**Modeling and Simulation** 建模与仿真, 2025, 14(5), 246-258
Published Online May 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.145390

Hans汉斯

# CE-FedAvg方案抗模型投毒攻击研究

**王　双[1]，刘　亚[1,2]，赵逢禹[3]，曲　博[4]**

[1]上海理工大学光电信息与计算机工程学院，上海
[2]香港狮子山网络空间安全实验室，香港
[3]上海出版印刷高等专科学校信息与智能工程系，上海
[4]港专学院网络空间科技学院，香港

收稿日期：2025年4月12日；录用日期：2025年5月4日；发布日期：2025年5月13日

## 摘　要

**物联网设备的广泛部署带来了数据量的爆炸式增长，联邦学习不仅可以将分散的物联网设备上的数据协同利用，也可以保护数据的安全。但物联网环境中设备众多、资源有限，模型训练效率成为难题。量化压缩技术虽然通过降低传输精度来减少通信成本，但可能存在着投毒攻击风险。文章研究了量化联邦学习CE-FedAvg抵抗模型投毒攻击的能力。具体来说，通过设置不同恶意客户端占比，对CE-FedAvg选择了无目标攻击、缩放攻击、分布式后门攻击以及微量攻击，并在MNIST和CIFAR-10数据集上进行实验。实验结果表明：模型投毒攻击会导致CE-FedAvg精准率、F1值、召回率等指标大幅下降，最后提出了一种利用模型更新的一致性来检测恶意客户端的防御策略，可有效识别联邦学习中的恶意客户端。**

## 关键词

**联邦学习，投毒攻击，模型投毒攻击，模型压缩**

# Research on the Resistance of the CE-FedAvg Scheme to Model Poisoning Attacks

**Shuang Wang[1], Ya Liu[1,2], Fengyu Zhao[3], Bo Qu[4]**

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai
[2]Hong Kong Lion Rock Cyberspace Security Laboratory, Hong Kong
[3]Department of Information and Intelligent Engineering, Shanghai Publishing and Printing College, Shanghai
[4]Institute of Cyberspace Technology, Hong Kong College of Technology, Hong Kong

Received: Apr. 12th, 2025; accepted: May 4th, 2025; published: May 13th, 2025

文章引用: 王双, 刘亚, 赵逢禹, 曲博. CE-FedAvg 方案抗模型投毒攻击研究[J]. 建模与仿真, 2025, 14(5): 246-258.
DOI: 10.12677/mos.2025.145390

## Abstract

**The widespread deployment of Internet of Things (IoT) devices has led to an explosive growth in data volume. Federated learning (FL) not only enables collaborative utilization of decentralized data from IoT devices but also enhances data security. However, in IoT environments with numerous devices and limited resources, model training efficiency becomes a critical challenge. Although quantization compression techniques reduce communication costs by lowering transmission precision, they may introduce risks of poisoning attacks. This paper investigates the resilience of the quantized federated learning method CE-FedAvg against model poisoning attacks. Specifically, by varying the proportion of malicious clients, CE-FedAvg is subjected to untargeted model poisoning attacks, scaling attacks, distributed backdoor attacks, and a little is enough attack. Experiments conducted on the MNIST and CIFAR-10 datasets demonstrate that model poisoning attacks significantly degrade CE-FedAvg's performance metrics, including accuracy, F1 score, and recall. Finally, a defense strategy leveraging the consistency of model updates to detect malicious clients is proposed, which effectively identifies adversarial participants in federated learning.**

## Keywords

**Federated Learning, Poisoning Attack, Model Poisoning Attack, Model Compression**



## 1. 引言

在数字化时代，物联网(Internet of Things, IoT)设备的广泛部署带来了数据生成量的爆炸式增长。这些数据为训练智能模型提供了丰富的资源，但同时在效率和安全性方面也带来了前所未有的挑战。如何有效利用这些数据来协作训练智能模型，同时保护用户隐私和数据安全，是当下急需解决的问题。联邦学习(Federated Learning, FL) [1] [2]作为一种分布式机器学习范式，允许多个参与方协作训练模型，而无需将敏感数据集中到一个中心位置。这不仅解决了数据隐私保护的问题，还降低了数据传输和存储的成本。然而在实际应用中，模型训练需要频繁地在参与方和中心服务器之间传输模型参数，这不仅耗时，还会消耗大量的网络资源，因此在网络带宽受限、设备计算能力有限的 IoT 环境下，联邦学习的模型训练效率面临着严峻挑战。为降低通信成本，研究者提出量化压缩技术，通过适度降低模型参数的精度，从而减少传输数据量，达到降低模型通信成本的目的。

CE-FedAvg [3]方案是 2020 年提出的适用于物联网环境下基于量化策略的联邦学习方案。该方案对模型参数进行细致量化处理，在不显著影响模型性能的前提下，极大程度减少了每次传输的数据量，大幅降低训练过程中的通信次数，有效提升了模型训练效率，缓解了网络传输压力，在物联网等实际场景中展现出良好的应用潜力。然而 CE-FedAvg 方案在应用量化压缩技术时，暴露出了安全性问题[4] [5]。攻击者能够敏锐地察觉到量化过程中的潜在漏洞，进而发起多种类型的模型投毒攻击[6] [7]，譬如：无目标投毒攻击、缩放攻击、分布式后门攻击和微量攻击。

无目标投毒攻击[8]没有特定指向，旨在扰乱模型训练流程，降低模型整体性能；缩放攻击[9]则更为隐蔽，攻击者通过巧妙调整模型更新幅度，使攻击不易被察觉，进而破坏全局模型的准确性；分布式后门攻击[10]由多个恶意客户端协同完成，它们在模型中植入特定后门，当满足特定条件时，后门被激活，

对模型产生严重破坏；微量攻击[11]的攻击者仅利用少量精心构造的恶意样本，就能对模型造成较大影响，干扰模型的正常学习过程。鉴于 CE-FedAvg 方案所代表的量化联邦学习模式在提升训练效率、降低通信成本方面的重要作用，提高其安全性已成为当务之急。

本文聚焦于 CE-FedAvg 压缩量化方案，深入评估其抵抗四种模型投毒攻击的能力，并提出防御手段来提高 CE-FedAvg 方案的安全性。具体来说，首先通过理论分析和实验评估，深入揭示了量化压缩可能引入的安全漏洞；其次通过设置不同比例的恶意客户端数量，在 MNIST 和 CIFAR-10 数据集上，对 CE-FedAvg 方案分别模拟进行无目标投毒攻击、缩放攻击、分布式后门攻击和微量攻击，模拟攻击结果显示 CE-FedAvg 方案在精确度、F1 值、召回率等方面都有下降。证明了 CE–FedAvg 方案在面对多种模型中毒攻击时，性能表现脆弱，需要有效的防御机制来提升其安全性和鲁棒性。最后，基于历史模型更新预测客户端的模型更新，并在多次迭代中标记与预测不一致的客户端为恶意，设计了一种利用模型更新的一致性来检测恶意客户端的防御机制，提高了 CE-FedAvg 方案抗模型攻击的能力。本研究的主要贡献包括：

1) 设置不同恶意客户端占比，分析不同模型投毒攻击对量化压缩下联邦学习方案 CE-FedAvg 性能的影响，揭示了 CE-FedAvg 方案的安全漏洞。

2) 设计了一种利用模型更新的一致性来检测恶意客户端的防御机制，有效抵御了针对量化压缩的上述四种模型攻击，增强了 CE-FedAvg 方案的安全性。

## 2. 相关工作

### 2.1. 量化压缩的联邦学习

近年来，量化压缩技术已成为联邦学习突破通信与计算瓶颈的关键技术，是实现高效率、低消耗模型训练的核心手段。随着联邦学习在多领域应用的拓展，其分布式架构下的数据传输与协同计算压力凸显，量化压缩技术的重要性愈发显著。当前量化方法可分为量化感知训练(Quantization-Aware Training，QAT) [12]和训练后量化(Post-Training Quantization, PTQ) [13]两大类别。QAT 是在模型训练的起始阶段便巧妙嵌入量化操作，其核心思想是让模型在训练进程中主动适应低精度数据表示，从而有效缓解量化对模型性能的负面影响。PTQ 则遵循截然不同的技术路径，它聚焦于模型训练完成后的量化处理。

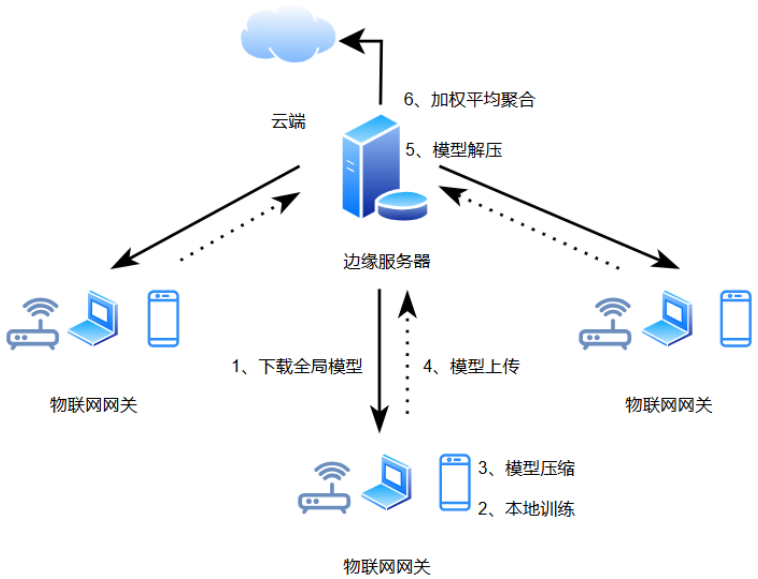


**Figure 1.** Architecture diagram of the CE-FedAvg scheme
**图 1.** CE-FedAvg 方案架构图

CE-FedAvg 方案是结合 QAT 和 PTQ 优势提出的一种改进的量化压缩联邦学习方案，如图 1 所示。该方案结合了量化技术和分布式 Adam 优化方法，引入了均匀量化和指数量化两种算法。均匀量化算法通过精确确定权重增量张量的取值范围，将权重增量映射为 8 位无符号整数。这种方法在有效压缩数据量的同时，最大程度地保留了权重更新的关键信息，保证了模型更新的稳定性和准确性。指数量化算法则针对 Adam 矩阵增量的特性进行设计，在不显著影响算法性能的前提下，降低了通信传输的数据量，提高了客户端与服务器之间的通信效率。CE-FedAvg 方案的量化策略与联邦学习的迭代训练过程高度契合，为解决联邦学习中的通信瓶颈问题提供了一个有效的途径。但还需要进一步考虑如何提高模型的安全性和鲁棒性，以应对恶意客户端的攻击。然而，随着联邦学习应用场景的不断拓展，恶意客户端的攻击威胁逐渐增大。尽管 CE-FedAvg 算法在量化压缩方面表现出色，但还需进一步深入研究如何提高模型的安全性和鲁棒性[14]，以有效抵御恶意客户端的攻击，确保联邦学习在复杂环境下的稳定运行。

### 2.2. 模型投毒攻击

在联邦学习场景下，各参与方通过交换模型参数来共同训练一个全局模型。模型投毒攻击正是利用了这一机制进行攻击的，即恶意客户端对本地训练后产生的模型更新参数进行恶意篡改，随后将这些被污染的参数上传至服务器。当服务器基于这些包含恶意信息的参数进行全局模型聚合时，整个模型的准确性和可靠性就会遭到严重破坏。这种攻击的危害极大，不仅会导致模型在正常任务上的性能大幅下降，还可能在特定场景下被攻击者利用，作出错误的决策。从攻击目的和方式上，模型投毒攻击主要分为无目标和有目标两类。无目标攻击旨在全面降低模型性能，而有目标攻击则是让模型在特定输入下给出攻击者预设的错误输出。

无目标投毒攻击全面破坏全局模型，使其在面对任何测试输入时，准确率都维持在较低水平。以 Fang 等人提出的针对联邦学习的无目标攻击框架为例，它将无目标攻击转化为一个优化问题。在恶意客户端上，攻击者通过一系列复杂的计算和策略设计，寻找最优的模型更新方式，目的是让攻击前后聚合模型更新之间的差异达到最大化。这种差异的最大化会严重干扰模型的收敛过程，使得模型无法准确学习到数据中的有效特征，从而导致模型性能全面崩塌。

缩放攻击属于有目标模型投毒攻击中的一种。攻击者首先会复制恶意客户端本地的训练样本，然后在这些样本中巧妙地嵌入特定的触发器，并为其分配攻击者预先设定的错误标签。基于这些被篡改的训练数据，攻击者计算出模型更新。更为关键的是，在将模型更新报告给服务器之前，攻击者会对其进行放大操作。这种放大使得恶意的模型更新在服务器进行全局模型聚合时，能够产生更大的影响，从而更有效地引导全局模型朝着攻击者期望的错误方向发展。

分布式后门攻击同样是有目标攻击，且极具隐蔽性。攻击者会将精心设计的触发器分解成多个局部模式，然后分别嵌入到不同恶意客户端的本地数据中。在联邦学习的模型聚合过程中，这些分散在各个恶意客户端的局部恶意信息会逐渐融合。当全局模型完成聚合后，就会被这些隐藏的恶意信息操控，使得模型在遇到包含特定触发器的测试输入时，给出攻击者预设的错误预测结果。分布式后门攻击由于其分布式的特点，很难被常规的检测手段发现。

微量攻击先采用与缩放攻击类似的方式计算模型更新，即通过篡改训练数据来获取恶意的模型更新参数。但与缩放攻击不同的是，攻击者会将计算得到的模型更新参数裁剪至一定范围。这一操作看似简单，实则效果显著，它使得传统的拜占庭鲁棒聚合规则难以识别和消除其中的恶意影响。即使联邦学习系统具备一定的容错机制，也难以抵御这种攻击，最终导致全局模型被成功破坏。

## 3. CE-FedAvg 抗模型投毒攻击方案设计

本节针对 CE-FedAvg 方案，设计抗模型投毒攻击方案并验证。在研究中，选用 MNIST 和 CIFAR-

10 数据集，因其数据规模、图像特征及分类难度的差异，可全面检验方案在不同场景下的表现。实验方案涵盖从数据准备、环境搭建到模型训练、攻击实施及结果评估的完整流程，精心选择多种攻击类型并设置参数。通过多维度评估指标，包括恶意客户端检测的检测率、精准率等以及全局模型的攻击成功率，全面衡量 CE-FedAvg 方案在抵御模型投毒攻击时的性能，为后续分析与总结提供坚实的数据基础。

### 3.1. 数据集

为深入探究 CE-FedAvg 量化联邦学习方案抵抗模型投毒攻击性能，本文在 MNIST 手写数字数据集和 CIFAR-10 图像分类数据集设计了一系列实验。具体信息如下：

MNIST 数据集包含 60000 张训练图像和 10000 张测试图像，图像大小为 28 × 28 像素，灰度值表示，用于识别 0~9 的手写数字。

CIFAR-10 数据集则更为复杂，由 10 个不同类别，共 60000 张 32 × 32 彩色图像组成，每个类别有 6000 张图像，广泛应用于图像分类任务研究。

这两个数据集在数据规模、图像特征和分类难度上各有特点，能够为实验提供丰富多样的数据支持。

### 3.2. 网络架构设计与参数选择

在处理 MNIST 数据集时，本文采用了一种针对性的网络架构设计。该架构从图像数据的特性出发，借助卷积层来提取手写数字图像中的关键特征。完成特征提取后，池化层开始发挥作用，它通过下采样操作，例如采用最大池化方法，在保留核心特征的同时，有效地减少了数据量，极大地降低了后续计算的复杂度。接着，全连接层将经过池化处理的特征向量进行整合，依据这些特征对数字的类别进行初步预测。最后，输出层基于全连接层的预测结果，给出最终的分类。网络模型的各层参数设计如表 1 所示。此外，设置训练迭代 200 次，学习率为 $2 \times 10^{-4}$。

当面对更为复杂的 CIFAR-10 数据集时，由于其包含 10 个不同类别的 32 × 32 彩色图像，对网络的学习能力提出了更高的要求。因此，本文选用了经典的 ResNet 网络架构，该架构可以让模型有能力学习到 CIFAR-10 数据集中各类图像更为复杂和细微的特征差异，为准确的图像分类提供了有力的保障。此外，设置学习率为 $10^{-3}$，训练迭代 200 次。

**Table 1.** Design of the CNN architecture for the global model on MNIST
**表 1.** MNIST 上全局模型的 CNN 架构的设计

| 层 | 大小 |
|---|---|
| 输入层 | 28 × 28 × 1 |
| 卷积层 | 3 × 3 × 30 |
| 池化层 | 2 × 2 |
| 卷积层 | 3 × 3 × 5 |
| 池化层 | 2 × 2 |
| 全连接层 | 100 |
| 输出层 | 10 |

整个实验均假设所有客户端都参与了 FL 训练的每次迭代，客户端数量固定为 100，恶意客户端占比分别设置为 0%、10%、20%、30%、50%，以观察恶意客户端数量变化研究 CE-FedAvg 方案抗模型投毒攻击的能力。

### 3.3. 方案设计

#### 3.3.1. 实验流程

在 MNIST 和 CIFAR-10 数据集上，对 CE-FedAvg 模拟进行无目标投毒攻击、缩放攻击、分布式后门攻击和微量攻击四种攻击，攻击的具体流程如下：

(1) 数据准备

分别下载并预处理 MNIST 手写数字数据集和 CIFAR-10 图像分类数据集。对 MNIST 数据集，将图像归一化到[0, 1]区间；对于 CIFAR-10 数据集，进行图像增强操作，如随机裁剪、水平翻转等，以扩充数据集并提升模型泛化能力。

(2) 环境搭建

在模拟的分布式环境中，设置好联邦学习框架，配置不同数量的客户端，本次实验将客户端数量设置为 100，模拟中心服务器，用于接收和聚合客户端上传的模型参数。

(3) 模型初始化

依据前文所述的网络架构，分别初始化针对 MNIST 数据集的卷积神经网络和针对 CIFAR-10 数据集 ResNet 网络，并设置好初始参数。

(4) 正常训练阶段

客户端接收全局参数后，执行本地 Adam 优化训练，生成权重和动量差异；随后对差异进行稀疏化(保留最大差异值)与混合量化(权重均匀量化、动量指数量化)，并通过 Golomb 编码压缩索引后上传；服务器解压后按数据量加权平均更新全局模型，迭代至收敛。

(5) 选择攻击类型

按照预定方案，选择一种无目标模型投毒攻击以及三种目标模型投毒攻击(缩放攻击、分布式后门攻击和微量攻击)。

(6) 设置攻击参数

无目标攻击以全面破坏模型性能为目标，不针对特定预测结果，无需设置攻击参数。三种目标模型中毒攻击针对不同攻击类型设置了特定参数，如下：

1) 缩放攻击：选择标签“0”作为目标标签，使攻击后的全局模型对嵌入触发器的测试输入错误预测为标签“0”，缩放参数设置为 1，以此增强攻击隐蔽性，让恶意模型更新在聚合时不易被发现。

2) 分布式后门攻击：同样选择标签“0”作为目标标签，确保攻击后模型按攻击者意图对特定输入进行错误分类。缩放参数设置为 1，降低攻击的可察觉性，使恶意更新融入正常模型更新中。

3) 微量攻击：选定标签“0”作为目标标签，引导模型对特定测试输入产生错误预测。

(7) 结果评估

服务器端运行恶意客户端检测机制，根据检测结果计算检测率、精准率、召回率和 F1 值。将检测出的恶意客户端数量与实际设置的恶意客户端数量进行对比，分析在不同恶意客户端占比情况下的性能表现。针对目标模型投毒攻击，计算攻击成功率(ASR)，通过分析 ASR 评估模型抵御目标攻击的能力。

#### 3.3.2. 评估指标

为了准确评估模型在遭受投毒攻击后的性能以及检测机制的有效性，本文围绕恶意客户端检测和全局模型学习效果两个关键方面设定了全面的评估指标。

(1) 恶意客户端检测指标

1) 检测率：该指标体现了检测机制对客户端性质的整体识别能力，计算公式为(正确识别的良性客户端数量 + 正确识别的恶意客户端数量)/客户端总数 × 100%，数值越高表明检测机制对所有客户端的识

别能力越强。

2) 精准率：主要关注检测结果中恶意客户端判断的准确性，即被判定为恶意客户端中实际确实为恶意的比例，通过(正确识别的恶意客户端数量/被判定为恶意客户端的数量) × 100%计算得出，精准率越高说明检测结果中误判为恶意的良性客户端越少。

3) 召回率：反映检测机制对恶意客户端的捕捉能力，指实际的恶意客户端被正确识别出来的比例，即(正确识别的恶意客户端数量/恶意实际客户端数量) × 100%，召回率越高意味着检测机制遗漏的恶意客户端越少。

4) F1 值：综合精准率和召回率，通过公式 F1 = 2 × (精准率 × 召回率)/(精准率 + 召回率)计算得出，能更全面地评估检测机制性能。F1 值兼顾了检测的准确性和完整性，取值范围在 0~1 之间，越接近 1 表示检测机制性能越好。

(2) 全局模型学习效果指标

攻击成功率(ASR)：专门针对目标模型投毒攻击而设定。在实验中，将特定触发器嵌入每个测试输入，ASR 表示被全局模型分类为目标标签的触发器嵌入测试输入的比例，即(被分类为目标标签的触发器嵌入测试输入数量/触发器嵌入测试输入总数) × 100%。较低的 ASR 表明全局模型在面对目标模型投毒攻击时，抗攻击能力更好，不容易被攻击者误导而输出错误的目标标签。

## 4. 实验结果与分析

本章节围绕 CE-FedAvg 抗模型投毒攻击方案的实验结果展开深度剖析。首先探究恶意客户端占比对攻击成功率的影响，发现随着占比提升，MNIST 和 CIFAR-10 数据集的攻击成功率均上升，且不同攻击类型表现各异。接着分析对检测率、精准率、召回率及 F1 值的影响，结果显示这些指标均随恶意客户端占比增加而下降，缩放攻击相对易检测，分布式后门攻击则较为隐蔽难测。最后，基于实验分析，提出基于 L-BFGS 算法的防御方法，通过预测模型更新、计算可疑分数来检测并移除恶意客户端，以保障联邦学习中全局模型的准确性与稳定性。

### 4.1. 恶意客户端占比对检测率的影响

在 MNIST 数据集下，针对不同恶意客户端占比的检测率实验中，整体来看，随着恶意客户端占比从 0%逐步提升至 50%，针对无目标模型攻击、缩放攻击、分布式后门攻击以及微量攻击的检测率均呈现出明显的下降趋势，这清晰表明恶意客户端数量的增加显著加大了检测机制的工作难度，恶意客户端占比越高，检测机制越难以从众多客户端中准确识别出恶意行为，如图 2 和表 2 所示。分不同攻击类型而言，缩放攻击在各恶意客户端占比下检测率相对较高，这或许是因为其攻击行为模式相对规律，检测机制较易捕捉到其特征，例如在恶意客户端占比为 10%时，检测率达 94.05%，高于其他三种攻击类型；分布式后门攻击的检测率相对较低，该攻击将恶意行为分散，更为隐蔽，如恶意客户端占比 50%时，检测率仅 69.90%，充分体现了检测的困难程度；无目标模型攻击与微量攻击的检测率变化趋势和数值都较为接近，两者攻击的随机性和隐蔽性导致它们在检测难度上相近，在恶意客户端占比变化过程中，两者检测率差距始终较小。

在 CIFAR-10 数据集基础上针对不同恶意客户端占比的检测率实验中，整体趋势明显，随着恶意客户端占比从 0%攀升至 50%，无目标模型攻击、缩放攻击、分布式后门攻击以及微量攻击的检测率也都持续走低。从不同攻击类型分析，缩放攻击在各个恶意客户端占比阶段检测率相对较高，可能是其在模型参数变动或数据特征上存在独特且易被捕捉的模式，例如：在 10%恶意客户端占比时，检测率为 89.95%，领先于其他攻击类型；分布式后门攻击的检测率一直处于低位，由于 CIFAR-10 数据集内容复杂、特征丰

富，该攻击利用这种特性将恶意后门隐藏得更深，使得检测机制难以从中提取恶意特征，当恶意客户端占比达 50%时，检测率仅 64.95%，在四种攻击中最低；无目标模型攻击和微量攻击的检测率相近且变化趋势一致，二者攻击的随机性与隐蔽性，使得随着恶意客户端占比上升，检测率同步下降。

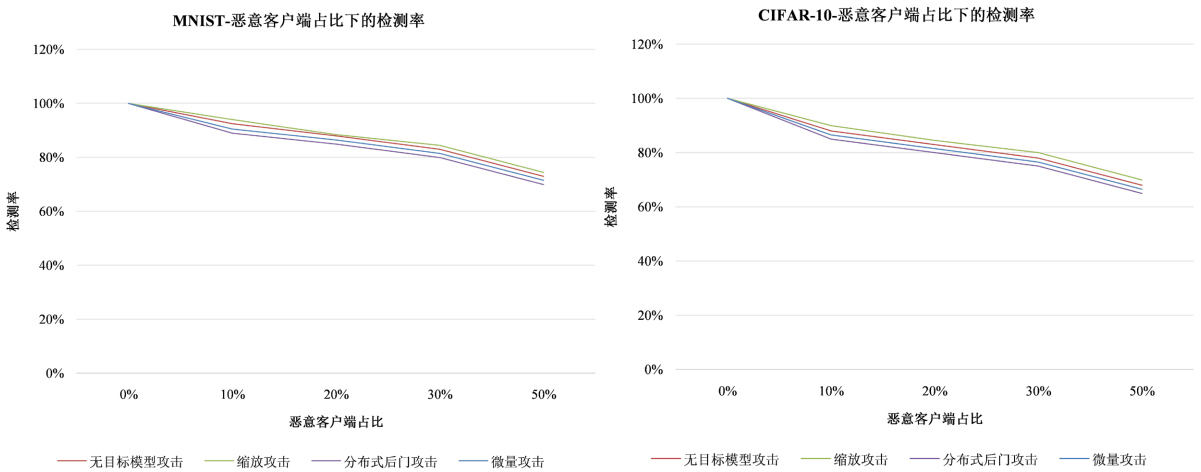


**Figure 2.** Influence of the proportion of malicious clients on detection rate
**图 2.** 恶意客户端占比对检测率的影响

**Table 2.** Table of data on the influence of the proportion of malicious clients on detection rate
**表 2.** 恶意客户端占比对检测率影响的数据表

| 数据集 | 恶意客户端占比 | 无目标模型攻击 | 缩放攻击 | 分布式后门攻击 | 微量攻击 |
|---|---|---|---|---|---|
| MNIST | 0% | 100% | 100% | 100% | 100% |
| | 10% | 92.47% | 94.05% | 88.98% | 90.53% |
| | 20% | 87.96% | 88.47% | 84.95% | 86.48% |
| | 30% | 82.98% | 84.50% | 79.94% | 81.45% |
| | 50% | 72.98% | 74.45% | 69.90% | 71.50% |
| CIFAR-10 | 0% | 100% | 100% | 100% | 100% |
| | 10% | 87.98% | 89.95% | 84.98% | 86.46% |
| | 20% | 82.95% | 84.47% | 79.96% | 81.47% |
| | 30% | 77.98% | 79.95% | 74.99% | 76.50% |
| | 50% | 67.98% | 69.90% | 64.95% | 66.45% |

## 4.2. 恶意客户端占比对精准率的影响

在 MNIST 数据集的实验研究中，针对不同恶意客户端占比下各类模型投毒攻击的精准率表现进行了系统性分析。实验数据显示，随着恶意客户端比例从 0%逐步提升至 50%，无目标模型攻击、缩放攻击、分布式后门攻击及微量攻击的精准率均呈现显著下降趋势(如图 3 和表 3 所示)。以无目标模型攻击为例，当恶意客户端占比从 10%增至 50%时，其精准率由 88.23%骤降至 66.07%，降幅达 22.16 个百分点；类似地，分布式后门攻击的精准率在同一阶段从 85.07%下滑至 65.04%。这种普遍性下降趋势表明，恶意客户端数量的增加不仅直接提升了攻击的破坏力，更通过引入大量干扰性模型更新，严重干扰了检测机制对恶意行为的判断能力。具体而言，当恶意客户端占比超过 30%时，检测机制对各类攻击的误判率平均上

升至 23.98%，导致良性客户端被错误标记的概率显著增加。值得注意的是，尽管所有攻击类型的精准率均随恶意客户端占比升高而恶化，但不同攻击类型间存在显著差异。缩放攻击在 MNIST 数据集上的表现尤为突出，其在 50%恶意占比下仍能保持 68.05%的精准率，显著高于其他攻击类型。

在 CIFAR-10 数据集上，精准率的下降趋势与 MNIST 具有相似规律，但整体数值更低且恶化速度更快。例如，当恶意客户端占比从 10%增至 50%时，缩放攻击的精准率从 85.03%降至 65.07%，降幅达 19.96 个百分点，显著高于 MNIST 数据集同期的 14.13 个百分点降幅。这一差异主要源于 CIFAR-10 数据集的复杂性：其 32 × 32 彩色图像包含更丰富的纹理特征与类别间重叠，导致正常模型更新的参数分布本身具有更高的方差。在此背景下，恶意客户端通过构造与正常数据分布部分重叠的对抗样本，使得恶意更新的统计特征更易隐藏于正常更新的高方差范围内。值得注意的是，在相同恶意客户端占比下，CIFAR-10 的精准率始终低于 MNIST (如 50%占比时，CIFAR-10 无目标攻击精准率为 63.05%，而 MNIST 为 66.07%)，这进一步印证了复杂数据环境下检测难度的升级——检测算法需在更高噪声背景下捕捉更细微的异常模式，对其敏感性与鲁棒性提出了双重考验。

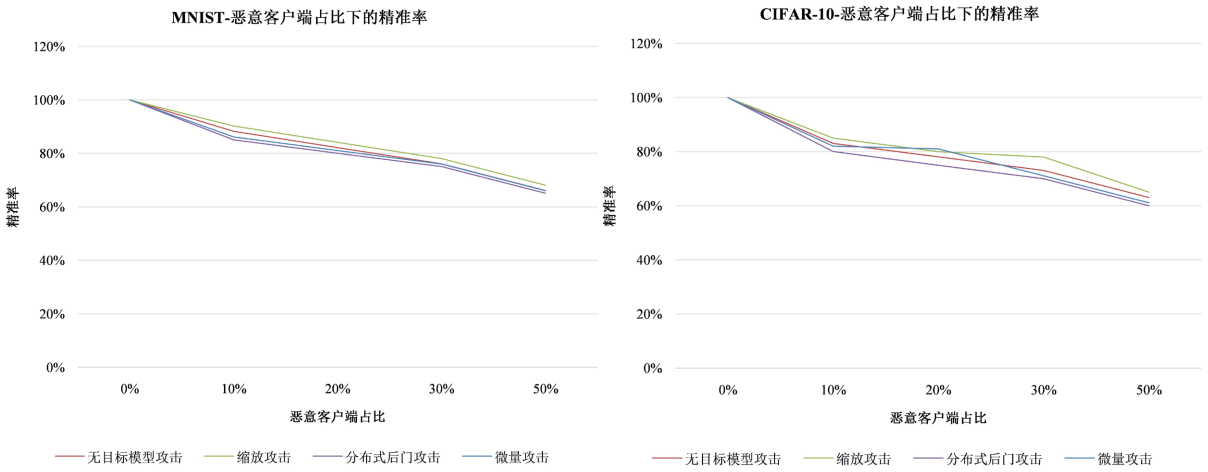


**Figure 3.** Influence of the proportion of malicious clients on precision
**图 3.** 恶意客户端占比对精准率的影响

**Table 3.** Table of data on the influence of the proportion of malicious clients on precision
**表 3.** 恶意客户端占比对精准率影响的数据表

| 数据集 | 恶意客户端占比 | 无目标模型攻击 | 缩放攻击 | 分布式后门攻击 | 微量攻击 |
|---|---|---|---|---|---|
| MNIST | 0% | 100% | 100% | 100% | 100% |
| | 10% | 88.23% | 90.18% | 85.07% | 86.12% |
| | 20% | 82.15% | 84.09% | 80.03% | 81.05% |
| | 30% | 76.08% | 78.05% | 75.03% | 76.02% |
| | 50% | 66.07% | 68.05% | 65.04% | 66.02% |
| CIFAR-10 | 0% | 100% | 100% | 100% | 100% |
| | 10% | 83.05% | 85.03% | 80.07% | 82.05% |
| | 20% | 78.08% | 80.05% | 75.03% | 81.01% |
| | 30% | 73.05% | 78.08% | 70.05% | 71.10% |
| | 50% | 63.05% | 65.07% | 60.05% | 61.12% |

### 4.3. 恶意客户端占比对召回率的影响

在 MNIST 数据集的召回率实验中，随着恶意客户端占比从 0%逐步提升至 50%，无目标模型攻击、缩放攻击、分布式后门攻击以及微量攻击的召回率均出现了明显的下降，如图 4 和表 4 所示。从攻击类型来看，缩放攻击在各恶意客户端占比下的召回率相对较高。例如，在 10%恶意客户端占比时，召回率达到 96.15%。这与缩放攻击在精准率方面的表现相呼应，原因在于其攻击特征相对明显，检测机制能够较为容易地识别出这些特征，从而全面发现实施该攻击的恶意客户端。相反，分布式后门攻击的召回率一直处于较低水平，在 50%恶意客户端占比时仅为 78.13%。这是因为该攻击方式具有很强的隐蔽性，恶意行为分散在数据中，检测机制难以全面察觉所有的恶意迹象，导致在发现所有实施该攻击的恶意客户端时面临较大困难。无目标模型攻击和微量攻击的召回率变化趋势和数值相近，它们的随机性和隐蔽性给检测机制在全面发现恶意客户端方面带来了相似的挑战，不同占比下两者的召回率同步下降。

CIFAR-10 数据集在召回率方面也呈现出与 MNIST 数据集相似的变化规律。随着恶意客户端占比从 0%增长到 50%，四种攻击类型的召回率均持续降低。缩放攻击依然保持相对较高的召回率，如在 10%恶意客户端占比时为 93.51%，这表明其攻击行为在该数据集中也具有一定的可识别性，但其优势随恶意客户端占比升高而减弱。例如，50%占比时召回率降至 80.49%，较 MNIST 降低 1.18 个百分点。这是因为 CIFAR-10 模型的深层网络结构对梯度缩放更具鲁棒性。分布式后门攻击的召回率持续偏低，在 50%恶意客户端占比时为 75.34%，低于 MNIST 数据集相同占比下的召回率。这主要是由于 CIFAR-10 数据集本身的复杂性，使得该攻击的隐蔽性更强，检测机制更难全面发现所有实施该攻击的恶意客户端。无目标模型攻击和微量攻击的召回率变化趋势和数值接近，在复杂的数据环境中给检测机制全面发现恶意客户端带来了相似的难题。

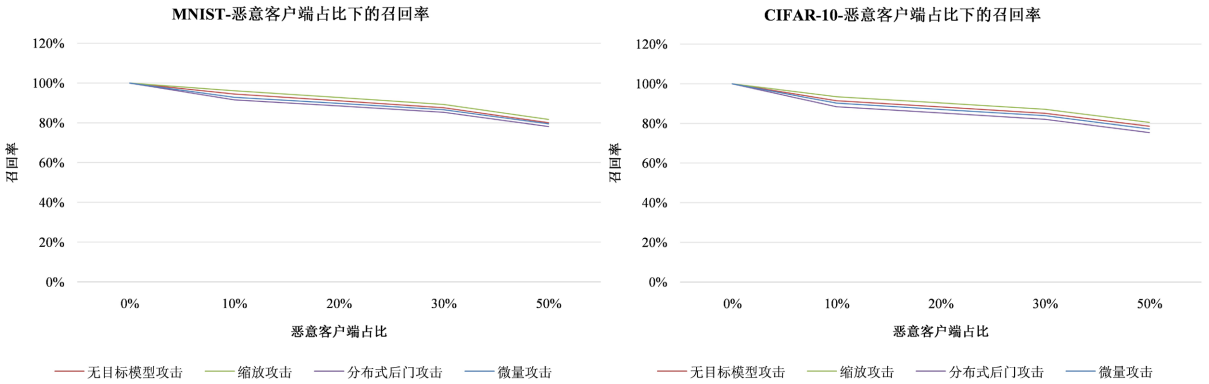


**Figure 4.** Influence of the proportion of malicious clients on recall rate
**图 4.** 恶意客户端占比对召回率的影响

**Table 4.** Table of data on the influence of the proportion of malicious clients on recall rate
**表 4.** 恶意客户端占比对召回率影响数据表

| 数据集 | 恶意客户端占比 | 无目标模型攻击 | 缩放攻击 | 分布式后门攻击 | 微量攻击 |
|---|---|---|---|---|---|
| MNIST | 0% | 100% | 100% | 100% | 100% |
| | 10% | 94.47% | 96.15% | 91.58% | 92.86% |
| | 20% | 91.05% | 92.68% | 88.47% | 89.77% |
| | 30% | 87.64% | 89.24% | 85.26% | 86.58% |
| | 50% | 80.08% | 81.67% | 78.13% | 79.47% |

续表

| | | | | | |
|---|---|---|---|---|---|
| | 0% | 100% | 100% | 100% | 100% |
| | 10% | 91.47% | 93.51% | 88.43% | 90.20% |
| CIFAR-10 | 20% | 88.28% | 90.32% | 85.25% | 87.04% |
| | 30% | 85.08% | 87.10% | 82.04% | 83.86% |
| | 50% | 78.57% | 80.49% | 75.34% | 77.17% |

### 4.4. 恶意客户端占比对 F1 值的影响

在 MNIST 和 CIFAR-10 数据集的实验中，F1 值的变化趋势紧密关联着精准率和召回率。整体来看，随着恶意客户端占比从 0%逐步提升至 50%，无论是 MNIST 还是 CIFAR-10 数据集，针对无目标模型攻击、缩放攻击、分布式后门攻击以及微量攻击的 F1 值都呈现出明显的下降趋势，如表 5 所示。这一趋势与精准率和召回率的下降趋势相呼应，进一步表明恶意客户端数量的增加，检测机制面对的干扰增多，导致精准率和召回率难以同时维持在较高水平，进而拉低了 F1 值。

**Table 5.** F1 score of the proportion of malicious clients
**表 5.** 恶意客户端占比的 F1 值

| 数据集 | 恶意客户端占比 | 无目标模型攻击 | 缩放攻击 | 分布式后门攻击 | 微量攻击 |
|---|---|---|---|---|---|
| | 0% | 100% | 100% | 100% | 100% |
| | 10% | 91.17% | 93.05% | 88.22% | 89.35% |
| MNIST | 20% | 86.28% | 88.18% | 84.08% | 85.10% |
| | 30% | 81.40% | 83.39% | 79.75% | 80.96% |
| | 50% | 72.37% | 74.28% | 70.89% | 71.98% |
| | 0% | 100% | 100% | 100% | 100% |
| | 10% | 87.14% | 89.15% | 84.07% | 85.88% |
| CIFAR-10 | 20% | 82.83% | 84.91% | 72.92% | 83.78% |
| | 30% | 78.59% | 82.35% | 75.70% | 76.85% |
| | 50% | 69.84% | 72.07% | 67.03% | 68.25% |

### 4.5. 恶意客户端占比对攻击成功率的影响

在针对 MNIST 和 CIFAR-10 数据集的目标模型投毒攻击实验中，攻击成功率(ASR)作为核心指标，直观量化了恶意客户端对全局模型的破坏程度。如表 6 中的实验数据所示，随着恶意客户端占比从 10%攀升至 50%，两类数据集的 ASR 均呈现指数级增长，模型抗攻击能力急剧衰减。以缩放攻击为例，MNIST 数据集在 50%恶意占比时 ASR 达到 36.91%，较 10%占比的 5.27%增长近 7 倍；而 CIFAR-10 的同类攻击表现更为严峻，ASR 从 12.38%激增至 55.21%，增幅达 3.45 倍。这一现象验证了恶意客户端规模与攻击效果的强正相关性——当恶意节点占比超过 30%时，其上传的污染参数在模型聚合中占据主导地位，直接扭曲全局模型的决策边界。

从攻击类型差异来看，MNIST 数据集的分布式后门攻击展现出更强的隐蔽性：在 50%恶意客户端占比时 ASR 达到 31.25%，远超缩放攻击的 36.91%。这是因为分布式后门将触发器分片植入不同客户端，规避了传统异常检测机制。而 CIFAR-10 的微量攻击在 50%恶意客户端时 ASR 高达 60.73%，显著高于 MNIST 的 42.67%，表明复杂图像数据中的梯度裁剪策略能更有效地绕过模型收敛约束。

**Table 6.** ASR of the proportion of malicious clients
**表 6.** 恶意客户端占比的 ASR

| 数据集 | 恶意客户端占比 | 缩放攻击 | 分布式后门攻击 | 微量攻击 |
|---|---|---|---|---|
| MNIST | 10% | 5.27% | 3.89% | 7.15% |
| | 20% | 12.45% | 9.74% | 15.32% |
| | 30% | 23.68% | 18.53% | 27.44% |
| | 50% | 36.91% | 31.25% | 42.67% |
| CIFAR-10 | 10% | 12.38% | 8.64% | 15.29% |
| | 20% | 24.73% | 19.87% | 28.46% |
| | 30% | 37.79% | 31.05% | 42.25% |
| | 50% | 55.21% | 47.82% | 60.73% |

## 4.6. 防御方法

基于上述针对联邦学习中模型投毒攻击的实验分析，为有效应对恶意客户端对全局模型的破坏，本文提出了一种全新的防御方法。在模型更新阶段，服务器利用 Cauchy 均值定理结合 L-BFGS 算法预测每个客户端的模型更新以此衡量模型更新的一致性。具体而言，通过计算预测更新与实际更新之间的欧几里得距离，为每个客户端分配可疑分数。然后，运用 Gap 统计量和均值聚类方法，依据可疑分数检测恶意客户端。一旦检测到恶意客户端，服务器将其移除并重新启动训练过程，从而保证剩余客户端能在相对安全的环境下进行联邦学习，保障全局模型的准确性和稳定性，实验结果如图 5 所示。使用测试精度(TACC)评估学习到的全局模型，它是全局模型在测试数据集上正确分类的样本比例，即(正确分类的测试样本数量/测试样本总数) × 100%，TACC 越高，表明模型对不同样本的分类能力越强，泛化性能越好。

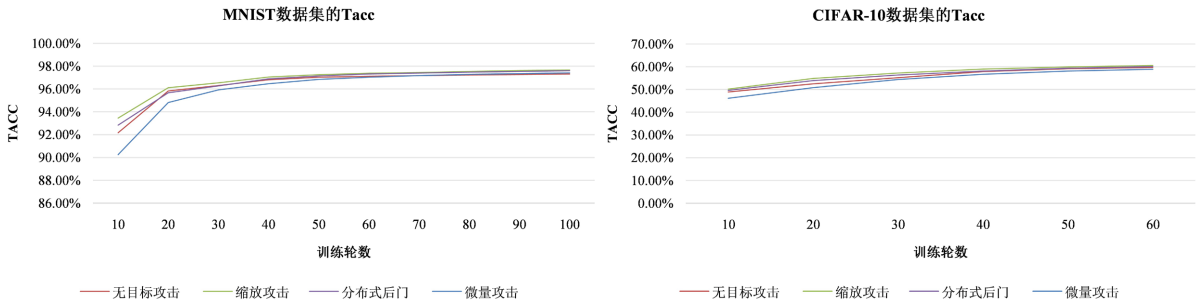


**Figure 5.** TACC of the different dataset
**图 5.** 不同数据集的 TACC

## 5. 结语

本文针对物联网环境下量化联邦学习的模型投毒攻击问题展开研究。通过理论分析与实验验证，揭示了 CE-FedAvg 方案在无目标攻击、缩放攻击、分布式后门攻击及微量攻击下的安全脆弱性，量化压缩导致的参数精度损失进一步放大了攻击危害。基于此，本文提出了一种基于模型更新一致性的恶意客户端检测机制，通过预测更新偏差计算可疑分数，结合聚类分析实现高效识别，在 MNIST 和 CIFAR-10 数据集上验证了防御策略的有效性。然而，当前防御方法仍存在局限性：依赖历史更新统计特征的检测机制难以应对动态演化的对抗性攻击，欧氏距离度量在异构数据分布场景中易引发误判，且可疑分数计算引入的额外开销可能制约实时性需求。未来研究需进一步探索动态自适应检测框架，通过融合在线学习

与对抗样本生成技术提升对攻击策略变化的响应能力；同时优化轻量化检测算法，结合知识蒸馏与硬件加速技术降低计算成本，以适应资源受限的物联网环境。

## 参考文献


[1] McMahan, B., Moore, E., Ramage, D., *et al*. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*), Fort Lauderdale, 20-22 April 2017, 1273-1282.

[2] Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019) Federated Machine Learning. *ACM Transactions on Intelligent Systems and Technology*, **10**, 1-19. https://doi.org/10.1145/3298981

[3] Mills, J., Hu, J. and Min, G. (2020) Communication-Efficient Federated Learning for Wireless Edge Intelligence in IoT. *IEEE Internet of Things Journal*, **7**, 5986-5994. https://doi.org/10.1109/jiot.2019.2956615

[4] 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.

[5] 赵亚茹, 张建标, 曹益皓, 等. 云边联邦学习系统下抗投毒攻击的防御方法[J/OL]. 软件学报, 1-21. https://doi.org/10.13328/j.cnki.jos.007266, 2025-03-15.

[6] Cao, X. and Gong, N.Z. (2022) MPAF: Model Poisoning Attacks to Federated Learning Based on Fake Clients. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), New Orleans, 19-20 June 2022, 3395-3403. https://doi.org/10.1109/cvprw56347.2022.00383

[7] Bhagoji, A.N., Chakraborty, S., Mittal, P., *et al*. (2019) Analyzing Federated Learning through an Adversarial Lens. 2019 *International Conference on Machine Learning*, Long Beach, 10-15 June 2019, 634-643.

[8] Fang, M., Cao, X., Jia, J., *et al*. (2020) Local Model Poisoning Attacks to {Byzantine-Robust} Federated Learning. 2020 29*th USENIX Security Symposium* (*USENIX Security* 20), Boston, 12-14 August 2020, 1605-1622.

[9] Bagdasaryan, E., Veit, A., Hua, Y., *et al*. (2020) How to Backdoor Federated Learning. 2020 *International Conference on Artificial Intelligence and Statistics*, Online, 26-28 August 2020, 2938-2948.

[10] Xie, C., Huang, K., Chen, P.Y., *et al*. (2019) DBA: Distributed Backdoor Attacks against Federated Learning. 2019 *International Conference on Learning Representations*, New Orleans, 6-9 May 2019.

[11] Baruch, G., Baruch, M. and Goldberg, Y. (2019) A Little Is Enough: Circumventing Defenses for Distributed Learning. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8 December 2019, 8635-8645.

[12] Zhang, D., Yang, J., Ye, D. and Hua, G. (2018) LQ-Nets: Learned Quantization for Highly Accurate and Compact Deep Neural Networks. In: *Lecture Notes in Computer Science*, Springer, 373-390. https://doi.org/10.1007/978-3-030-01237-3_23

[13] Banner, R., Nahshan, Y. and Soudry, D. (2019) Post Training 4-Bit Quan-Tization of Convolutional Networks for Rapid Deployment. *Advances in Neural Information Processing Systems*, **32**, 7948-7956.

[14] 陈晋音, 曹志骐, 郑海斌, 等. 面向模型量化的安全性研究综述[J/OL]. 小型微型计算机系统, 1-23. http://kns.cnki.net/kcms/detail/21.1106.tp.20250117.1440.018.html, 2025-03-15.