

Securing Decentralized Federated Learning: An Integrated Approach with Blockchain, Trusted Execution Environment, and Internal Attack Detection

Youheng He, Xiaoxiao Wu, Taotao Wang

College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China

Abstract

Federated learning is a distributed learning paradigm that ensures data privacy and security by training models locally with data on client devices and aggregating models on an aggregation server. Although Federated learning can reduce privacy risks, the existence of aggregation servers still introduces a centralized structure to Federated learning, which often leads to security issues. To address security issues in traditional centralized Federated learning, this paper first proposes a decentralized Federated learning framework that adopts a round-robin scheme to select a temporary aggregation server from all clients in each round. We utilize blockchain smart contract technology to achieve the random selection of temporary aggregation servers to avoid malicious intervention and early prediction. However, the system is still vulnerable to other attacks, so a strong privacy protection model is needed to protect the round-robin decentralized Federated learning. To achieve this, a three-layer security protection mechanism is introduced. We adopt blockchain and trusted execution environment as the first and second layers, respectively, to establish a trusted environment to protect the decentralized Federated learning from external attacks. The third layer combines an internal attack detection algorithm to combat potential internal threats. To overcome the limitations of existing Federated learning experiments, we use Raspberry Pi to build a practical experimental platform for the proposed Federated learning framework and implement the three-layer security protection mechanism on this Federated learning platform. This platform helps evaluate the performance of the Federated learning algorithm and provides a realistic testing environment for evaluating the effectiveness of the three-layer security protection mechanism.

Keywords: *Federated Learning; Blockchain; Trusted Execution Environment; Internal Attack Detection; Decentralization*

去中心化联邦学习的安全防护：融合区块链、可信执行环境与内部攻击检测的方法

何有桁，吴晓晓，王滔滔

深圳大学电子与信息工程学院，广东 深圳，518060

摘要：联邦学习作为一种分布式学习范式，通过在客户端设备上使用本地数据训练模型，并将模型聚合到聚合服务器上，从而确保数据隐私和安全。尽管联邦学习可以降低隐私风险，但聚合服务器的存在仍然为联邦学习引入了集中式结构，这通常会导致安全问题。为了解决传统集中式联邦学习中的安全问题，本文首先提出了一种去中心化的联邦学习框架，该框架采用轮询机制，在每一轮中从所有客户端中选择一个临时聚合服务器。我们利用区块链智能合约技术来实现临时聚合服务器的随机选择，以避免恶意干预和提前预测。然而，该系统仍然容易受到其他攻击，因此需要一个强大的隐私保护模型来保护这种轮换式去中心化联邦学习。为此，我们引入了一种三层安全保护机制。我们分别采用区块链和

可信执行环境作为第一层和第二层，以建立一个可信环境来保护去中心化联邦学习免受外部攻击。第三层结合了内部攻击检测算法来应对潜在的内部威胁。为了克服现有联邦学习实验环境的局限性，我们使用树莓派为所提出的联邦学习框架构建了一个实用的实验平台，并在该平台上实现了三层安全保护机制。该平台有助于评估联邦学习算法的性能，并为评估三层安全保护机制的有效性提供了一个真实的测试环境。

关键词：联邦学习；区块链；可信执行环境；内部攻击检测；去中心化

1 引言

联邦学习通过数据不出本地的方式缓解了数据共享带来的隐私泄露问题，已经在医疗、交通、和物联网等场景中得到广泛关注^{[1][2]}。但传统联邦学习通常仍依赖中心聚合服务器完成全局模型的更新，这使系统在结构上保留了明显的中心化特征。一方面，聚合服务器一旦失效、被入侵或遭受拒绝服务攻击，训练过程将受到阻塞，甚至导致全局模型被攻击；另一方面，中心服务器对模型收集与广播过程具有绝对控制权，也使联邦学习天然面临单点信任问题。与此同时，恶意客户端还可能通过构造异常梯度或上传污染模型干扰全局模型收敛^{[3][4]}。

为减轻中心服务器可能带来的安全风险，已有研究开始探索去中心化联邦学习的实现方式，相关思路主要包括基于点对点通信的协同更新，以及基于区块链的模型管理机制^{[5][6]}。前一种方式依靠节点之间直接交换信息完成协作，后一种方式则将部分原本由中心服务器承担的功能交由智能合约、链上状态记录和共识机制实现。这些方法在不同的场景下各有优势和安全性局限性，因此需要根据具体需求进行仔细选择和优化。因此，进一步探索去中心化的联邦学习方法并解决潜在的安全问题，对于推动联邦学习的发展和支持其在各个领域的广泛应用至关重要。

为了构建能够抵御外部和内部攻击的安全去中心化联邦学习，我们将区块链、可信执行环境和内部攻击检测算法的技术融合到一个集成方法中。本文的具体贡献包括：

- 1) 构建了一种轮换式去中心化联邦学习框架，利用区块链智能合约在每轮训练中随机选择临时聚合节点，从而解决固定中心服务器带来的中心化结构问题。
- 2) 设计了一种由区块链身份认证、TEE 远程证明以及内部攻击检测组成的三层安全机制，有效地保护去中心化联邦学习系统免受外部和内部威胁。
- 3) 基于树莓派搭建原型系统，并在真实设备环境下对训练性能、隐私保护效果以及内部攻击检测能力进行了实验验证。

2 相关工作

联邦学习安全研究主要集中在隐私保护、可信执行和去中心化架构三个方向。在隐私保护方面，差分隐私通过在局部梯度或模型参数中添加噪声来抵御模型反演攻击，是目前应用最广泛的隐私保护方法之一^[7]。进一步地，同态加密和安全多方计算能够在不暴露明文参数的情况下完成聚合，但常伴随较高的计算与通信开销，因此在资源受限的边缘设备上部署难度较大^[8]。从工程实现角度看，如何在精度损失、系统开销和隐私强度之间取得平衡，仍是联邦学习隐私保护研究的核心问题。

在可信执行方面，TEE 为敏感代码和数据提供硬件级隔离环境，使本地训练能够在相对可信的运行空间内完成。相较于纯密码学方案，TEE 不需要在每次聚合中执行大量加密与解密，因此在性能开销上更适合边缘设备部署。已有研究已将 TEE 用于联邦学习聚合和本地训练保护，并通过远程证明机制验证参与方执行环境的完整性^{[9][10]}。不过，TEE 主要解决的是本地训练是否运行于可信环境这一问题，对于已经通过远程证明但仍提交恶意更新的内部节点，其防护能力有限。

在去中心化架构方面，点对点通信机制和区块链机制是两条典型路线。点对点机制强调节点间直接协作，适合动态组网和无中心管理场景；区块链机制则通过智能合约和分布式账本为模型状态管理、访问控

制和行为审计提供支持^{[5][6][11]}。尽管去中心化联邦学习具有诸多优势，但仍存在潜在的风险和挑战，例如对抗节点可能会引入有害模型，且假设半诚实的参与者是攻击者在现实世界中可能过于乐观。因此，有必要构建多层协同防护机制，以提升去中心化联邦学习系统的整体鲁棒性。

3 轮换式去中心化联邦学习框架

本文提出的轮换式去中心化联邦学习框架取消固定中心聚合服务器，由所有客户端轮流担任临时聚合节点。每轮训练开始前，系统通过区块链智能合约从客户端集合中随机选取一个节点作为本轮临时聚合者，其职责包括接收其余客户端的模型更新、完成聚合并广播新的全局模型。该机制一方面避免了单一中心节点长期掌控聚合过程，另一方面借助区块链的公开可审计特性提高了聚合节点选取过程的透明性。

如图 1 所示，我们展示了基于区块链的轮换式去中心化联邦学习框架在医疗场景中的示意图。该框架无需中央聚合服务器即可运行，所有客户端相互通信并以轮换方式充当临时聚合服务器。在每次迭代中，随机选择一个客户端作为聚合服务器，负责协调其他客户端的本地模型更新，并将它们组合成一个全局模型。为了确保临时聚合服务器选择的公平性和公正性，并防止某些节点被过度选择或排除，我们采用区块链智能合约来实现临时聚合服务器的随机选择。具体而言，在每次迭代之前，智能合约会随机选择一个客户端作为临时聚合服务器，所有参与者都可以查询选择结果。这种方法确保了临时聚合服务器选择过程的透明性和公平性，最大限度地降低了人为操纵或数据篡改的风险。

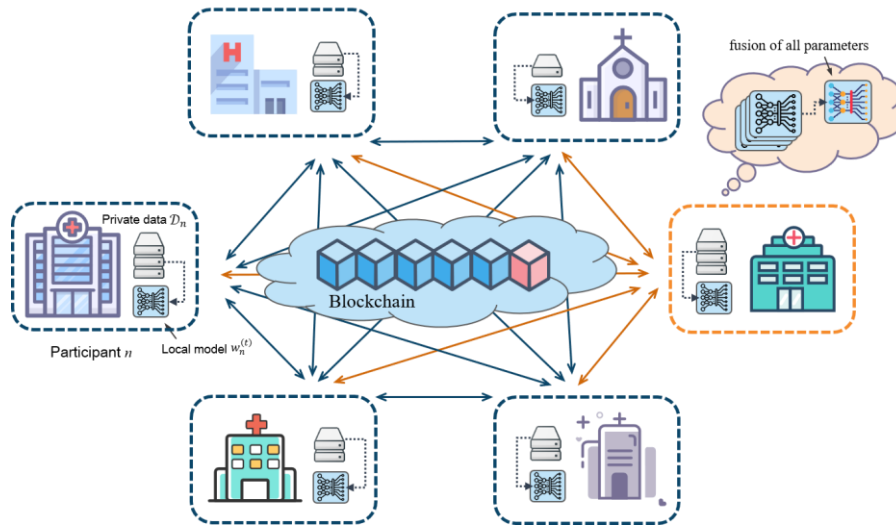


图 1 轮换式去中心化联邦学习框架示意

在训练流程上，各客户端接收上一轮全局模型后使用本地数据独立完成训练，并将本地更新发送给本轮临时聚合节点。聚合完成后，聚合者向各客户端广播更新后的全局模型。为避免聚合节点向不同客户端下发不一致结果，可在广播前将全局模型摘要信息写入链上，客户端在接收模型后再与链上记录进行一致性校验，从而提高模型分发过程的可追踪性。

4 三层安全防护机制

图 2 展示了所提出的三层安全防护机制的架构，该机制由以下三个独立层组成：基于区块链的身份认证，TEE 远程证明以及内部检测和定位算法。三层防护机制主要内容如下。

4.1 威胁模型

本文考虑两类攻击者。外部攻击者不具备合法系统凭据，但能够窃听信道、发起拒绝服务攻击；内部攻击者则是已经通过认证的参与者，其能够篡改本地数据或构造恶意模型更新以降低全局模型性能。本文假设攻击者无法破解标准密码学原语，也无法破坏区块链账本不可篡改性与 TEE 硬件完整性。在这一假设

下，区块链负责构建可信准入与可信记录，TEE 负责提供可信执行，而内部攻击检测机制则用于约束已经进入系统的恶意参与方。

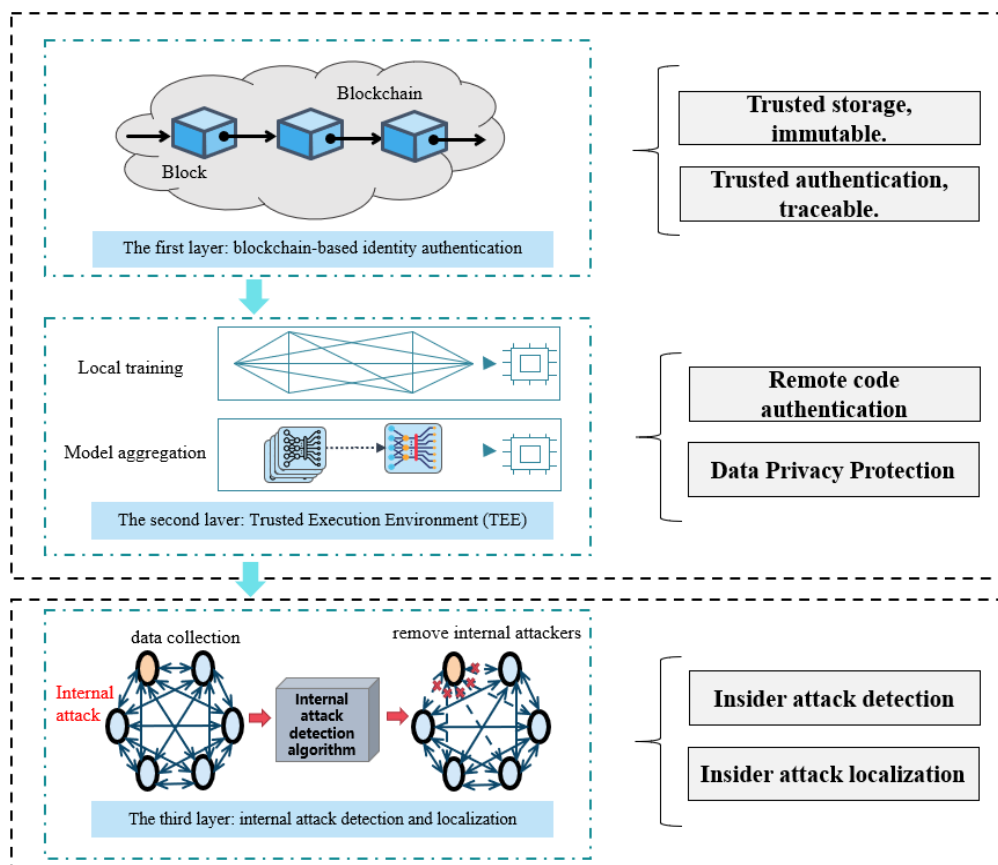


图 2 三层安全防护机制架构

4.2 第一层：基于区块链的身份认证

第一层保护用于阻断未授权外部节点进入系统。本文设计身份认证合约，用链上地址与设备信息建立映射关系，并维护授权状态表。管理员完成节点注册，其他节点可查询某一地址是否已授权及其基础网络信息；当某节点退出训练或被确认为恶意参与者时，可由管理员撤销其权限。借助链上记录，系统能够实现参与身份的可查询、可追踪和可审计。

与传统集中式认证相比，链上身份认证的优势在于结果对所有参与方公开一致，且无法被单个管理节点静默篡改。特别是在多组织协同训练场景中，这种可审计性能够降低准入过程的不透明风险，为后续的异常溯源提供可信依据。

4.3 第二层：TEE 远程证明

第二层保护用于确保本地训练过程运行在可信环境内。本文采用 SGX 类 TEE 机制，在客户端本地将敏感数据和关键训练代码置于受保护区域运行，并通过远程证明验证节点的真实性和完整性。为降低联机验证开销，原型系统采用 DCAP 类证明思路：临时聚合节点发起挑战，客户端生成证明报告并上传至区块链，聚合节点下载证明报告完成验证，再将验证结果写回链上。其他节点可进一步读取验证结果，用于建立对参与方执行环境的信任关系。

需要说明的是，TEE 主要解决的是代码是否在受保护环境内执行这一问题，并不能天然保证执行的代码没有恶意。因此，即便某节点通过远程证明，也不能完全排除其在允许范围内提交异常更新的可能，这正是第三层内部攻击检测机制存在的必要性。

4.4 第三层：内部攻击检测与定位

即使节点已完成身份认证和 TEE 验证，仍可能通过恶意数据或恶意更新实施内部攻击。因此，第三层保护聚焦于异常模型更新的识别。设共享模型中共包含 $\bar{K}$ 个参数，本文从中选取 K 个参数，对其在多轮训练中的变化轨迹进行统计，构造空间差异（spatial difference, SD）检测指标：

$$\varphi_{ij}^k = \sum_{t=0}^T (\mu_j^k(t) - \bar{\mu}_i^k(t)) \quad (1)$$

$$D_1^i = \frac{1}{|V_i|} \sum_{j \in V_i} \left(\frac{1}{K} \sum_{k=1}^K |\varphi_{ij}^k| \right)^2 \quad (2)$$

其中 V_i 表示节点 i 的邻居集合， $\bar{\mu}_i^k(t)$ 为其邻域内第 k 个参数在时刻 t 的均值。若 D_1^i 超过阈值 δ_l ，则判定网络中存在内部攻击。进一步地，可构造定位指标：

$$\widetilde{\varphi}_{ij}^k = \sum_{t=0}^T (\mu_j^k(t) - \mu_i^k(t)) - \varphi_{ii}^k \quad (3)$$

$$L_1^{ij} = \left(\frac{1}{K} \sum_{k=1}^K |\widetilde{\varphi}_{ij}^k| \right)^2 \quad (4)$$

当 L_1^{ij} 超过阈值 ϵ_l 时，将节点 j 识别为可疑节点。

该策略的直观含义是：在正常训练条件下，各客户端更新虽受本地数据差异影响，但总体仍围绕共同优化方向波动；若某个节点持续输出带有目标偏移的恶意更新，则其参数轨迹会与邻域平均状态出现累积偏离，这种偏离可被统计量捕获。

5 系统原型设计与实验评估

5.1 平台原型结构

为验证所提方法的可实现性，本文使用树莓派搭建去中心化联邦学习原型系统，每台设备作为一个参与客户端。区块链模块采用私有链实现，TEE 采用基于 SGX 的可信执行方案，训练、认证与检测流程均在该平台上完成。该原型平台既保留了真实边缘设备的资源约束特性，也为评估所提安全机制提供了可重复实验环境。图 3 给出了平台原型结构示意图。

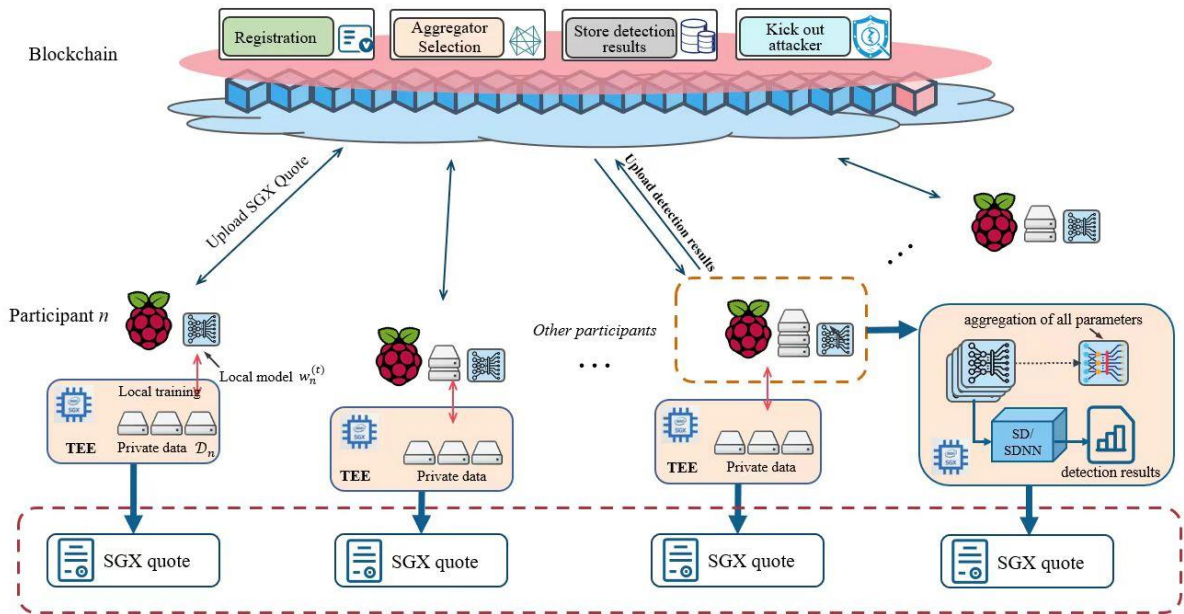


图 3 系统原型结构示意图

5.2 实验设置

实验选用 MNIST-10 和糖尿病视网膜图像（Diabetic Retinopathy Image，DRI）两个数据集。MNIST-10 使用卷积神经网络模型，客户端上传本地模型权重；DRI 使用 SqueezeNet 结构，客户端上传局部梯度。参与客户端数量均设为 10。对差分隐私相关方法，梯度裁剪阈值设为 $C = 2$ ，噪声尺度设为 $\sigma = 2$ ，采样概率为 $q = 0.03425$ ，动量系数为 $\beta = 0.9$ 。其中，MNIST-10 任务的学习率设为 0.002、批大小为 50、本地迭代次数为 5；DRI 任务的学习率设为 0.01、批大小为 50、本地迭代次数为 1。内部攻击场景中，设置 1 个恶意节点持续提交偏离正常分布的更新向量，以模拟数据注入和模型投毒行为。

5.3 身份认证与远程证明分析

在身份认证实验中，树莓派节点首先通过授权交易向链上提交设备地址、名称与网络信息，随后其他节点可通过查询交易读取其授权状态与连接信息。实验现象表明，认证合约能够稳定地完成节点注册和查询操作，并为后续联邦训练提供一致的参与方身份视图，表 1 和表 2 分别展示了注册节点和查询节点的交易信息，这些数据由区块链智能合约存储并用于授予设备相应权限。在远程证明实验中，客户端根据挑战生成证明报告，聚合节点下载并验证证明报告后，将验证结果写入区块链。该过程证明了链上记录与可信执行机制可以协同工作，既保持证明结果可追踪，也避免了纯中心化验证带来的可信争议。

表 1 注册节点

Key	Value
Transaction hash	0x98c3...abacc
From	0xAb848...15835cb2
To	Authentication.authorizeDevice (address,string,string,string)
Transaction cost	134067
Input	0x526 ... 00000
Decoded input	{"address deviceAddress": "0xAb8483...cb2", "string name": "node1", "string publicKey": "0x6bf287102 ... 9e67b8", "string ipAddress": "198.168.1.136"}
Decoded output	{"0": "bool: true"}

表 2 查询节点认证信息

Key	Value
From	0x5B38Da...eddC4
To	Authentication.authenticate(address)
Transaction cost	11496
Input	0x08e ... 35cb2
Decoded input	{"address deviceAddress": "0xAb8483...835cb2"}
Decoded output	{"0": "bool: true", "1": "tuple(string,bytes32,string): node1, 0x6bf287102 ... 9e67b8, 198.168.1.136"}

5.4 内部攻击影响与检测性能

从图 6 和图 7 可以看出，在不存在恶意节点时，联邦学习模型精度整体保持稳定；当引入内部攻击节点后，模型精度出现明显波动甚至下降，说明恶意更新确实能够持续拉偏全局模型收敛方向。对于 MNIST-10，在模型权重上传场景下，攻击发生后全局精度可能在高低区间之间大幅振荡；对于 DRI，在梯度上传场景下，模型精度在短暂上升后逐步下降并在较低区间波动。这说明内部攻击在不同聚合对象下表现形式不同，但都会对训练稳定性产生实质影响。但引入我们的三层防护模型后，系统可以有效检测出恶意节点，从而去除恶意攻击带来的负面效果

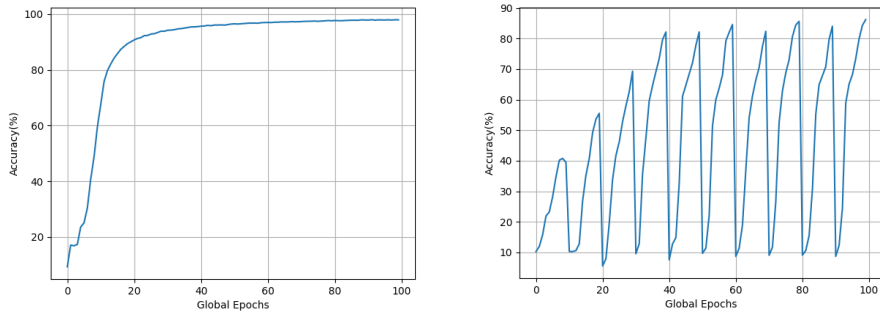


图 6 MNIST-10 数据集内部攻击影响与检测结果示意：无攻击精度：被攻击精度

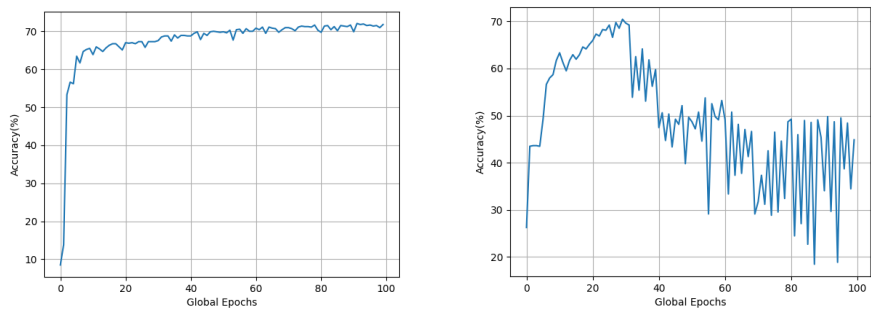


图 7 DRI 数据集内部攻击影响与检测结果示意：无攻击精度：被攻击精度

如图 8 和图 9 所示，基于 ROC 曲线的评估结果进一步表明，SD 方法均能较好地检测区分正常场景与攻击场景，以及定位恶意节点。并且当观测参数数量从 $K = 1$ 增加到 $K = 5$ 时，算法性能进一步提升。这说明增加观测维度有助于揭示恶意更新在不同参数位置上的累积偏移。

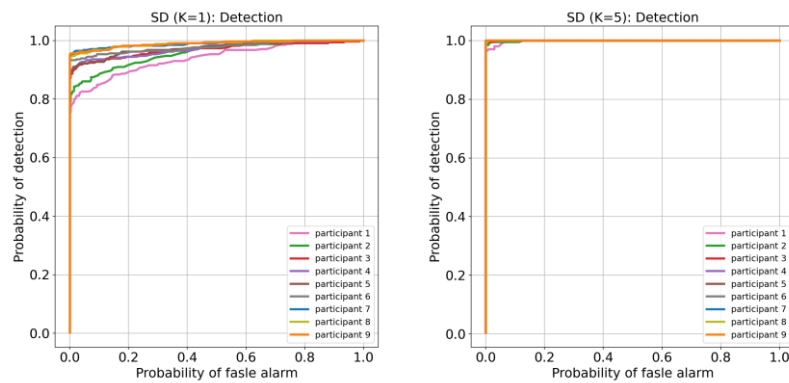


图 8 MNIST-10 数据集下的 SD 检测的 ROC 曲线

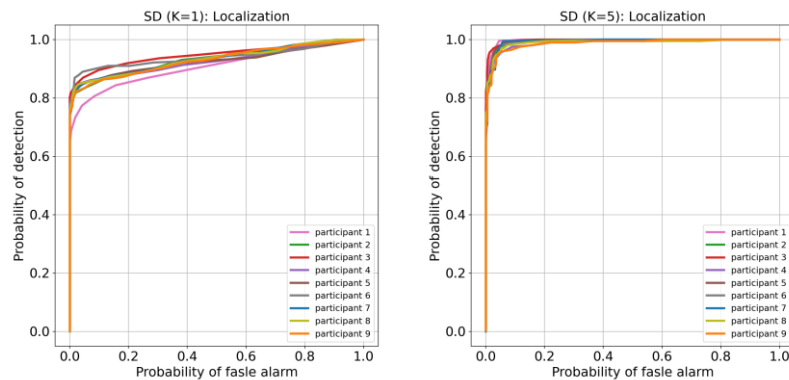


图 9 MNIST-10 数据集下的 SD 定位的 ROC 曲线

6 结论

本文面向去中心化联邦学习场景，提出了一种融合区块链、可信执行环境和内部攻击检测的综合安全防护方法。通过轮换式临时聚合机制，系统弱化了固定中心聚合服务器带来的结构性风险；通过三层安全机制，系统分别从身份认证、可信执行和内部异常识别三个层面提升了整体安全性；通过树莓派原型平台实验，验证了所提方法在模型训练性能、差分隐私保护和攻击检测效果上的可行性。

需要指出的是，内部攻击检测算法仍依赖于对异常更新模式的可区分性。当攻击者针对检测逻辑设计自适应绕过策略时，现有方法的效果可能下降。此外，TEE 在大规模异构设备环境中的部署成本仍然较高。后续可结合可验证计算、零知识证明和更轻量的链下协同机制，进一步提高系统在复杂真实场景中的可扩展性与通用性。

参考文献

- [1] Konečný J, McMahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv preprint arXiv:1610.05492, 2016.
- [2] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19.
- [3] Douceur J R. The sybil attack[C]//International workshop on peer-to-peer systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 251-260.
- [4] Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems[C]//European symposium on research in computer security. Cham: Springer International Publishing, 2020: 480-501.
- [5] Roy A G, Siddiqui S, Pölsterl S, et al. Braintorrent: A peer-to-peer environment for decentralized federated learning[J]. arXiv preprint arXiv:1905.06731, 2019.
- [6] Li Y, Chen C, Liu N, et al. A blockchain-based decentralized federated learning framework with committee consensus[J]. Ieee Network, 2020, 35(1): 234-241.
- [7] Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and trends® in theoretical computer science, 2014, 9(3-4): 211-487.
- [8] Acar A, Aksu H, Uluagac A S, et al. A survey on homomorphic encryption schemes: Theory and implementation[J]. ACM Computing Surveys (Csur), 2018, 51(4): 1-35.
- [9] Sabt M, Achemlal M, Bouabdallah A. Trusted execution environment: What it is, and what it is not[C]//2015 IEEE Trustcom/BigDataSE/Ispa. IEEE, 2015, 1: 57-64.
- [10] Ohrimenko O, Schuster F, Fournet C, et al. Oblivious {Multi-Party} machine learning on trusted processors[C]//25th USENIX Security Symposium (USENIX Security 16). 2016: 619-636.
- [11] Ramanan P, Nakayama K. Baffle: Blockchain based aggregator free federated learning[C]//2020 IEEE international conference on blockchain (Blockchain). IEEE, 2020: 72-81.