A Robust Privacy-Preserving Federated Learning Model Against Model Poisoning Attacks

内容纲要

1、 背景介绍

2、 模型技术

3、 实验验证

内容纲要

1、 背景介绍

2、 模型技术

3、 实验验证

1、 背景介绍

Journal：IEEE Transactions on Information Forensics and Security

Author：Abbas Yazdinejad; Ali Dehghantanha; Hadis Karimipour; Gautam Srivastava; Reza M. Parizi

Cite this：A. Yazdinejad, A. Dehghantanha, H. Karimipour, G. Srivastava and R. M. Parizi, "A Robust Privacy-Preserving Federated Learning Model Against Model Poisoning Attacks," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 6693-6708, 2024, doi: 10.1109/TIFS.2024.3420126.

1、 背景介绍

论文的核心目标/任务

在应对模型投毒攻击时，现有的防御策略主要集中于检测可疑的明文本地梯度。然而，检测非独立同分布的加密梯度对现有的方法构成了重大挑战。此外，在隐私保护联邦学习中，尤其是在加密梯度的背景下，解决计算复杂性和通信开销问题变得至关重要。

我们的方法引入了一个**internal auditor**（内部审计员），用于评估加密梯度的相似性和分布，以区分良性梯度和恶意梯度，并采用**Gaussian Mixture Model**（高斯混合模型）和**Mahalanobis Distance**（马氏距离）进行容错聚合。利用同态加密来确保机密性，同时最小化计算和通信开销。

本文主要贡献如下：

1）设计并实现了一种新颖的**PPFL**框架，该架构使用了加法同态加密（**AHE**）来确保数据的保密性。我们提出的模型**paillier**密码系统为加密梯度提供加法同态加密，以抵御模型中毒攻击，同时保护用户、服务器和梯度的隐私。

2）集成了一个内部审计机制，显著提高了检测和预防此类攻击的准确性，降低了将良性梯度误判为有害梯度的风险，简化了预测和预防有害梯度的过程。

3）引入了一种新的审计协议，应对异构数据的挑战，且抵消模型中毒攻击过程中恶意用户的影响。

4）引入了优化技术，最小化计算和通信的成本

1、 背景介绍

预备知识

高斯混合模型（**Gaussian Mixture Models, GMMs**）是一种统计模型，用于将数据集表示为多个高斯分布的混合，每个高斯分布都有其自身的均值和方差。在数据被认为是由多个高斯源生成的场景中，这些模型表现出色。在我们的隐私保护联邦学习（**PPFL**）模型背景下，**GMMs**提供了一种复杂的方法来分析梯度分布的异质性，使我们能够将梯度分类为不同的簇。这种分类对于区分良性梯度（通常与模型的学习目标一致）和对抗性梯度（试图破坏模型性能）特别有用。

马氏距离（**Mahalanobis Distance, MD**）是一种多维度量标准，用于衡量一个点与一个分布之间的距离。与欧几里得距离不同，马氏距离具有尺度不变性，并考虑了数据集的相关性。在我们的模型中，马氏距离被用作梯度分布中异常值检测的强大工具。它评估给定梯度与预期分布的偏离程度，从而有效地识别出异常或可能由恶意活动导致的梯度。这使得马氏距离特别适用于确保隐私保护联邦学习（**PPFL**）中学习过程的完整性，因为在该过程中，梯度数据可能是多样的且受到各种来源的影响。

1、 背景介绍

问题表述

给定N个客户端，其数据集Di可能包括IID或非IID数据分布，目标是通过最小化损失函数L来训练全局模型M，PPFL中对抗性攻击问题可以表示在隐私、鲁棒性和效率相关约束下最小化L（M），公式如下：

$$minimize_L(M) = (\frac{1}{N}) \times \sum_{i}^{N} L(D_i, M) \qquad (1)$$

在最小化公式的时候，最重要的是考虑以下约束：1）计算出梯度（g）后，使用经过验证的梯度（gi）来更新全局模型M。2）对于i=1……N，有gi = ▽(Di, M)，其中梯度经过加密以保持机密性。3）满足一些性能约束，包括但不限于数据安全、模型稳定性等。

1、 背景介绍

模型威胁

本文重点研究了联邦学习中恶意用户可利用的安全和隐私威胁。

威胁1：诚实但好奇（HBC）的对手

威胁描述：在联邦学习（FL）中，服务器知晓所有本地梯度和密文，虽系统基于其HBC行为假设运行，但服务器可能发起隐私攻击，如通过数据重建或推理攻击来获取用户数据隐私，其核心是寻求敏感的全局模型信息。

应对目标：保护本地梯度的保密性，因为服务器或其他对手可能通过共享梯度和全局参数暴露敏感用户数据，而在服务器传输前对单个梯度进行加密可提供一定程度的保密性。

威胁2：注入有毒梯度

威胁描述：拜占庭参与者可能伪装成提交源自异构数据的梯度，实则提交欺诈性梯度，从而破坏模型的完整性。

应对目标：改进对加密梯度的审查机制，以区分良性和恶意用户，增强系统对中毒攻击的抵御能力。这需要更精准地识别和过滤异常梯度，确保参与模型训练的梯度真实可靠，维护模型的准确性和稳定性。

内容纲要

1、 背景介绍

2、 模型技术

3、 实验验证

2、 模型技术

模型架构

文章中模型的基本实体包括数据所有者（用户）、审计实体和聚合服务器

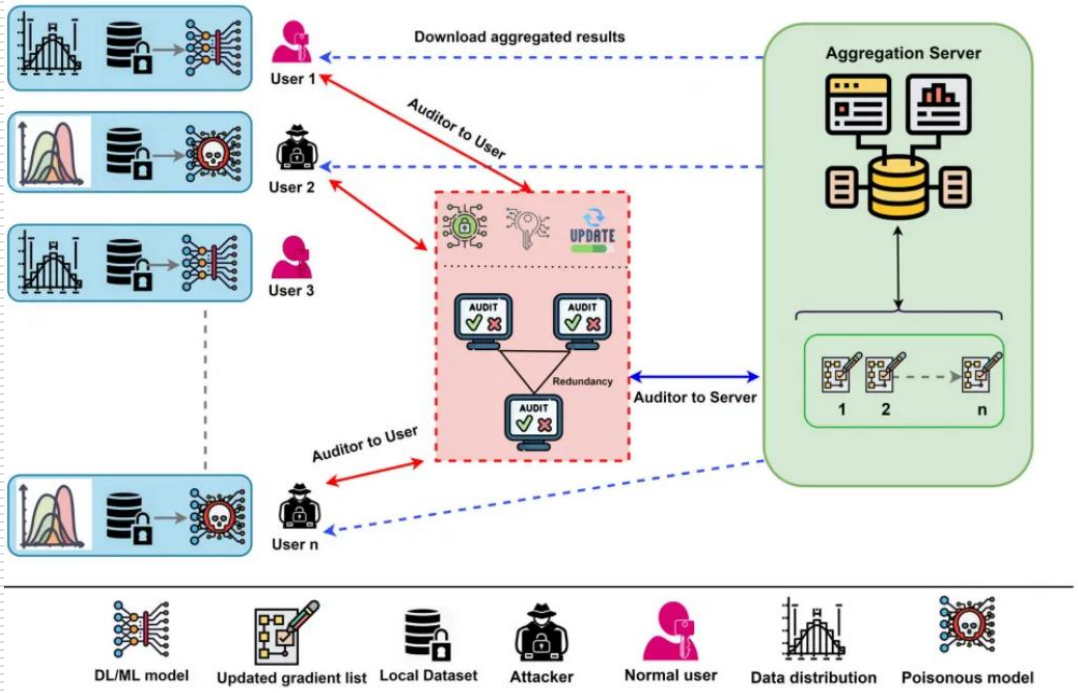


Fig. 1. Architecture of the proposed model.

2、模型技术

技术概述

如图展示了模型的技术概述：在服务器端对加密梯度进行内部审计以检测恶意梯度

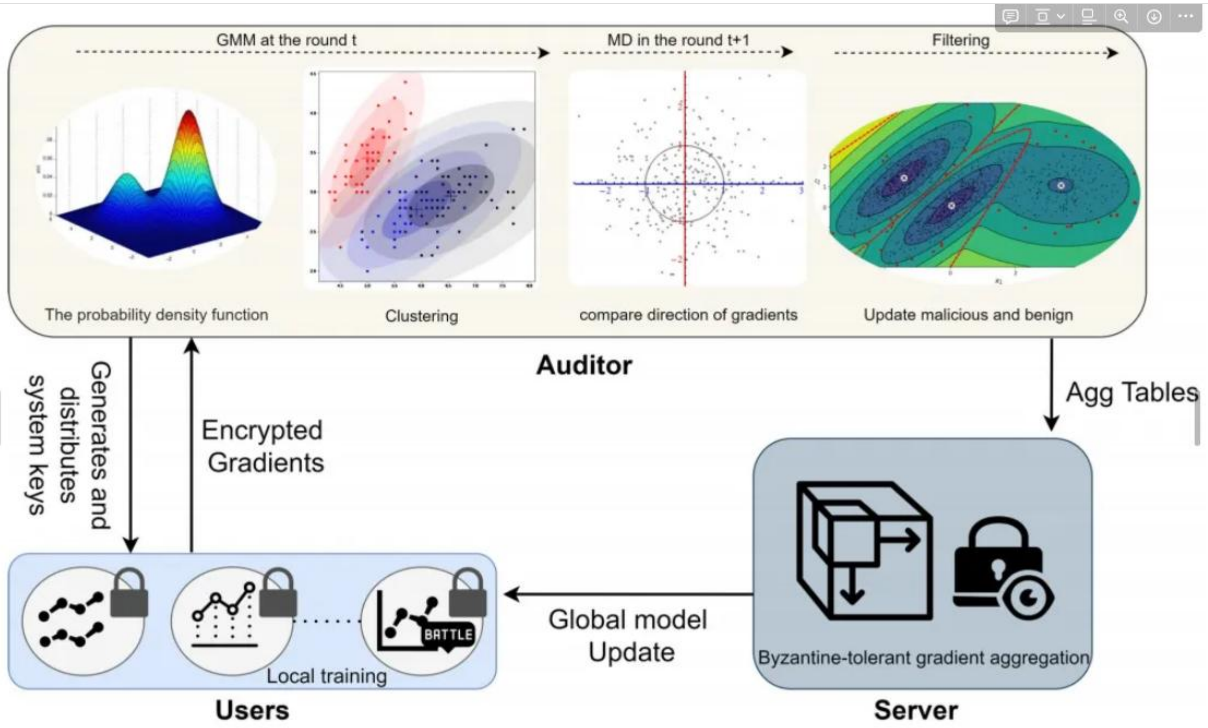


Fig. 2. Technical synopsis of the proposed system model.

2、模型技术

实现内部审计主要分为两大步骤：利用GMM和MD进行梯度分析以及基于AHE的加密与安全计算，具体实现如下：

1. 利用GMM和MD进行梯度分析

- 审计员将本地模型mi和全局模型M作为向量实体进行存储。由于本地提交的梯度为多维向量，审计实体采用高斯混合模型（GMM）来分析梯度分布，以此区分真实梯度和对抗性梯度（分类）。（GMM特别适用于处理非独立同分布（non - IID）数据中的复杂梯度分布情况，它能够将分布建模为多个高斯分布的混合，有效捕捉数据的异质性，从而更准确地识别出与正常学习趋势偏离的梯度，即可能的恶意梯度。）

- 马氏距离（MD）在该模型中被用作强大的异常值检测工具。MD考虑了数据集中的相关性和尺度不变性，能够衡量一个梯度与预期分布之间的偏离程度。通过计算每个梯度到其所属簇（由GMM确定的正常或恶意簇）均值的距离，审计实体可以有效识别出那些偏离正常分布的异常梯度，从而降低将良性梯度误判为恶意梯度的风险，尤其在non - IID数据场景中，这种方法能够更精确地监测和处理梯度数据。

2、模型技术

文章中Paillier密码系统的实现包括密钥生成、加密和解密过程，具体如下：

1. 密钥生成

首先，选择两个大素数a和b，计算它们的乘积 i=ab，i即为模数。接着，计算欧拉函数$\psi(i) = (a-1)(b-1)$和Carmichael函数 $\eta(i) = lcm(a-1, b-1)$.（其中lcm表示最小公倍数）。需要确认$gcd(i, \psi(i)) = 1$(gcd表示最大公约数)，若不满足，则重新选择模数i。然后，在$Zi2$中随机选择一个整数 f，使得i能整除 f在 $Zi2$中的阶。最后，Paillier公钥设置为 (i,f)，私钥设置为 $(\eta(i), \beta)$。若 $f = r + 1$（其中 r=ab），则 $\xi(i) = \eta(i)$，且 $\beta = \mu(i) - 1 \pmod i$，这种情况下可简化密钥生成过程。

2. 加密过程

对于要加密的消息 $k \in Zi$，加密过程如下：密码软件或硬件中的加密安全随机数生成器产生一个随机值 $n \in Zi$，n 作为随机化因子。使用Paillier加密算法计算密文

$$E(k) = c = f^n r^m \pmod{i^2}$$

其中r=ab,f和r是公钥的元素。

3. 解密过程

当需要解密时，密码软件或硬件执行以下操作：输入密文 $c \in Zi*2$。使用Paillier解密算法计算原始消息 ，计算公式为：

$$k = L\left(c^{\mu_i} \pmod{i^2}\right) \times \beta \pmod{i}$$

其中 β是私钥的一部分， L和 μi 是在Paillier密码系统中定义的函数。

2、 模型技术

文章中的系统设置主要涉及四个关键步骤，用户注册、密钥分发，模型分发以及审计表的生成，具体如下：

1. 用户注册（**Registration**）

对希望参与联邦学习（**FL**）的用户进行登记管理。用户向审计实体表明参与意愿，审计实体负责记录用户的相关信息。随后审计员创建并分发密钥，包括公钥和私钥份额，并为每个用户分配一个**ID**。

2. 密钥分发（**Key Distribution**）

审计实体为每个用户生成一对非对称密钥（**pk,qk**），即公钥**pk**和私钥**qk**。公钥用于用户加密本地梯度，确保数据在传输过程中的保密性；私钥则由审计实体保留，用于后续解密和验证操作。（例如，通过特定的密码学算法（如**Paillier**密码系统）生成密钥对，以保障加密和解密的有效性和安全性。

3. 模型分发（**Model Distribution**）

初始时，审计实体使用自己的公钥对初始全局模型进行加密，生成加密后的全局模型 **M0**。将加密后的全局模型 分发给每个注册用户。用户接收到加密模型后，使用自己的私钥进行解密，得到初始全局模型，从而可以在本地基于该模型和自己的数据集进行训练。

4.审计表生成（**Audit Table Generation**）

对于每个用户 **n**，审计实体使用其公钥**pkn**生成一个加密表（**Encrypted_Table** ）。该加密表用于存储用户特定数据的加密表示，例如用户在训练过程中产生的梯度更新或模型参数等信息。

Algorithm 1 System Setup

Require: Number of users N

User registration, key and model distribution, and audit table creation.
Step 1: Register users and assign unique IDs
for user $n \in \{1, 2, \ldots, N\}$ **do**
  Register user n and assign a unique ID Tag_ID_n
end for
Step 2: Distribute keys
for user $n \in \{1, 2, \ldots, N\}$ **do**
  Generate a Paillier cryptosystem key pair (pk_n, sk_n)
end for
Step 3: Distribute the initial global model $M(0)$ to users
Encryption $E = (KeyGen, Enc, Dec)$ with AHE:

- $Enc_{pk}(x_1) \cdot Enc_{pk}(x_2) = Enc_{pk}(x_1 + x_2)$
- $Enc_{pk}(x_1)^t = Enc_{pk}(t \cdot x_1)$

Step 4: Generate the audit table
Initialize an empty $Update\_Table$
for user $n \in \{1, 2, \ldots, N\}$ **do**
  Generate an $Encrypted\_Table_n$ for user n using their public key pk_n
  $(Tag\_ID_n, Encrypted\_Table_n)$ to the $Update\_Table$
end for
Internal auditing and make the $Update\_Table$

2、 模型技术

模型优化算法

文章中的优化算法主要包括去除冗余、减少通信轮数和降低计算成本三个方面

$$M_{\text{new}} = M_{\text{old}} + \eta \sum_{g \in G} f(g) \cdot g$$

$$t \text{ be the ratio} = \frac{\text{local updates after optimization}}{\text{local updates before optimization}}$$

Algorithm 3 Optimization of PPFL Model

Procedure: OPTIMIZEPPFL($M_{\text{old}}, G, n, r, AHE, t$)
Initialization: $M_{\text{new}} \leftarrow M_{\text{old}}$
Optimization 1: Removing redundancy
for $g \in G$ **do**
  $f(g) \leftarrow$ frequency of gradient g
  // Hash each gradient g to identify duplicates
  // Create a set G_{unique} of unique gradients
  // Use only gradients in G_{unique} for the next step
  $M_{\text{new}} \leftarrow M_{\text{new}} + \eta f(g) \cdot g$ // For $g \in G_{\text{unique}}$
end for
Optimization 2: Reducing the communication round
$t \leftarrow$ ratio of local updates
$r' \leftarrow \frac{r}{t}$
Optimization 3: Reducing the computing cost
Choose an efficient AHE such that $C'(M) < C(M)$
return M_{new}, r', AHE

内容纲要

1、 背景介绍

2、 模型技术

3、 实验验证

3、 实验验证

实验环境配置

实验使用Ubuntu 20.04操作系统，并基于PyTorch和Python 3.9构建实验环境。使用了Paillier库来配置联邦学习系统中的密码学原语。

数据集选择

实验选用MNIST、KDDCup和Amazon数据集。在处理非独立同分布（non - IID）数据时，对于每个数据集采用了特定的方法。以MNIST数据集为例，在实验中采用标签倾斜（label skew）方法，即每个用户仅存储来自单个类别的数据样本。而在独立同分布（IID）情况下，数据则均匀地分配给各个用户，以对比模型在不同数据分布条件下的表现

训练参数设置

针对不同的数据集，根据其特点定制了相应的训练参数，以优化模型的训练过程。对于MNIST和KDDCup数据集，设置训练迭代次数为300次，批处理大小为50；对于Amazon数据集，由于其具有较高的特征数量和较少的类别样本，将训练迭代次数设置为60次，批处理大小设置为10。MNIST和KDDCup数据集的学习率设置为0.01，而Amazon数据集的学习率设置为0.005。

为了评估我们的方法的性能，我们利用加密的本地梯度模拟了一系列模型投毒攻击。这些模拟研究了不同程度的对抗性存在，用不同的攻击率α表示 , 我们将其设置为 10%, 20%, 30%和 50%。

3、 实验验证

准确性分析

评估指标及含义

目标准确性（**Target Accuracy, T_acc**）：在针对性攻击场景下，**T_acc**反映了模型对目标标签（即攻击者试图影响的特定类别数据）的测试性能，用于衡量模型在处理被攻击特定类别数据时的分类准确性。（例如，在模拟对**MNIST**数据集中特定类别的攻击时，**T_acc**可以准确显示模型对这些目标类别的分类正确程度，直观反映模型在针对性攻击下对关键类别数据的处理能力。）

其他标签准确性（**Other Label Accuracy, O_acc**）：评估模型在非源和非目标标签（即除了攻击者针对的类别之外的其他类别数据）上的学习性能，用于检测模型在处理其他正常类别数据时是否受到攻击的干扰而出现误分类等问题。通过分析**O_acc**，可以全面了解模型在针对性攻击下对整体数据分类性能的影响，判断模型是否在保护目标类别数据的同时，依然能够准确处理其他类别数据。

总体准确性（**Overall Accuracy, ACC**）：**ACC**衡量模型在所有标签上的平均学习准确性，是一个综合反映模型在整个数据集上分类能力的指标。它综合考虑了模型对各类别数据的分类情况，不受攻击目标的限制，能够提供一个关于模型在面对各种攻击和数据分布情况下整体性能的直观度量。例如，在不同数据集和攻击场景下，**ACC**可以帮助评估模型的防御方法是否在整体上有效地抵御了中毒攻击，保持了较高的分类准确性。

3、 实验验证

实验结果分析

TABLE I

EVOLUTION OF ACCURACY IN NON-IID SETTINGS

Datasets	Baseline				Proposed Model			
	Targeted Attack		Untargeted Attack		Targeted Attack		Untargeted Attack	
	T_acc	Oracy	ACC		T_acc	Oracy	ACC	
MNIST	0.078	0.89	0.81	0.784	0.987	0.979	0.965	0.963
KDDCup	0.015	0.94	0.782	0.761	0.995	0.992	0.989	0.967
Amazon	0.015	0.93	0.771	0.743	0.994	0.997	0.998	0.996

在非独立同分布（non - IID）设置下，对MNIST、KDDCup和Amazon数据集进行了针对性和非针对性攻击实验（攻击率为50%）。以MNIST数据集为例，基线模型在针对性攻击下的T_acc为0.078，O_acc为0.89，ACC为0.81；而所提出模型的T_acc显著提高到0.784，O_acc达到0.987，ACC为0.979。在非针对性攻击中，所提出模型的ACC也达到了0.963，相比基线模型有很大提升。类似地，在KDDCup和Amazon数据集的实验中，所提出模型同样表现出色，在针对性和非针对性攻击下的各项准确性指标均明显优于基线模型。这表明所提出模型在non - IID数据场景下，无论是对目标类别还是其他类别数据的分类准确性都有显著提升，能够有效抵御中毒攻击对模型准确性的破坏。

3、 实验验证

实验结果分析

TABLE II

EVOLUTION OF ACCURACY IN IID SETTING

Datasets	Baseline				Proposed Model			
	Targeted Attack		Untargeted Attack		Targeted Attack		Untargeted Attack	
	T_acc	Oracy	ACC		T_acc	Oracy	ACC	
MNIST	0.09	0.2	0.15	0.11	0.989	0.995	0.97	0.968
KDDCup	0.049	0.6	0.58	0.52	0.997	0.995	0.992	0.991
Amazon	0.06	0.51	0.4	0.235	0.995	0.989	0.921	0.941

在独立同分布（IID）设置下，同样对不同数据集进行了针对性和非针对性攻击实验。结果显示，所提出模型在各种攻击场景下的准确性指标依然优于基线模型。例如，在MNIST数据集中，针对性攻击下所提出模型的T_acc为0.989，ACC为0.97；非针对性攻击下ACC为0.968。这说明模型在IID数据分布情况下，也能保持较高的准确性，不受攻击影响，进一步证明了模型的有效性和稳定性。

3、 实验验证

鲁棒性分析

评估方法：

通过分析模型在不同数量用户（$|U| \in [10,23,50]$ ）和不同攻击率（从10%到50%）下的准确性轨迹来评估模型的鲁棒性

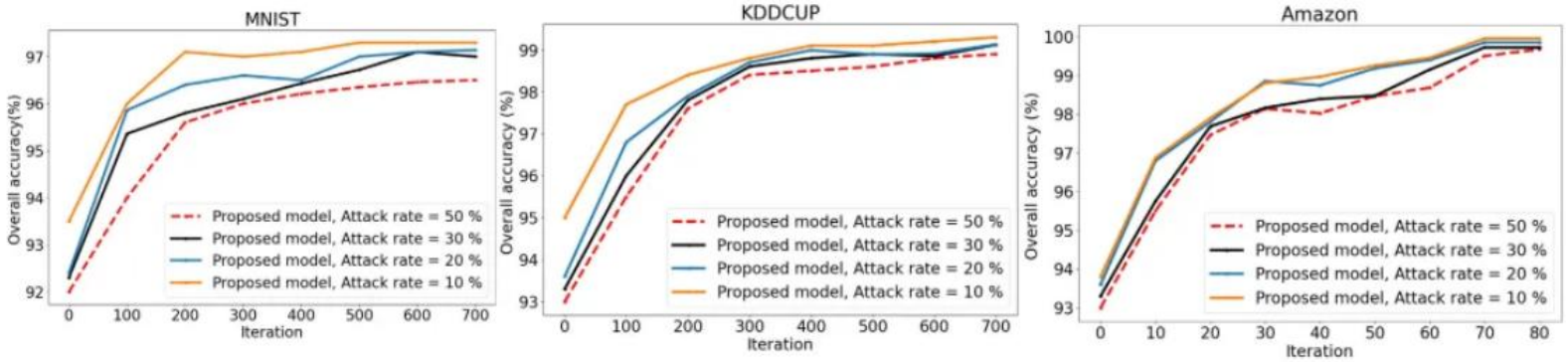
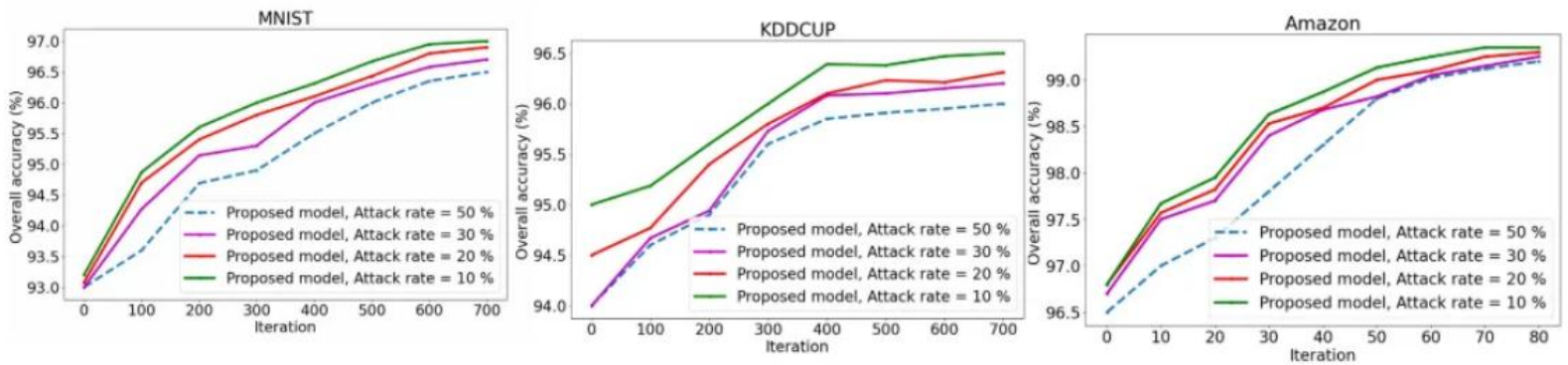


Fig. 4. Performance with targeted attack and non-IID data.



3、 实验验证

在non - IID设置下，以MNIST数据集为例，随着攻击率从10%增加到50%，所提出模型在不同用户数量下的准确性轨迹相对稳定。这表明模型在面对不同程度的攻击时，能够保持较为稳定的性能，不会因为攻击强度的增加而出现大幅度的准确性下降，体现了模型对恶意攻击的较强抵抗能力。在KDDCup和Amazon数据集的实验中也观察到了类似的结果（图4和图5），进一步证明了模型在non - IID数据环境下针对不同攻击强度和用户数量变化时的鲁棒性。

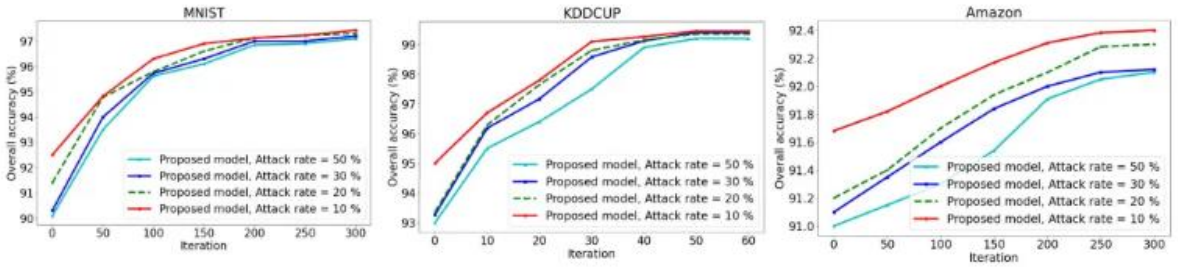


Fig. 6. Performance with targeted attack and IID data.

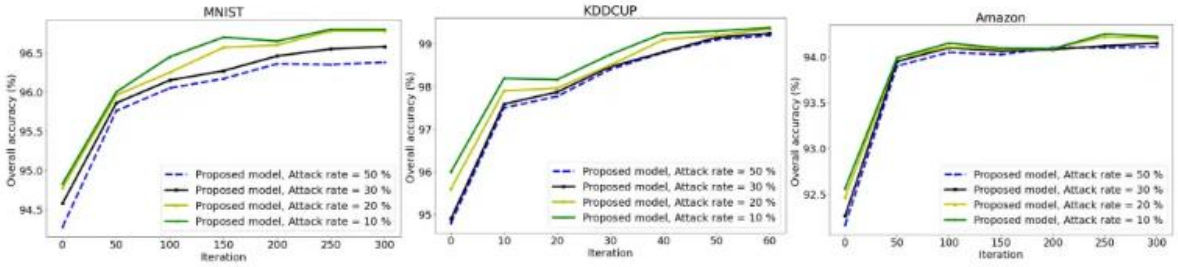


Fig. 7. Performance with targeted attack and IID.

3、 实验验证

恶意警报（MA）分析

评估指标及方法：

使用接收者操作特征（**Receiver Operating Characteristic, ROC**）曲线来评估防御算法在联邦学习系统中识别恶意活动的有效性。**ROC**曲线以假阳性率（**False Positive Rate**）为横坐标，真阳性率（**True Positive Rate**）为纵坐标，通过绘制不同阈值下的分类结果，能够全面展示模型在区分良性和恶意梯度时的性能。曲线下面积（**Area Under the Curve, AUC**）越大，表示模型在检测恶意活动方面的性能越好。

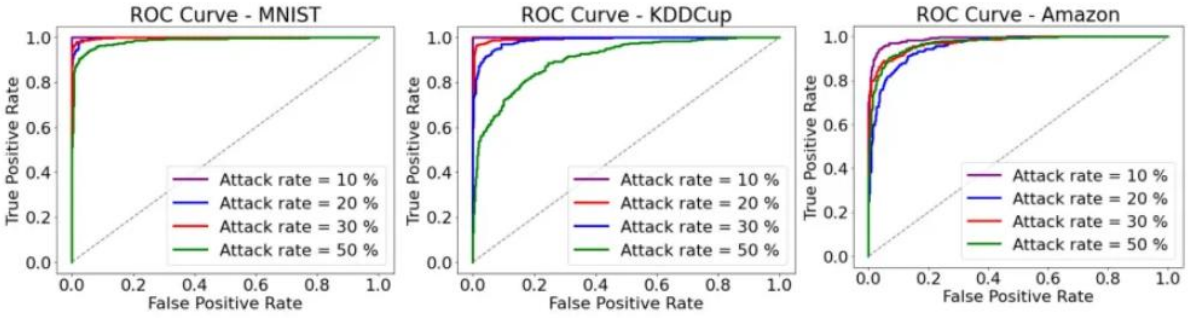
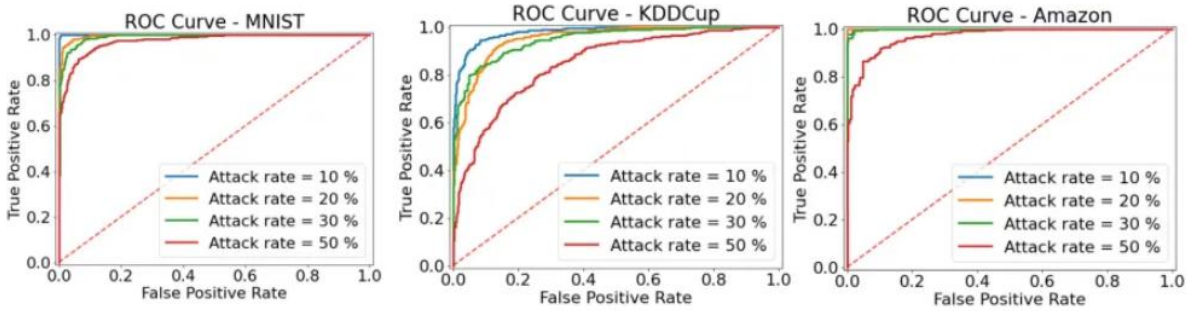


Fig. 8. Analyzing malicious alarms on IID data.



3、 实验验证

在MNIST、KDDCup和Amazon数据集的独立同分布（IID）和非独立同分布（non - IID）设置下，对模型进行了实验（图8和图9）。实验结果表明，所提出模型在不同数据集和数据分布设置下的ROC曲线表现良好，AUC值较高。例如，在MNIST数据集的实验中，模型在不同攻击比例下的ROC曲线均显示出较高的真阳性率和较低的假阳性率，AUC值较高，这意味着模型能够准确地区分良性和恶意梯度，有效检测出系统中的恶意活动。在KDDCup和Amazon数据集的实验中也得到了类似的结果，进一步证明了模型内部审计机制在识别恶意梯度方面的准确性和可靠性。

3、 实验验证

TABLE III

ACCURACY COMPARISON BETWEEN EXISTING SCHEMES

Work	Attack		Sitting
	Targeted	Untargeted	
Trimmed-means	52.7	53.4	non-IID
	82.3	85.7	IID
Krum	71.4	78.3	non-IID
	90.7	90.4	IID
Auror	32.5	49.9	non-IID
	89.2	87.8	IID
PEFL	46.4	57.4	non-IID
	88.2	81.7	IID
Sybils	89.7	89.5	non-IID
	74.5	70.8	IID
FL-trust	37.8	43.4	non-IID
	75.1	69.7	IID
ShieldFL	90.1	89.4	non-IID
	90.3	90.2	IID
Propose model	96.5	96.3	non-IID
	97	96.8	IID

为了证明我们的模型的有效性，我们将其准确性与几种现有的方案进行了比较，包括**Trimmed-means**、**Krum**、**Auror**、**PEFL**、**Sybils**、**FL-trust**和**ShieldFL**。比较使用了**MNIST**数据集，攻击率为**50%**，并进行了**500**次训练迭代。同时考虑了**IID**和非**IID**的数据设置。如表三所示，在**IID**和非**IID**数据设置下，我们的模型在靶向和非目标攻击方面都优于现有的方案。特别是，在非**IID**设置中，它对目标攻击（**96.5%**）和非目标攻击（**96.3%**）都获得了显著更高的准确性。类似地，我们的模型在**IID**设置中表现出出色的性能，目标攻击的准确率为**97%**，对非目标攻击的准确率为**96.8%**。当精度低于**60%**，攻击率设置为**50%**时，表现出次优性能。