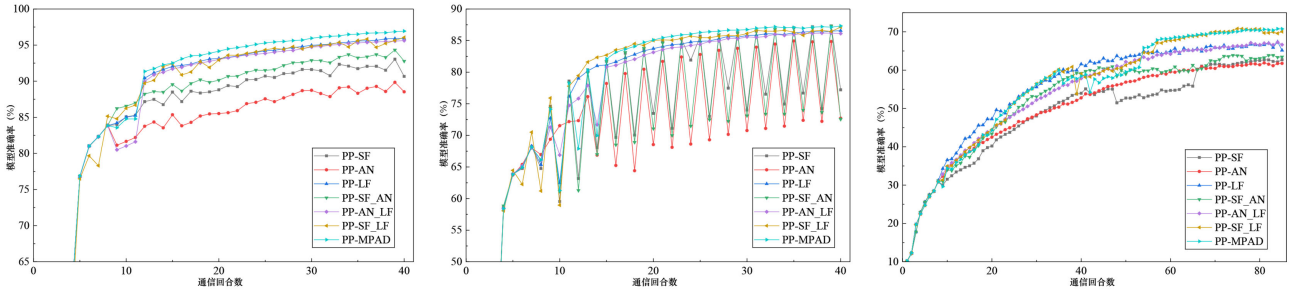
**Figure 2.** Accuracy curves of different variant methods on various datasets in Table 1
**图 2.** 在不同数据集上表 1 中各变体方法的准确率曲线

**Table 2.** Experimental results of PP-MPAD variants replacing various components
**表 2.** PP-MPAD 算法替换不同模块的变体方法实验结果

| 变体方法 | 符号翻转攻击 | 噪声注入攻击 | 定向攻击 |
|---|---|---|---|
| PP-SF | √ | | |
| PP-AN | | √ | |
| PP-LF | | | √ |
| PP-SF_AN | √ | √ | |
| PP-AN_LF | | √ | √ |
| PP-SF_LF | √ | | √ |

为进一步探究 PP-MPAD 投毒攻击检测模块的内部贡献，各变体方法如表 2 所示，其中“√”表示该方法能检测对应的攻击类型。如变体方法 PP-SF 仅检测符号翻转攻击，隐私保护方法与 PP-MPAD 保持一致均使用自适应多链安全聚合。图 3 展示了 PP-MPAD 与表 2 所示变体的性能对比。结果显示，在不同数据集上，PP-MPAD 算法的准确率均优于所有变体方法，这充分验证了其全方位、多维度防御策略的必要性与有效性。与仅检测单一攻击类型的变体方法(PP-SF、PP-AN、PP-LF)相比，PP-MPAD 在准确率方面优势明显。其中，PP-AN 性能最差，原因在于其无法识别符号翻转与标签翻转攻击者，导致模型收敛方向受到严重干扰，整体性能下降。PP-SF 的性能优于 PP-AN，说明抑制符号翻转攻击在一定程度上能够修正模型方向。在单类检测变体中，PP-LF 表现最佳，这是因为实验设定中标签翻转攻击者数量

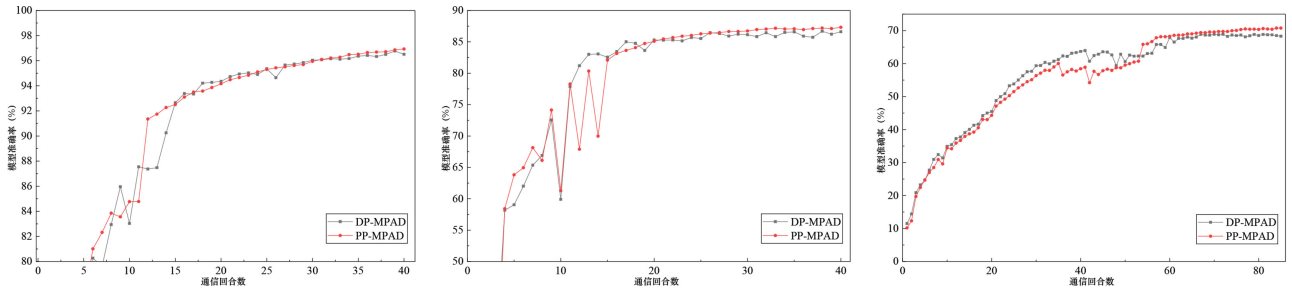
最多，对模型收敛方向的干扰最为显著。与进行双类攻击检测的变体方法(PP-SF_AN、PP-AN_LF 及 PP-SF_LF)相比，PP-MPAD 依然保持优势。其中，PP-SF_AN 因漏检数量较多的标签翻转攻击者而性能最差，PP-SF_LF 略优于 PP-AN_LF，进一步验证了抑制符号翻转攻击能够更好地修正模型方向。总体而言，在复杂的混合攻击环境下，随着检测覆盖面的扩大，模型性能显著提升，充分体现了全面检测与防御的必要性。PP-MPAD 算法通过比较不同梯度历史间的差异化特征，精准识别并过滤多类恶意模型更新，有效保障联邦学习系统在复杂攻击环境下的整体鲁棒性与安全性。




(a) MINIST 数据集　(b) Fashion-MINIST 数据集　(c) CIFAR10 数据集

**Figure 3.** Accuracy curves of different variant methods on various datasets in Table 2
**图 3.** 在不同数据集上表 2 中各变体方法的准确率曲线

为进一步验证 PP-MPAD 在自适应多链安全聚合过程中的无损特性，本文在相同的攻击设定下对比了 PP-MPAD 与变体方法 DP-MPAD 的性能。其中，DP-MPAD 与 PP-MPAD 采用相同的攻击检测算法，但通过向筛选出的正常客户端更新添加拉普拉斯噪声(隐私预算为 0.1)来实现差分隐私保护。图 4 给出在不同数据集上算法的准确率曲线。其实验结果显示，PP-MPAD 的性能优于 DP-MPAD，这主要是由于差分隐私中噪声的引入不可避免地降低了模型效用，而 PP-MPAD 只是引入可消除的随机掩码因此可以实现无损聚合，使得在筛选出正常客户端后能保持更优的模型性能。




(a) MINIST 数据集　(b) Fashion-MINIST 数据集　(c) CIFAR10 数据集

**Figure 4.** Accuracy curves of different method variants on various datasets
**图 4.** 在不同数据集上各变体方法的准确率曲线

## 5.2. 对比实验

本文选取经典的联邦平均算法 FedAvg [3]、基于鲁棒聚合的防御算法 FedSECA [9]、基于余弦相似度的防御算法 CosDefence [12]、基于梯度历史的防御算法 MUD-HoG [15]以及基于聚类的防御算法 DPFLA [32]作为比较算法，与 PP-MPAD 进行对比实验，以全面评估所提方法在抵御多类投毒攻击方面的有效性。

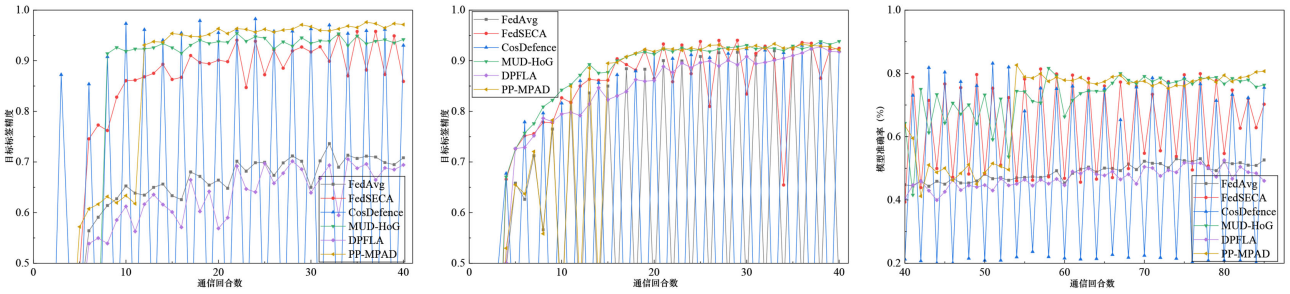
为评估算法在不同攻击强度下的综合防御性能，本文在 Fashion-MNIST 和 MNIST 以及 CIFAR10 数据集上，比较了 PP-MPAD 算法与另外五种算法在不同恶意客户端占比下的模型准确率，实验结果如表 3

所示。随着恶意客户端比例的增加，各算法的模型准确率出现不同程度的下降，充分说明投毒攻击对模型性能具有显著破坏作用。相比之下，PP-MPAD 在恶意客户端占比较高的场景下仍能维持较高的准确率，展现其在不同攻击强度及多样化攻击模式下具有更高的鲁棒性与稳定性。在三个数据集上，PP-MPAD 的模型准确率均优于其他基准算法，其在 MNIST 和 Fashion-MNIST 以及 CIFAR10 数据集上的平均准确率分别达到 96.87%和 87.76%以及 70.91%，相较于次优算法 MUD-HoG 分别提升了 0.47%和 0.43%以及 1.06%。PP-MPAD 的性能优势源于其在多个维度上充分挖掘客户端梯度长短历史信息间的差异化特征，从而能够精准识别并剔除多类型恶意客户端。相比之下，MUD-HoG 虽具备多类投毒攻击检测能力，但在数据非独立同分布和恶意客户端占比较高的场景下，其检测过程中依赖的中值基准不够稳健，聚类方法也易受干扰，导致检测结果出现偏差，整体性能不及 PP-MPAD。CosDefense 仅依赖瞬时余弦相似度作为检测依据，在数据非独立同分布且多类攻击并存的场景下，难以有效区分正常与恶意客户端，易造成大量误判，引发准确率剧烈波动，模型性能不稳定。FedAvg 由于缺乏防御机制，表现出明显的性能波动。FedSECA 通过一致性比率诱导符号选举并缩放客户端更新，在一定程度上抑制了异常值，对非定向攻击具备一定防御能力，整体性能优于 CosDefense 和 FedAvg，尤其是在恶意客户端占比较高时表现更为明显，但对标签翻转等定向攻击的鲁棒性不足，导致算法在混合攻击场景下的防御能力受限。DPFLA 通过聚类算法将较小类中客户端视为攻击者，然而，在多类攻击并存的复杂环境中，单一聚类方法难以全面区分不同攻击行为，往往仅能识别出异常特征最为显著的攻击者，导致其不同数据集中表现不同，但无论在哪种数据集中，其检测效果始终不及 MUD-HoG 和 PP-MPAD。PP-MPAD 采用多维度、针对性检测策略能够在混合攻击场景下更准确全面地识别不同类型的攻击者。

**Table 3.** Accuracy comparison of different algorithms models (%)
**表 3.** 不同算法模型准确率比较(%)

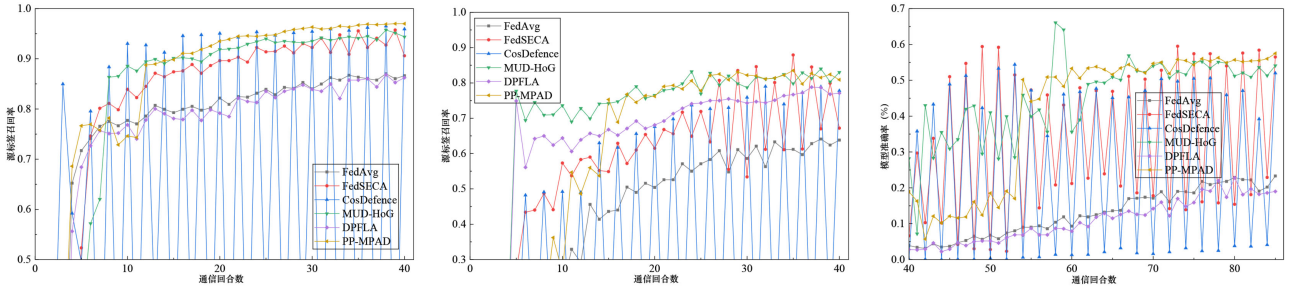
| 数据集 | 算法 | 恶意客户端占比 12.5% | 恶意客户端占比 20.0% | 恶意客户端占比 27.5% | 恶意客户端占比 35.0% | 恶意客户端占比 42.5% | 恶意客户端占比 47.5% |
|---|---|---|---|---|---|---|---|
| MNIST | FedAvg | 96.20 | 95.41 | 94.08 | 93.04 | 90.72 | 88.34 |
| | FedSECA | 96.47 | 96.12 | 96.00 | 95.89 | 94.73 | 94.86 |
| | CosDefense | 86.38 | 86.58 | 76.93 | 79.14 | 86.03 | 77.79 |
| | MUD-HoG | 96.66 | 96.60 | 96.47 | 96.31 | 96.14 | 96.22 |
| | DPFLA | 96.22 | 95.40 | 93.89 | 92.98 | 90.39 | 87.18 |
| | PP-MPAD | 96.92 | 96.76 | 96.79 | 96.83 | 96.93 | 96.96 |
| Fashion-MNIST | FedAvg | 87.41 | 85.41 | 84.79 | 83.21 | 78.08 | 76.75 |
| | FedSECA | 85.80 | 85.78 | 86.11 | 85.63 | 85.20 | 83.39 |
| | CosDefense | 68.33 | 69.29 | 70.82 | 69.84 | 78.46 | 78.22 |
| | MUD-HoG | 87.75 | 87.73 | 87.77 | 87.29 | 87.01 | 86.41 |
| | DPFLA | 87.58 | 87.05 | 87.23 | 86.46 | 85.94 | 85.67 |
| | PP-MPAD | 87.81 | 87.87 | 87.82 | 87.87 | 87.31 | 87.87 |
| CIFAR10 | FedAvg | 69.07 | 67.50 | 65.3 | 64.05 | 62.21 | 61.57 |
| | FedSECA | 67.95 | 66.41 | 65.92 | 66.39 | 66.58 | 63.81 |
| | CosDefense | 52.21 | 53.92 | 53.42 | 59.57 | 59.15 | 59.83 |
| | MUD-HoG | 71.60 | 71.16 | 70.57 | 69.18 | 68.47 | 68.14 |
| | DPFLA | 68.86 | 67.51 | 64.55 | 63.04 | 61.48 | 60.92 |
| | PP-MPAD | 71.32 | 71.65 | 71.32 | 70.67 | 70.65 | 69.87 |

为进一步评估算法在抵御定向攻击方面的有效性，当恶意客户端占比为 42.5%时，比较了目标标签“7”的精度与源标签“2”的召回率，结果分别如图 5 和图 6 所示。实验结果表明，在 MNIST 和 CIFAR10 数据集上，PP-MPAD 的目标标签精度和源标签召回率始终优于其他基准算法，充分展现出其在复杂攻击环境下对定向攻击具备更高的检测准确性与更强的防御能力。MUD-HoG 对于定向攻击的防御效果次于 PP-MPAD，主要由于在数据非独立同分布条件下，其聚类方法稳定性较差，导致定向攻击检测存在偏差。在 Fashion-MNIST 数据集上，PP-MPAD 与 MUD-HoG 的精度与召回率表现相近，均展现出较强的定向攻击防御能力。相比之下，其余四种基准算法在两个数据集上的精度与召回率曲线均波动明显，稳定性较差。其中，CosDefense 在非独立同分布数据下仅依赖瞬时余弦相似度进行判别，难以在混合攻击环境中准确识别定向攻击，误判率较高导致曲线波动最剧烈，难以有效收敛，训练过程缺乏可信度。FedAvg 未引入防御机制，整体表现较差。FedSECA 虽对部分攻击有效，但对定向攻击缺乏鲁棒性，因此在精度和召回率方面表现不佳。DPFLA 算法在不同数据集上表现不同，其在 MNIST 和 CIFAR10 数据集上表现不佳，这是由于聚类算法只检测出来一部分噪声注入攻击者，导致标签翻转攻击对模型有较大影响，而在 Fashion-MNIST 数据集上表现仅次于 MUD-HoG 和 PP-MPAD，这是因为算法检测出来了一部分标签翻转攻击者使得标签翻转攻击者影响降低。这种显著的性能差异揭示出单一聚类算法在多类攻击场景下的局限性，其往往只能识别出异常特征最突出的攻击者，因此在多类攻击并存的情况下，对于定向攻击的检测不具稳定性导致定向攻击防御效果缺乏可靠性。上述实验结果验证了 PP-MPAD 在定向攻击检测方面具备更高的识别准确率与稳定性。




(a) MINIST 数据集　(b) Fashion-MINIST 数据集　(c) CIFAR10 数据集

**Figure 5.** Precision curves of each algorithm for target label “7” across different datasets
**图 5.** 在不同数据集上各算法目标标签“7”的精度曲线

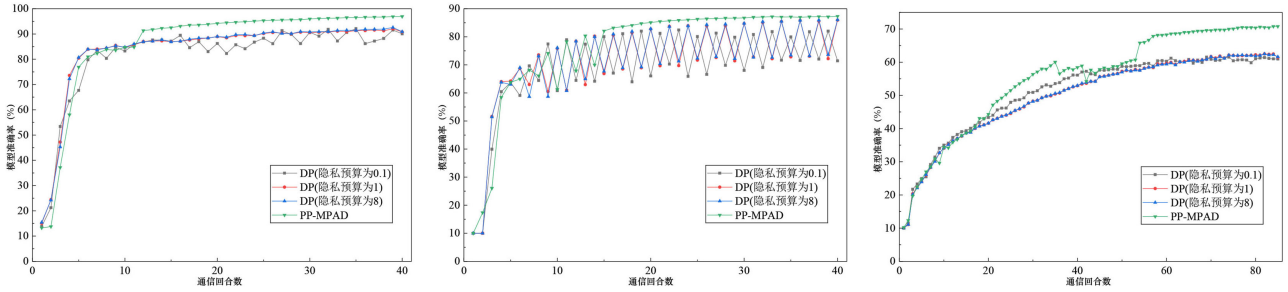



(a) MINIST 数据集　(b) Fashion-MINIST 数据集　(c) CIFAR10 数据集

**Figure 6.** Curves of each algorithm for source label “2” across different datasets
**图 6.** 在不同数据集上各算法源标签“2”的召回率曲线

在投毒攻击的对抗性环境中，仅依赖通用的隐私保护技术(如差分隐私)难以实现有效防御，当恶意客户端占比为 42.5%时，将 PP-MPAD 方案与仅采用差分隐私(Differential Privacy, DP)进行梯度保护的方法

进行比较，其中设置隐私预算分别为 0.1、1 和 8 [22]。图 7 展示了不同算法随通信回合数的模型准确率变化曲线。实验表明，无论隐私预算取值如何，PP-MPAD 的模型准确率均显著高于单纯依赖差分隐私的方法。导致这一结果的原因主要有两点，其一是差分隐私本身不具备识别和抵御恶意梯度更新的能力，无法阻止投毒攻击对模型的破坏，其二是差分隐私中注入的统计噪声会进一步损害模型性能，这在隐私预算由 8 降至 0.1 时模型准确率的下降中得到验证。实验结果充分说明，在复杂的对抗性场景下，必须将针对性的防御机制与隐私保护技术深度融合，才能实现系统的整体鲁棒性。




(a) MINIST 数据集　(b) Fashion-MINIST 数据集　(c) CIFAR10 数据集

**Figure 7.** Accuracy curves of each algorithm across different datasets
**图 7.** 在不同数据集上各算法的准确率曲线

## 6. 结束语

针对非完全可信联邦学习环境中的多类型混合投毒攻击的有效识别和隐私保护问题，本文提出 PP-MPAD 算法。由于不同投毒攻击在目标、强度和行为模式上的差异显著，PP-MPAD 通过比较不同明文梯度历史间的异常差异实现对客户端类型的精准识别，以抵抗符号翻转、噪声注入、标签翻转多类投毒攻击。考虑服务器可通过梯度长短历史信息间接推断出客户端原始梯度，PP-MPAD 采用周期性投毒攻击检测策略以有效减少隐私泄露的可能性。同时，结合基于链式安全多方计算的简单部署和无损计算优势，提出自适应多链安全聚合方法，在隐私保护的同时提升聚合效率。实验结果表明，PP-MPAD 算法能够有效识别并抵御多类投毒攻击，模型性能显著优于现有方法，并在隐私保护方面实现了无损聚合，兼顾了数据机密性与模型效用。考虑 Top-K 稀疏化、量化压缩等方法对梯度传输通信开销的降低，未来研究将探索梯度压缩技术与 PP-MPAD 相融合，从而在保持检测精度的同时大幅降低计算与通信成本，提升算法在带宽受限场景下的实用性。

## 基金项目

新疆维吾尔自治区自然科学基金项目(2023D01C20)；国家自然科学基金项目(62363032)。

## 参考文献

[1] Rashidi, G., Bounias, D., Bujotzek, M., Mora, A.M., Neher, P. and Maier-Hein, K.H. (2024) The Potential of Federated Learning for Self-Configuring Medical Object Detection in Heterogeneous Data Distributions. *Scientific Reports*, **14**, Article No. 23844. https://doi.org/10.1038/s41598-024-74577-0

[2] Abrams, L. (2024) UnitedHealth Says Data of 100 Million Stolen in Change Healthcare Breach. https://www.secrss.com/articles/71760

[3] Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T. and Bacon, D. (2016) Federated Learning: Strategies for Improving Communication Efficiency. https://arxiv.org/pdf/1610.05492

[4] 北京交通大学. 融合自适应权重分配和个性化差分隐私的联邦学习方法[P]. 中国专利, CN202210198444.5. 2022-06-07.

[5] Guerraoui, R. and Rouault, S. (2018) The Hidden Vulnerability of Distributed Learning in Byzantium. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 3521-3530.

[6] Kaur, H., Rani, V., Kumar, M., Sachdeva, M., Mittal, A. and Kumar, K. (2023) Federated Learning: A Comprehensive Review of Recent Advances and Applications. *Multimedia Tools and Applications*, **83**, 54165-54188. https://doi.org/10.1007/s11042-023-17737-0

[7] Blanchard, P., El Mhamdi, E.M., Guerraoui, R. and Stainer, J. (2017) Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 119-129.

[8] Xie, C., Koyejo, O. and Gupta, I. (2018) Generalized Byzantine-Tolerant SGD. https://arxiv.org/pdf/1802.10116

[9] Benjamin, J.G., Asokan, M., Yaqub, M. and Nandakumar, K. (2025) FedSECA: Sign Election and Coordinate-Wise Aggregation of Gradients for Byzantine Tolerant Federated Learning. 2025 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Nashville, 11-12 June 2025, 1771-1780. https://doi.org/10.1109/cvprw67362.2025.00165

[10] You, X., Liu, Z., Yang, X. and Ding, X. (2022) Poisoning Attack Detection Using Client Historical Similarity in Non-IID Environments. 2022 12*th International Conference on Cloud Computing, Data Science & Engineering* (*Confluence*), Noida, 27-28 January 2022, 439-447. https://doi.org/10.1109/confluence52989.2022.9734158

[11] Isik-Polat, E., Polat, G. and Kocyigit, A. (2023) ARFED: Attack-Resistant Federated Averaging Based on Outlier Elimination. *Future Generation Computer Systems*, **141**, 626-650. https://doi.org/10.1016/j.future.2022.12.003

[12] Yaldiz, D.N., Zhang, T. and Avestimehr, S. (2023) Secure Federated Learning Against Model Poisoning Attacks via Client Filtering. https://arxiv.org/pdf/2304.00160

[13] Yin, C. and Zeng, Q. (2024) Defending against Data Poisoning Attack in Federated Learning with Non-IID Data. *IEEE Transactions on Computational Social Systems*, **11**, 2313-2325. https://doi.org/10.1109/tcss.2023.3296885

[14] 蒋伟进, 杨璇, 李碧霞. 基于可解释贡献异常检测与动态裁剪的联邦学习投毒攻击防御方法[J]. 计算机学报, 2025, 48(12): 2855-2874.

[15] Gupta, A., Luo, T., Ngo, M.V. and Das, S.K. (2022) Long-Short History of Gradients Is All You Need: Detecting Malicious and Unreliable Clients in Federated Learning. In: Atluri, V., Di Pietro, R., Jensen, C.D. and Meng, W., Eds., *Lecture Notes in Computer Science*, Springer, 445-465. https://doi.org/10.1007/978-3-031-17143-7_22

[16] He, C., Liu, G., Guo, S. and Yang, Y. (2022) Privacy-Preserving and Low-Latency Federated Learning in Edge Computing. *IEEE Internet of Things Journal*, **9**, 20149-20159. https://doi.org/10.1109/jiot.2022.3171767

[17] Wang, B., Li, H., Guo, Y. and Wang, J. (2023) PPFLHE: A Privacy-Preserving Federated Learning Scheme with Homomorphic Encryption for Healthcare Data. *Applied Soft Computing*, **146**, Article 110677. https://doi.org/10.1016/j.asoc.2023.110677

[18] 李瑞琪, 贾春福, 王雅飞. 基于 NTRU 的多密钥同态代理重加密方案及其应用[J]. 通信学报, 2021, 42(3): 11-22.

[19] Wu, X., Zhang, Y., Shi, M., Li, P., Li, R. and Xiong, N.N. (2022) An Adaptive Federated Learning Scheme with Differential Privacy Preserving. *Future Generation Computer Systems*, **127**, 362-372. https://doi.org/10.1016/j.future.2021.09.015

[20] 曹世翔, 陈超梦, 唐朋, 等. 基于函数机制的差分隐私联邦学习算法[J]. 计算机学报, 2023, 46(10): 2178-2195.

[21] Li, H., Li, X., Liu, X., Wang, B., Wang, J. and Tian, Y. (2026) FedSam: Enhancing Federated Learning Accuracy with Differential Privacy and Data Heterogeneity Mitigation. *Computer Standards & Interfaces*, **95**, Article 104019. https://doi.org/10.1016/j.csi.2025.104019

[22] Li, Y., Zhou, Y., Jolfaei, A., Yu, D., Xu, G. and Zheng, X. (2021) Privacy-Preserving Federated Learning Framework Based on Chained Secure Multiparty Computing. *IEEE* Internet *of Things Journal*, **8**, 6178-6186. https://doi.org/10.1109/jiot.2020.3022911

[23] Cui, Y. and Zhu, J. (2023) Privacy Preserving Federated Learning Framework Based on Multi-Chain Aggregation. In: Wang, X., *et al*., Eds., *Lecture Notes in Computer Science*, Springer, 693-702. https://doi.org/10.1007/978-3-031-30637-2_46

[24] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., *et al*. (2017) Practical Secure Aggregation for Privacy-Preserving Machine Learning. *Proceedings of the* 2017 *ACM SIGSAC Conference on Computer and Communications Security*, Dallas, 30 October 2017-3 November 2017, 1175-1191. https://doi.org/10.1145/3133956.3133982

[25] Chen, X., Yu, H., Jia, X. and Yu, X. (2023) APFed: Anti-Poisoning Attacks in Privacy-Preserving Heterogeneous Federated Learning. *IEEE Transactions on Information Forensics and Security*, **18**, 5749-5761. https://doi.org/10.1109/tifs.2023.3315125

[26] Liu, X., Li, H., Xu, G., Chen, Z., Huang, X. and Lu, R. (2021) Privacy-Enhanced Federated Learning against Poisoning

Adversaries. *IEEE Transactions on Information Forensics and Security*, **16**, 4574-4588. https://doi.org/10.1109/tifs.2021.3108434

[27] Shen, X., Liu, Y., Li, F. and Li, C. (2024) Privacy-Preserving Federated Learning against Label-Flipping Attacks on Non-IID Data. *IEEE Internet of Things Journal*, **11**, 1241-1255. https://doi.org/10.1109/jiot.2023.3288886

[28] Le, J., Zhang, D., Lei, X., Jiao, L., Zeng, K. and Liao, X. (2023) Privacy-Preserving Federated Learning with Malicious Clients and Honest-but-Curious Servers. *IEEE Transactions on Information Forensics and Security*, **18**, 4329-4344. https://doi.org/10.1109/tifs.2023.3295949

[29] 姚玉鹏, 魏立斐, 张蕾. 一种隐私保护的抗投毒攻击联邦学习方案[J]. 计算机工程, 2025, 51(6): 223-235.

[30] Liu, J., Li, X., Liu, X., Zhang, H., Miao, Y. and Deng, R.H. (2024) DefendFL: A Privacy-Preserving Federated Learning Scheme Against Poisoning Attacks. *IEEE Transactions on Neural Networks and Learning Systems*, **35**, 13955-13969.

[31] 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.

[32] Feng, X., Cheng, W., Cao, C., Wang, L. and Sheng, V.S. (2024) DPFLA: Defending Private Federated Learning against Poisoning Attacks. *IEEE Transactions on Services Computing*, **17**, 1480-1491. https://doi.org/10.1109/tsc.2024.3376255

[33] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V. (2020) Federated Optimization in Heterogeneous Networks. *Proceedings of Machine Learning and Systems*, **2**, 429-450.