基于 NVIDIA FLARE 的支付交易隐私保护联邦欺诈检测

Holger R. Roth[*1]，Sarthak Tickoo[2], Mayank Kumar[5], Isaac Yang[1], Andrew Liu[1], Amit Varshney[2], Sayani Kundu[2], Iustina Vintila[4], PeterMadsgaard[4], JurajMilcak[4],Chester Chen[*1]，Yan Cheng[1], Andrew Feng[1],Jeff Savio[*1]，Vikram Singh[3], Craig Stancill[5], Gloria Wan[2], Evan Powell[5], Anwar Ul Haq[+*4]，Sudhir Upadhyay[+*2]，以及 Jisoo Lee[+*3]

[1]NVIDIA，圣克拉拉，美国[2]J.P. Morgan，纽约，美国[3]纽约银行，纽约，美国[4]加拿大皇家银行，多伦多，加拿大[5]DeepTempo，旧金山，美国

arXiv:2603.13617v1 [cs.LG] 13 Mar 2026

摘要

与欺诈相关的金融损失持续攀升，而监管、隐私和数据主权方面的限制，正日益降低集中式欺诈检测系统的可行性。联邦学习（FL）已成为一种颇具前景的范式，使各机构能够在不共享原始交易数据的情况下协同训练模型。然而，在现实的非独立同分布金融数据分布下，其实际效用仍未得到充分验证。

本文介绍了一项面向行业的多机构概念验证研究，评估了使用 NVIDIA FLARE框架进行支付交易联邦异常检测的效果。我们模拟了一个由异质金融机构组成的现实联邦，各机构观察到不同的欺诈类型模式，并严格执行数据隔离。我们采用通过联邦平均（FedAvg）训练的深度神经网络，结果表明，联邦模型的平均 F1 分数达到 0.903，显著优于本地训练模型（0.643），且接近集中式训练性能（0.925），同时完整保留数据主权。

我们进一步分析了收敛行为，结果显示，在 10 轮联邦通信轮次内即可取得良好性能，凸显了 FL 在对延迟和成本敏感的金融环境中的运营可行性。为支持在受监管环境中部署，我们使用基于Shapley的特征归因评估模型可解释性，并确认联邦模型依赖语义连贯、与领域相关的决策信号。最后，我们通过差分隐私随机梯度下降（DP-SGD）引入样本级差分隐私，并证明隐私–效用权衡良好：在欺诈检测性能仅中度下降的情况下，有效隐私预算低于 $\varepsilon = 10.0$ 。总体而言，这些结果提供了实证证据，表明 FL 能够实现有效的跨机构欺诈检测，在维持严格数据隔离并支持形式化隐私保证的同时，达到接近集中式的性能。

[*]通讯作者：{hroth,chesterc,jsavio}@nvidia.com；anwar.ulhaq@rbc.com；sudhir.x.upadhyay@jpmorgan.com；Jisoo@bny.com。
[+]这些作者贡献相同。

1 引言

欺诈检测仍是金融服务业（FSI）面临的核心挑战。近期分析指出，在美国，商户每发生 1 美元的欺诈活动，平均需承担 4.61 美元的成本[1]。随着数字交易量持续增长、欺诈手段日益复杂，金融机构面临越来越大的压力，必须部署自适应的数据驱动型检测系统[2,3]。然而，传统方法高度依赖集中式数据聚合；在现代监管约束、严格的数据主权要求以及各机构风险态势下，这种策略已越来越难以为继[4]。这些局限性长期以来阻碍了跨组织协作，导致检测能力分散，识别新兴欺诈类型模式的效果也有所下降。

联邦学习（FL）提供了一种颇具吸引力的替代方案，使各机构能够在不交换原始客户数据或交易数据的情况下训练共享的机器学习模型[5,6]，，如图 1所示。原则上，联邦学习能够在保障严格隐私保障和机构自主权的同时，促进各机构形成集体智能。然而，联邦学习在欺诈检测中的实际可行性仍是一个悬而未决的问题；欺诈检测所涉及的数据分布高度不平衡、非独立同分布，且因机构而异[7,8]。

在本研究中，我们使用 NVIDIA FLARE[9] (NVFlare) 联邦计算框架，对用于支付异常检测的联邦学习进行了全面的多方技术评估。我们的概念验证实验模拟了一个

具有异质欺诈特征的真实多机构环境，展示了通过联邦平均训练的深度神经网络[10] (FedAvg)能够达到接近集中式模型的性能（**0.903 vs 0.925**， ΔF1=0.022），同时显著优于仅本地训练（**0.643 → 0.903**， +40%）。值得注意的是，这一性能伴随着快速收敛——通常在几个联邦轮次内即可实现——并且在异常类型各异的客户端之间表现出强大的跨域泛化能力。

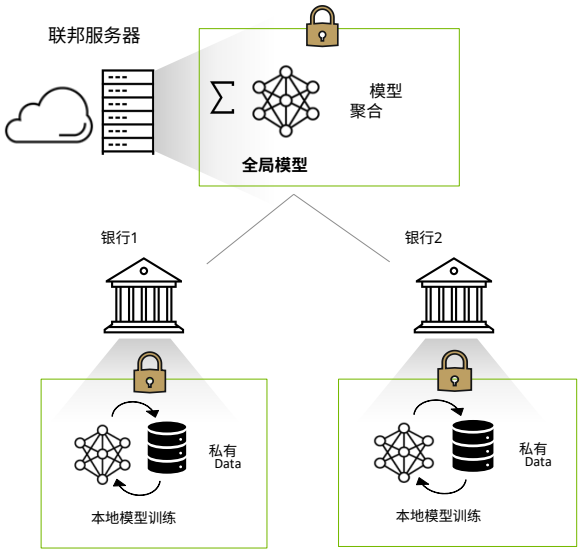


图 1. 用于协作式欺诈检测的联邦学习。

为确保模型在受监管的金融场景中值得信赖且适用，我们还使用基于 Shapley 值的特征归因可解释性分析[11,12]对训练所得模型进行了进一步验证。特征归因研究证实，联邦模型保留了可解释且语义连贯的决策路径，符合领域预期以及透明人工智能的监管要求[12]。

除了验证准确率和可扩展性外，我们还通过**差分隐私（DP）训练**[13]评估了增强隐私的机制，并采用安全采样策略实现样本级差分隐私。研究结果表明，私有联邦学习能够通过**通过子采样实现的隐私放大**[14]，显著降低有效隐私损失，同时保持具有竞争力的模型性能。这一结果凸显了将形式化隐私保证融入联邦欺诈检测流程的可行性，为实现符合监管要求的隐私保护提供了切实可行的路径。

总而言之，本研究提供了严谨的实证证据，表明联邦学习是一种切实可行、高性能且保护隐私的多机构欺诈检测范式。我们的贡献包括：

- 展示了多个模拟金融机构在不共享敏感数据的情况下开展协作式深度学习。
- 对 FedAvg 与集中式及仅本地训练进行了定量基准测试，并在非独立同分布场景中建立了稳健的性能基线。
- 通过基于 Shapley 值的归因分析验证模型的可解释性，从而提高透明度并促进监管合规。
- 评估了差分隐私技术，展现出良好的隐私—效用权衡和可衡量的隐私放大效果。
- 评估 NVFlare 在高安全性、分布式金融服务行业环境中的可扩展性、通信效率和运行准备度。

本研究表明，联邦学习在维护绝对数据主权的同时，能够实现接近集中式模型的模型性能，从而推动新一代协作式欺诈检测系统的发展，满足技术和监管方面的迫切需求。

2 结果

我们构建了一个纯合成数据生成框架，可在联邦站点之间控制统计异质性、组合基于规则的异常注入，并调整带有标签噪声的类别平衡（参见方法）。每家机构

都获得了一个欺诈类别分布各异的数据集，从而构建出能够代表现实世界机构间差异的非独立同分布环境。

每个客户端均使用三个数据集分区：

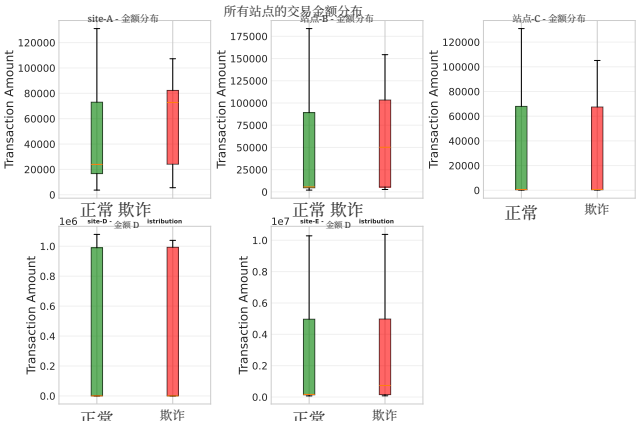
1. **缩放数据集：** 用于计算适用于所有异常类别的特征归一化参数。2. **训练数据集：** 用于本地联邦学习轮次。3. **测试数据集：** 保留用于本地模型验证和性能监控。

这种配置既确保了受控实验，又忠实地呈现了实际运行的欺诈检测系统中遇到的异质性。图 2c 和 2d 分别展示了训练集和评估集中不同参与银行的欺诈类型分布。为进一步刻画模拟联邦的异质性，我们考察了交易金额的分布特征以及各机构的欺诈发生率。图 2a 汇总了数据集的关键统计信息，包括总交易量、欺诈数量和欺诈率；这些数据共同凸显了参与客户端的非独立同分布特征，以及由此给联邦优化带来的挑战。欺诈率是根据应用标签噪声后最终的欺诈标签，在合并后的训练集+和测试集上计算得出的。作为补充，图 2b 展示了各站点欺诈交易与非欺诈交易的美元支付金额差异，凸显了机构间在规模和离散程度方面的显著差异。

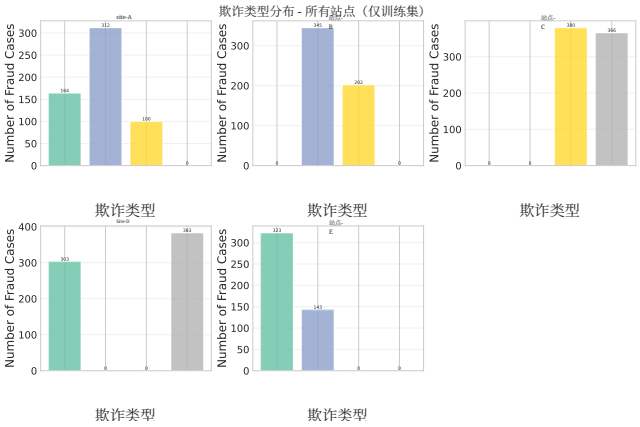
2.1 数据配置

(a) 参与银行的欺诈率，其中站点-E的欺诈率最高。平均交易金额以千美元为单位，总金额以十亿美元为单位。

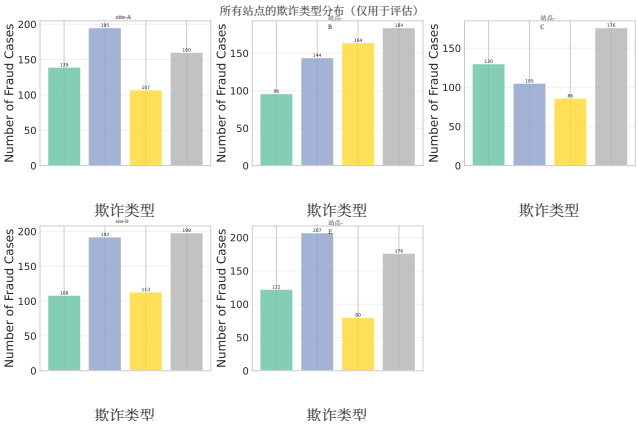
Site	总计 记录数	欺诈 数量	Rate (%)	Avg 金额	总计 金额
Site-A	200,000	1,077	0.54	$40.28K	$8.06B
站点-B	200,000	1,060	0.53	$36.99K	$7.40B
站点-C	200,000	1,162	0.58	$25.28K	$5.06B
Site-D	200,000	1,196	0.60	$336.98K	$67.40B
站点-E	200,000	995	0.50	$1768.23K	$353.65B



(b) 正常交易和欺诈交易的美元金额。



(c) 欺诈类型（训练集）。



(d) 欺诈类型（评估集）。

图 2。 跨站点的数据分布。A、站点级欺诈率和金额汇总统计。**B、**正常交易与欺诈交易的金额分布。**C–D、**训练集和评估集中异常类型分布。

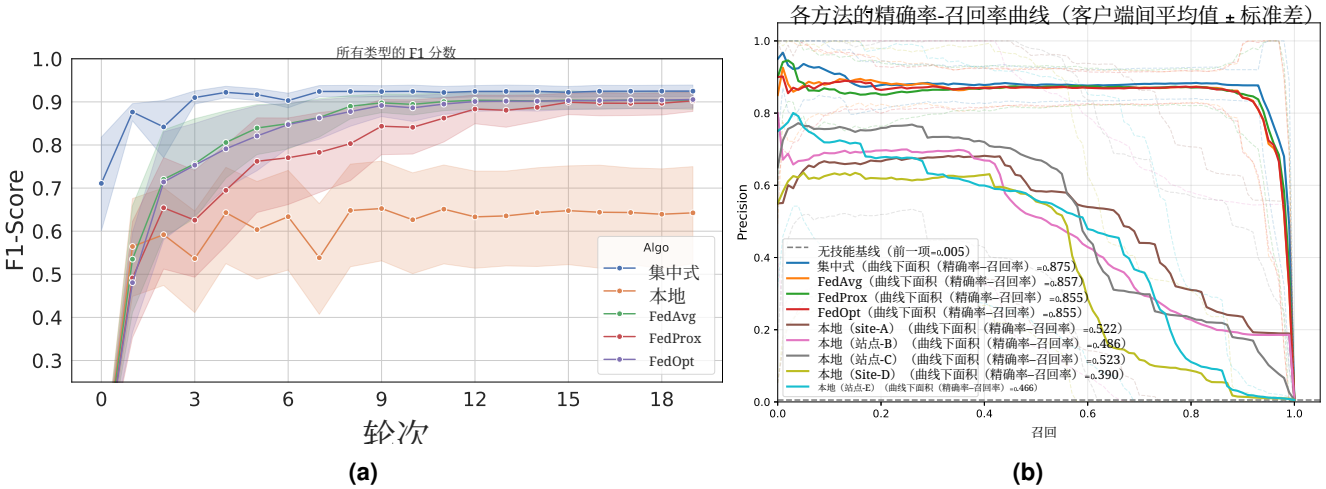
2.2 实验 1：联邦学习性能与收敛

使用 FedAvg 训练的联邦模型[10] 在所有客户端和欺诈类型上的表现始终优于本地训练模型。客户端的平均 F1 分数从 0.643（本地训练）提升至 0.903（FedAvg），增幅达 40%，如表 1所示。我们还比较了先进的联邦学习算法，例如 FedOpt[15]，（它使用服务器上的 SGD 优化器更新全局模型）和 FedProx[5]，（它在本地训练期间添加近端项，以防止模型偏离当前全局模型过远）。结果表明，在我们的场景中，无需使用这些先进算法也能获得更高的性能。

表 1. 所有客户端和欺诈类型经过 20 轮次后的平均二元 F1 分数。

算法	平均 F1 分数
集中式	0.925
FedAvg	0.903
FedOpt	0.905
FedProx	0.903
本地	0.643

在比较模型在训练轮次中的收敛情况时，FedAvg 始终优于更先进的联邦优化器，如图所示。3A。此外，精确率–召回率曲线（图3B）证实了其在类别不平衡情况下性能有所提升：FedAvg 的表现接近集中式性能，且显著优于本地训练模型。



(a) (b)

图 3。 集中式与本地与联邦学习：收敛性和精确率–召回率性能。A, 各客户端和异常类型在联邦轮次中的平均二元 F1 分数（阴影带：客户端间的标准差）。**B,** 在各客户端测试集上评估每个检查点模型的精确率–召回率曲线；实线表示平均值，阴影带表示客户端间的 ±1 标准差。图中包含无技能基线，并在图例中报告曲线下面积（精确率–召回率）值。

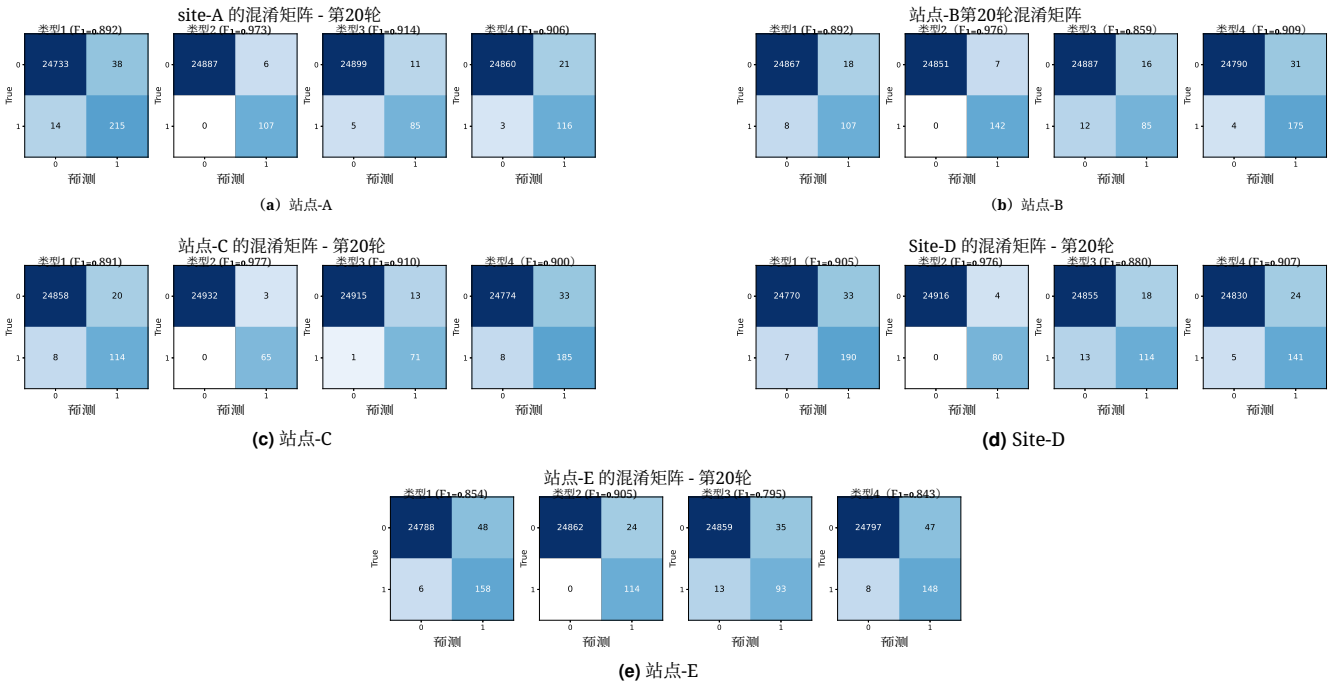
接下来，我们将考察五家机构各自使用 FedAvg 训练的最终全局模型所生成的混淆矩阵（图4）。在这种“按类型”评估中，我们针对给定异常类型与正常类型（采用一对其余方式）对应的子集计算二元 F1 分数；而在图 3中，我们报告各类型和客户端的平均值。第20轮的矩阵显示，在大多数欺诈类别中，真阳性率始终较高，而假阳性率则明显较低。这些结果表明，联邦模型有效地平衡了精确率和召回率，在尽量减少将合法交易误分类的同时，对欺诈活动保持了较高的敏感度。对于实际运行的欺诈检测系统而言，这种平衡至关重要，因为未能识别欺诈行为（假阴性）和错误拦截合法交易（假阳性）都会造成巨大的财务成本和客户体验成本。

总而言之，我们可以从定量结果中得出以下结论。

- **性能优势：** FedAvg 的表现优于任何仅使用本地数据进行训练的银行，同时与集中式训练的表现相当，验证了它能够在协作提升欺诈检测能力的同时保护数据隐私。

• **快速收敛：** 学习曲线显示，模型通常在 3-5 个联邦训练轮次内迅速趋于稳定，这表明从异构数据源中进行知识聚合的效率很高，并有望缩短训练时间。

• **跨域学习：** 客户端在其本地训练数据中未出现的欺诈类型上表现出显著的性能提升。例如，主要拥有类型 1 欺诈数据的客户端在联邦训练后，在类型2欺诈上的 F1 分数达到较高水平（>0.94）。



(a) 站点-A

(b) 站点-B

(c) 站点-C

(d) Site-D

(e) 站点-E

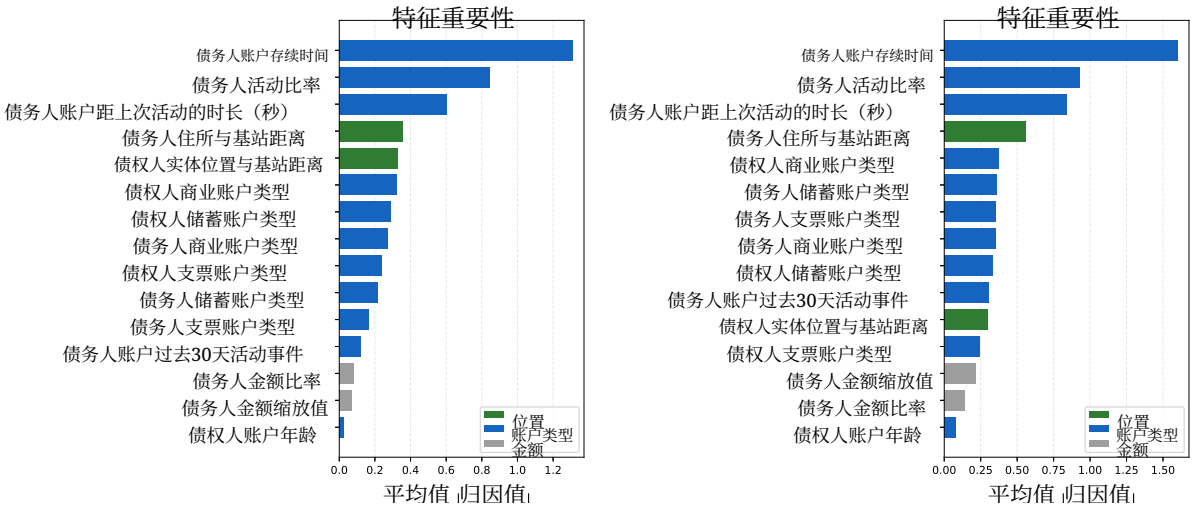
图 4. 各参与银行最终 FedAvg 全局模型的混淆矩阵 .

2.3 模型可解释性与特征重要性

使用 Captum 进行特征重要性分析[1] 证实，联邦模型学习到了符合领域预期的有意义模式，如图 5所示。

对于基于位置的欺诈（类型 1），地理特征在重要性得分中占主导地位。对于账户年龄和基于交易金额的欺诈（类型2），模型优先考虑了这些特征，这证实模型的决策具有可解释性，并以相关金融指标为依据。

[1] https://captum.ai



(a) 集中式训练　　**(b)** 联邦学习

图5. 基于 Shapley 值的归因（Captum GradientShap）的特征重要性图。

2.4 实验 2：差分隐私联邦学习

采用差分隐私随机梯度下降（DP-SGD）可提供强有力的形式化隐私保证，实现的有效隐私预算为 $\varepsilon = 10.0$，同时保持了有竞争力的模型性能，如图6所示。这些发现表明，在效用仅适度下降的情况下，也可以将稳健的隐私保护融入联邦学习流程。我们使用 Opacus[2] 库[16]，在本地训练期间于每个客户端上实施差分隐私随机梯度下降。隐私预算会随着联邦学习轮次的推进而累积，每个 epsilon 值代表整个联邦学习训练过程所消耗的预算。最终模型性能如表2所示。

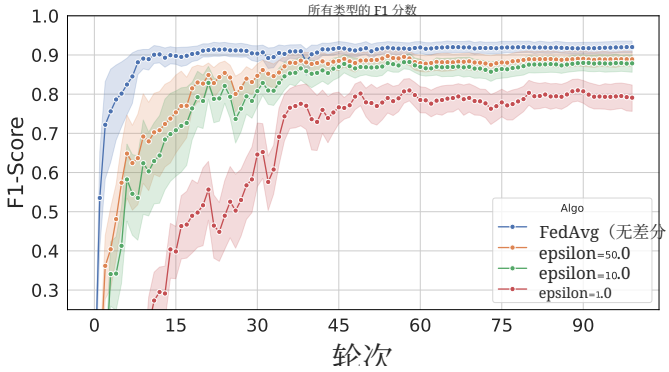


图 6. FedAvg 与 DP-FedAvg： 所有欺诈类型的 F1 分数在各客户端间的平均值。阴影区域表示各指标的标准差。

表 2. 使用差分隐私随机梯度下降时第 100 轮后的平均 F1 分数。不同隐私预算下所有客户端和欺诈类型的平均二元 F1 分数。

算法	隐私	平均 F1 分数
FedAvg	无差分隐私	0.920
FedAvg	$\varepsilon = 50.0$	0.889
FedAvg	$\varepsilon = 10.0$	0.879
FedAvg	$\varepsilon = 1.0$	0.791

[2] https://opacus.ai/

3 讨论

本研究使用 NVIDIA FLARE，评估了联邦学习（FL）在异构金融机构间开展协作式欺诈检测的效果，并探讨了四个问题：联邦学习相较于集中式训练的性能、收敛速度、模型可解释性，以及差分隐私下的隐私—效用权衡。

性能与跨机构协作。经过 20 轮后，FedAvg 的平均 F1 分数达到 0.903，接近集中式性能（0.925，ΔF1 = 0.022），同时显著优于仅本地训练（0.643，+40%）。这一提升的主要驱动因素是跨域泛化：每家机构在本地训练期间只观察到部分欺诈类型，这与现实情况一致，即各家机构接触到的欺诈类型模式有限。联邦机制直接解决了这一问题，客户端对其本地训练数据中未出现的欺诈类型也达到了 F1 $>$ 0.94，这一点得到了第 20 轮的分类别混淆矩阵（图 4）的证实。这一结果表明，联邦学习在金融欺诈检测中的主要运营价值，并非对已知欺诈类型进行渐进式改进，而是能够检测新出现或代表性不足的类型模式，避免单家机构将其遗漏。

方法选择。与梯度提升等基于树的方法相比，我们优先选择了深度神经网络，因为其基于梯度的连续参数空间可直接兼容 FedAvg 式权重平均——这是基于树的模型所不具备的特性。紧凑的三层架构（64、32、16 个神经元）在模型表达能力与每轮通信开销之间取得了平衡。为应对所有站点严重的类别不平衡（欺诈率为 0.50–0.60%，图 2a），我们选择 FocalLoss 而非标准交叉熵，使梯度更新聚焦于较难识别的欺诈少数类，而不是占主导地位的合法交易类别。

在联邦学习聚合算法中，FedAvg、FedProx 和 FedOpt 最终收敛至相近的 F1 分数（分别为 0.903、0.903 和 0.905，见表 1），其中 FedAvg 的收敛轨迹最为稳定。FedProx 的近端正则化和 FedOpt 的服务器端动量，在客户端漂移严重时最有益；但鉴于各站点的特征分布在结构上相似，这并非此处的主要情况。因此，对于异质性适中的联邦，FedAvg 是合适的默认选择；当欺诈类型存在显著差异，或各机构间的数据量高度不对称时，则应考虑更复杂的算法。

收敛性。模型在 3–5 个联邦轮次内趋于稳定，超过第 10 轮后提升幅度很小（图 3A）。这种快速收敛将通信开销限制在初始的短时间内，使联邦学习在对延迟和成本敏感的金融环境中具备实际可行性。通过 NVIDIA DGX Cloud 部署于 H200 GPU 实例上的 NVFlare Scatter-Gather 工作流，在该实验配置下支持了这一轮次节奏，且未观察到明显瓶颈。

可解释性与监管对齐。GradientShap 归因分析证实，联邦模型的决策信号与底层欺诈规则一致：位置特征在类型 1欺诈中占主导，而账户年龄和交易金额则在类型2中起主导作用（图5）。联邦归因概况与集中式归因概况高度一致，表明相较于使用汇总数据训练的模型，联邦聚合不会扭曲学习到的表征。这些结果表明，在本概念验证的范围内，该方法满足 SR 11-7 等模型风险管理框架的可解释性要求；此类框架要求模型决策以符合领域知识且可审计的特征信号为依据。

隐私—效用权衡。随着隐私预算收紧，差分隐私随机梯度下降导致 F1 值单调下降：在 100 轮内，从 0.920（无差分隐私）降至 0.879（$\varepsilon = 10.0$）和 0.791（$\varepsilon = 1.0$）（表2）。这种权衡在实践中是否可接受，取决于各机构的运营阈值——在交易量较高时，即使 F1 值下降 0.04，也会对检测到的欺诈和产生的误报造成可衡量的影响；部署前应根据机构具体的风险承受度进行评估。此外，还必须区分本研究采用的两种隐私机制。差分隐私随机梯度下降可针对对训练模型发起的推理攻击提供样本级保护，而联邦学习的架构保障则确保原始交易数据不会离开各机构。两者针对的是不同的威胁面，结合使用比单独采用任何一种机制能提供更全面的隐私保护——对于需要遵守通用数据保护条例等框架规定的合规要求及数据主权法规的机构而言，这一点尤为重要。

局限性。这些结果的适用范围受到三项限制。首先，所有实验均使用了通过基于规则的异常注入生成的合成数据，无法体现现实世界中欺诈的复杂性，包括对抗性适应、概念漂移或真实标签模糊性。在生产部署之前，仍有必要使用真实交易数据进行验证。其次，联邦由五个数据量相等且欺诈类型分区明确的站点组成；实际部署中的不对称性会更明显，类型模式也会有所重叠。第三，整个研究过程中均假设参与方行为诚实；研究未评估系统对恶意梯度更新或模型投毒的鲁棒性。后续工作应优先考虑使用真实交易数据进行影子模式试点验证、在更严苛的非独立同分布条件下开展压力测试，以及集成安全聚合协议，以应对诚实参与方这一假设。

结论。总体而言，本研究报告的结果有力地表明，采用差分隐私增强的联邦学习是一种可行且可扩展的协作式欺诈检测解决方案，能够在维护数据主权并提供更高隐私保障的同时，实现接近集中式性能的效果。

4 方法

4.1 实验目标

本实验按照以下设计目标进行：

1. **非独立同分布数据分布：** 模拟真实、异质的数据分布，反映各机构面临不同欺诈模式的情况。2. **跨域泛化：** 评估联邦学习模型检测参与方本地数据中未出现的欺诈类型的能力。3. **隐私保护协作：** 验证在不共享原始交易数据的情况下进行有效模型训练。4. **可扩展性评估：** 在 NVIDIA DGX Cloud 上使用多个参与方（5 个客户端站点）测试该框架。5. **算法比较：** 对 FedAvg、FedProx 和 FedOpt 进行基准测试，以确定最佳聚合策略。6. **模型可解释性：** 通过夏普利分析确保预测结果可解释。

4.2 数据

合成交易生成器。生成合成数据集，以模拟真实的支付交易场景，并确保隐私合规、受控实验和可复现性。数据生成方法旨在直接支持总体实验目标，并为评估联邦学习在欺诈检测中的表现提供受控且保护隐私的基础。[8]。

合成数据的依据。开发稳健的支付异常检测模型面临的根本障碍——尤其是在联邦环境中——是缺乏带标签的高保真交易数据。现实世界的支付数据集受到严格的监管与隐私限制，无法公开共享；即使可以获取，也存在严重的类别不平衡：欺诈交易通常不到交易总量的 1%。联邦学习进一步加剧了这一难题，因为要验证协作训练是否能利用多样化的本地数据分布生成更优的全局模型，就必须让不同参与方的数据分布具有异质性。为应对这些限制，我们开发了一个完全可配置的纯合成数据生成框架。该框架具有以下特性：

- 可控的统计异质性：每个模拟机构（站点）都从独立参数化的分布中抽取特征值，从而能够系统地改变联邦参与方之间的数据特征。
- 可组合的基于规则的异常注入：通过一组模块化扰动函数引入异常，这些函数既可单独应用，也可组合使用，以生成复杂程度各异的多维欺诈特征。
- 可调的类别平衡和标签噪声：异常记录数的比例和正确标注的概率均可配置，因此能够生成涵盖从理想化情形（异常比例高、标签完美）到现实情形（极度不平衡、含噪标签）的数据集。

跨机构非独立同分布分布选择与参数配置 为模拟不同金融机构之间的异质性，每个客户端站点均通过以下两项参数化：**(i)** 训练期间观测到的异常类型子集；**(ii)** 站点特定的特征分布。生成器配置定义了多个站点，并为其设置不同的异常类型暴露范围，使得没有任何单一站点能够观测到所有异常类型。这会形成非独立同分布的联邦学习场景，在该场景中，跨站点协作能够提升对代表性不足或未见过的异常类型的检测能力。

特征分布和可调参数。生成器采用简单且易于解释的分布，并且可针对每个站点独立配置：

- 交易金额仅使用对数正态分布，原因有三：（i）它能确保样本严格为正且有限——正态分布 $\mathcal{N}(\mu, \sigma^2)$ 会赋予$P(X < 0) = \Phi(-\mu/\sigma)$，导致金额以非微不足道的$\sigma/\mu$取负值；（ii）其右偏重尾形态符合实证支付金额分布特征，而均匀分布会产生不切实际的平坦密度；（iii）根据乘法中心极限定理，它自然产生于乘法过程（汇率、数量与价格的乘积）。伽马分布同样具有非负性，但相比直接指定算术平均值，其形状/规模参数化不够直观，而且尾部较轻，无法充分涵盖极端异常金额。

- 坐标扰动和类别选择使用均匀分布，以表示在有界区间内的最大无知——没有哪种塔楼偏移幅度比其他幅度更受偏好，而带符号扰动则无法采用对数正态分布。
- 对活动事件计数进行离散均匀抽样，会迫使分类器学习范围边界，而不是利用分布形状。

对数正态分布以期望的算术平均值 $\bar{x}$ 和标准差 σ_{raw}指定，并通过以下方式转换为底层参数：

$$\sigma^2 = \ln\left(1 + \frac{\sigma_{\text{raw}}^2}{\bar{x}^2}\right), \qquad \mu = \ln(\bar{x}) - \frac{\sigma^2}{2}$$

随后，通过改变这些参数（例如，站点-A 的正常个人金额： $\bar{x} = 20,000$；站点-E 的正常个人金额： $\bar{x} = 150,000$）以及塔坐标扰动界限（$[-0.75, 1.25]^\circ$ 与 $[-10, 10]^\circ$），实现跨站点异质性。

基础数据生成 每条记录包含若干属性

- 参与者属性（借方/贷方）：基本身份信息（姓名、出生日期、地址）以及来自真实土地位置的实际地理坐标均通过 Faker 选取。
- 账户创建时间戳在以下范围内均匀分布：$(t\text{dob}, t\text{now} - 12\text{w}]$, 最近活动时间戳在以下范围内 $[\max(t\text{create}, t\text{now} - 1\text{w}), t\text{now}]$
- 账户类型从{SAVINGS, CHECKING, BUSINESS}中均匀分布地选取
- 基站坐标由实际位置扰动得出：纬度基站 $=$ 纬度实际 $+\Delta$, $\Delta \sim$ 均匀分布(l, h)。
- 活动事件从按账户类型划分的范围中抽取（商业账户： $[50, 1.5\text{M})$， 支票账户： $[2, 500)$， 储蓄账户 : $[2, 50)$)
- 财务属性：借方金额根据账户类型，按照站点特定的对数正态分布生成；贷方金额通过A贷方 $= A$借方 $\cdot r$(C借方, C贷方)并使用欧洲央行汇率计算。
- 所有行均以欺诈标记 $= 0$ 初始化。

交易金额分布。 交易金额采用对数正态分布建模，以体现现实世界中支付金额的重尾特征。身份信息/地址使用 Faker 生成；时间字段遵循上述参数化采样规则；金额则通过站点特定的对数正态分布生成。

异常类型规则。 我们重点关注贯穿整个实验使用的四种结构化异常族。属性根据异常类型进行分组

- 基站/实际位置不匹配（类型 *1*）- 基站坐标偏移到超出正常数值范围的位置。对于坐标$c \in [c\text{min}, c\text{max}]$, 可用偏移量 $\delta = (c\text{max} - c) + p$ 其中p 是站点特定的（通常为负值）扰动因子，用于将新坐标推向范围极值。候选范围$R = (c\text{max} - \delta, c\text{max})$ 或 $(c\text{min} \ \delta - |c\text{min}|)$ 均按均匀分布采样。此方法应用于全部四个基站坐标（双方的纬度/经度）。
- 类型 *2* - 新建账户 + 异常金额- 此处将账户创建时间重置为付款前几分钟：$t\text{newcreate} = t\text{init} - (h \cdot 3600 + m \cdot 60 + s)$，其中$h \in \{0..5\}$、$m \in \{0, 1, 4, 9, 16, 25\}$、$s \in \{1, 10, 20, 30, 40, 50\}$。交易金额从异常对数正态分布中重新采样（例如，站点-E： $\bar{x}\text{anom} = 750{,}000$ 对比 $\bar{x}\text{normal} = 150{,}000$）。
- 类型 *3* - 休眠账户重新激活 - 将最近活动时间提前$d \in \{90, 120, 150, 180\}$ 天（另加亚日级抖动），形成 3–6 个月的休眠间隔。
- 类型 *4* - 异常交易频率 - 将活动次数提高至：商业账户 $[1\text{M}、5\text{M})$， 支票账户 $[525, 2000)$， 储蓄账户 $[75, 2000)$——与正常范围仅有少量重叠。

行选择与标签噪声 行选择支持按欺诈行百分比 α和欺诈类型重叠比例 β控制重叠。如果 $\mathscr{D}$是数据集，其中有 $\beta > 0$, $n_F = \lceil |\mathscr{D}|\alpha \rceil \beta \rceil$ 行从先前已标记的记录中抽取，$n_N = \lceil |\mathscr{D}|\alpha \rceil - n_F$ 行从干净记录中抽取，从而生成复合多类型异常。注入后会添加标签噪声：假设p^{apply} 是应应用欺诈标签的行所占比例，则从欺诈标记行中随机抽取的记录里，有一定比例（$1 - p^{\text{apply}}$）的标签会被翻转为 0，从而产生特征异常但标签为非欺诈的记录。

最终生成的包含n行的数据集极度不平衡，欺诈行总数为n_F，$n_{Fn} \in [0.001, 0.01)$ 。

表3. 各客户端站点使用的异常类型 **ed across client sites.** 注入合成支付数据集的结构化异常族 .

异常类型	描述	示例
1: 基于位置	预期位置与实际交易位置之间存在显著地理偏差 预期交易位置与实际交易位置	预期位置：距离纽约市 500 英里；起源地： 阿塞拜疆
2: 基于账户年龄	来自比例失衡账户的高价值交易 新账户	账户年龄：5 小时；金额：15,000 美元
3: 基于不活跃	长期不活跃后出现的异常交易 账户不活跃	最近活动：6 个月前；突然出现高/低- 交易金额
4: 账户类型异常活动	短时间内交易频率过高 时间窗口	10 分钟内有 50 笔支票账户 或储蓄账户交易

表4。 重点特征及其关联的异常类型。 此处仅显示 60+ 特征的一个子集（与异常规则最相关的特征）。

特征	Type	描述	数据类型
欺诈标记_	目标	表示该交易是否异常（历史上标记为欺诈）。 已标记为欺诈。	int
借方地理纬度_ _	类型 1	借方预期位置的纬度。	浮点数
债务人地理经度_ _	类型 1	债务人预期位置的经度。	浮点数
债务人基站纬度_ _	类型 1	交易中使用的蜂窝基站纬度。	浮点数
债务人基站经度_ _	类型 1	交易中使用的蜂窝基站经度。	浮点数
付款发起时间戳_ _	类型2	付款发起的时间戳。	日期时间
债务人金额_	类型2	借方账户的扣款金额。	浮点数
借方账户创建时间戳_ _ _	类型2	借方的账户创建时间戳。	日期时间
债务人账户最后活动时间戳_ _ _ _	类型 3	借方账户的最近活动时间戳。	日期时间
借方账户类型_ _	类型 4	借方账户的类型（例如支票账户、储蓄账户）。	str
借方账户过去 30 天的活动事件数量_ _ _ _ _	类型 4	过去 30 天内的账户活动事件数量。整数	

客户端预处理。 根据 ≈ 50 个原始属性构建特征向量：

- 时间特征：这些特征由借方和贷方的账户年龄（分钟）（$t_{\text{init}} - t_{\text{create}}$ ~~60）~~ 以及距上次活动的间隔（秒）计算得出，随后进行对数变换（$x' = \ln x$），以压缩正常（$\sim 10^5$分钟）与异常（~ 1 分钟）账户年龄之间跨越多个数量级的范围。

- 地理空间特征：我们计算物理坐标与基站坐标之间的哈弗辛距离

$$a = \sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat_1)\cos(lat_2)\sin^2\left(\frac{\Delta lon}{2}\right), \quad d = 2R \cdot \text{atan2}(\sqrt{a}, \sqrt{1-a})$$

  与 $R = 6334.08$ 公里。

- 类别编码：账户类型采用独热编码，转换为三个二元指示变量（商业账户、支票账户、储蓄账户）。
- 缩放：数值特征通过标准化缩放器进行缩放
- 目标：欺诈_标记 $\in \{0, 1\}$

4.3 模型与训练

深度学习与传统机器学习： 出于多方面的方法论和实践考量，我们选择了深度神经网络（DNN），而非传统机器学习模型（如逻辑回归、随机森林、XGBoost）。首先，深度神经网络具有更强的表征能力，能够从高维交易数据中自动提取复杂的非线性特征交互，无需进行大量手动特征工程。其次，它们的

基于梯度的优化和连续参数空间与 FedAvg 等联邦聚合策略天然契合，因此深度神经网络尤其适用于联邦学习范式。

此外，深度神经网络学习到的层次化特征表示支持在异构的非独立同分布客户端数据集之间进行更有效的知识迁移——这在多机构金融环境中十分常见。从基础设施的角度来看，深度神经网络能够在包括 NVIDIA DGX Cloud 提供的分布式 GPU 资源在内的现代加速器上高效扩展。最后，深度学习方法在欺诈检测中始终能够达到业界领先的性能，进一步证明了其适合部署于生产级金融系统。

网络架构选择： 所选架构由三个全连接隐藏层组成，分别包含 64、32 和 16 个神经元，体现了在模型表达能力与运行约束之间经过审慎权衡的结果。该配置能够提供足够的容量来建模多样化的欺诈模式，同时保持紧凑的参数化设计，从而减轻过拟合问题——这在非独立同分布数据分布下尤为重要。较小的模型规模还能够减少联邦参数交换期间的通信开销，进而提升往返效率。此外，此规模的网络表现出良好的收敛行为，可减少实现稳定性能所需的联邦训练轮次。

模型摘要：

- **模型类型：** 使用 PyTorch 实现的深度神经网络（DNN）。
- **架构：** 由以下部分组成的全连接前馈网络：

  – **输入层** （大小：输入特征的数量）– **隐藏层 1：** 64 个神经元（层归一化 + ReLU 激活）– **隐藏层 2：** 32 个神经元（层归一化 + ReLU 激活）– **隐藏层 3：** 16 个神经元（层归一化 + ReLU 激活）– **输出层：** 2 个神经元

- **超参数：** 学习率 = 5e-4，联邦学习轮次 = 20，每轮训练周期数 = 1，批量大小 = 64，损失函数 =FocalLoss。
- **工作流：** 实验采用 NVFlare Scatter-Gather 工作流：服务器从头初始化全局模型并聚合模型更新（例如通过 FedAvg），客户端则在各自的私有数据集上进行本地训练。

4.4 样本级差分隐私

我们在本地训练期间使用 Opacus 的 PrivacyEngine 模块，实施样本级 (ε, δ) 差分隐私。差分隐私要求：对于任意两个数据集 $\mathscr{D}$ 和 $\mathscr{D}'$，若二者仅相差一个样本，则对于任意可测输出集合 $\mathscr{O}$，

$$\Pr[\mathscr{A}(\mathscr{D}) \in \mathscr{O}] \leq e^{\varepsilon} \Pr[\mathscr{A}(\mathscr{D}') \in \mathscr{O}] + \delta,$$

其中 $\mathscr{A}$ 表示随机化学习算法。

为满足这一保证，我们采用逐样本梯度裁剪实现高斯机制。对于参数为 θ 的模型，我们对每个训练步骤进行如下修改：

1. 逐样本梯度。令 g_i 表示小批量 B 中样本 i 所贡献的梯度。将梯度裁剪至固定的 ℓ_2 范数上限 C：

$$\tilde{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_2}\right).$$

2. 添加校准高斯噪声。随后对小批量的聚合梯度进行扰动：

$$\hat{g} = \frac{1}{|B|} \sum_{i \in B} \tilde{g}_i + \mathscr{N}(0, \sigma^2 C^2 I),$$

其中 σ 为噪声乘数。

3. 隐私核算。 Opacus 使用 Rényi 差分隐私（RDP）组合来跟踪所有优化步骤中的累积隐私损失。给定目标（ε， δ）和总训练轮数 E，隐私引擎会自动选择噪声乘数 σ，使最终的 RDP 保证转换为所需的（ε， δ）界限。

集成到训练中。实际上，训练循环通过以下方式实现差分隐私：

```
from opacus import PrivacyEngine

... 模型，优化器，训练_加载器 = 隐私_引擎.make_私有_通过_epsilon( 模块=模型，优化器=优化器，数据_加载器=训练_加载器，周期数=E，目标_epsilon=1.0，目标_delta=1e-5，最大_梯度_范数=1.0， )
```

这会将模型、优化器和数据加载器封装起来，并应用上述的梯度裁剪、噪声注入和隐私核算流程。该过程可确保每家机构的贡献在单个训练样本层面保持差分隐私，在保留整体学习动态的同时提供强有力的形式化隐私保证。

4.5 评估

我们使用 F1 分数作为主要指标评估模型性能，因为在欺诈检测场景中，假阳性和假阴性都会造成重大损失，而该指标能够平衡精确率和召回率。对于按类型进行的分析，我们将每种异常类型视为一个独立的二元检测任务（类型 k 与非欺诈），并以欺诈为正类计算二元 F1 分数（$y=1$）。通过对模型输出取最大值索引获得预测标签，并报告按异常类型和客户端站点汇总的平均 F1 值。为进行更全面的评估，我们还将准确率、精确率和召回率作为次要指标进行计算。我们通过收敛行为监控模型训练进度；收敛行为定义为性能指标在连续多个通信轮次中稳定在预设方差阈值内的时点。收敛后，我们使用基于 Shapley 值的归因（Captum GradientShap）评估模型可解释性，以验证所学特征归因是否符合已知的异常金融活动模式（使用 500 个样本进行基线归因）。

评估摘要：

- **主要指标：**欺诈检测的二元 F1 分数，其中欺诈为正类（$y=1$）。通过对模型输出应用最大值索引来获取预测类别标签。次要指标包括准确率、精确率和召回率。
- **收敛标准：**各项性能指标在连续轮次中趋于稳定（$\sigma \leq$ 阈值）。
- **可解释性：**模型收敛后，使用基于 Shapley 值的归因分析模型决策，以确认特征重要性与预期异常规则的一致性。

受控比较：为确保本地、集中式和联邦（FedAvg）训练之间的公平比较，我们在所有设置中采用相同的实验方案。所有模型均使用相同的网络架构和优化器/超参数（例如学习率、批量大小和本地训练轮数）。每个客户端均采用相同的特征预处理流程（包括归一化/缩放），并使用基于其缩放数据集拟合的参数（该数据集仅由训练数据生成，不存在测试集泄漏）；每次运行也使用相同的一组输入特征。评估在留出测试集上进行，且本地、集中式和联邦模型均采用相同的指标实现流程，因此观察到的性能差异可归因于训练范式，而非数据处理、模型容量或评估方法的变化。

4.6 基于 Shapley 值的特征归因

Shapley 值提供了一种有原则的方法，可将模型预测归因于各个输入特征。对于具有特征集 fF 的模型，特征i 的 Shapley 值 ϕ_i 定义如下：

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{i\}}(x) - f_S(x) \right].$$

此表达式计算特征 i 对每个可能联盟 S 的边际贡献，并进行加权，使所有特征排序得到对称处理。

对于给定输入x，我们将模型输出表示为

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i,$$

其中，ϕ_0 是期望模型预测，M 是特征的数量。这种可加性分解满足效率、对称性、虚设性和可加性——这些公理唯一地刻画了 Shapley 值。

精确计算在 $|F|$ 中呈指数级增长，因此实际应用依赖近似方法。在本研究中，我们使用`GradientShap`（来自 Captum）[3][17],，通过沿基线 $\mathbf{x}_{\text{base}}$ 与输入 $\mathbf{x}$ 之间的路径积分输入梯度，为可微模型估算归因值：

$$\phi_i \approx \frac{1}{K} \sum_{k=1}^{K} (x_i - x_{i,\text{base}}) \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}^{(k)}},$$

其中，$\mathbf{x}^{(k)}$ 为插值样本。由此得到忠实且局部准确的特征归因，适用于受监管的金融应用。

4.7 联邦学习软件

接下来，我们将说明如何在参与的金融机构中配置、分发和运行 NVFlare，包括安全配置、作业编排、依赖管理和实时监控。这些组件共同构成了一个可复现、可用于生产的软件环境，支持大规模可靠实验。

软件分发与系统运行： 联邦学习框架通过 NVFlare 仪表板进行部署和管理，该仪表板提供集中式界面，用于在所有参与站点间分发软件、进行配置和编排。这种方法可确保多机构联邦实验的一致性、安全性和可复现性。

启动套件生成： 管理员使用 NVFlare 仪表板为服务器和每个客户端生成经过加密签名的启动套件。这些套件包含安全接入所需的配置文件、身份验证凭据和证书。

客户端配置： 启动套件被安全地传输至每个参与机构，各机构的客户端在那里安装 NVFlare 软件包并初始化其环境。每个套件中提供的配置指定了网络端点、资源分配和机构特定的安全策略。

作业提交： 由模型架构、超参数、训练脚本和数据接口组成的实验工作流被打包为 NVFlare 作业。作业通过仪表板提交，仪表板会自动将其分发给所有客户端。集成的版本控制和回滚机制确保更新能够安全且可复现地部署。

软件包与依赖管理： Python 依赖项（例如 PyTorch、Captum、Pandas）通过专用 Docker 容器进行管理。自定义训练代码和模型定义被封装为 Python 模块，NVFlare 会自动在所有站点之间同步这些组件。分布式软件包集合包括客户端与服务器可执行文件、模型定义（例如 model.py）、训练逻辑（例如 client.py、central_train.py）、数据预处理工具，以及评估或可视化脚本。

这种基于仪表板的分发机制统一了各机构间的执行方式，并降低了多方联邦学习部署固有的运维复杂性。

实时监控框架： 为实现对联邦训练过程的全面可观测性，我们集成了 MLFlow，将其用作实时实验跟踪与可视化系统。MLFlow 界面可实时呈现全局及各站点性能，便于比较不同的联邦优化算法（例如 FedAvg、FedProx 和 FedOpt）。仪表板有助于及早发现收敛问题、客户端掉线或数据异常，并支持系统化的超参数优化。该监控框架让联邦系统的运行情况更加透明，为科学严谨性和运维可靠性提供支持。

服务器部署： 我们在 NVIDIA DGX Cloud 上配置了 MLFlow 跟踪服务器，并提供持久化存储和安全的基于网页的访问，供授权协作者使用。所有指标、超参数和制品均集中记录，以确保可复现性和可审计性。

[3] https://captum.ai/

表 5. NVIDIA DGX Cloud 上的系统设置和基础设施。

组件	规格
部署模式	DGX Cloud 上的多站点联邦拓扑
服务器	一个 NVFlare 服务器实例（仅使用中央处理器）
客户端	五个客户端站点，每个站点配备一个专用的 NVIDIA H200 图形处理器实例（由 DGX Cloud 管理）
联邦学习框架	NVIDIA FLARE (NVFlare) 软件开发工具包
运行时间	基于 NVIDIA NGC PyTorch 容器构建的 Docker 镜像

客户端集成： 每个 NVFlare 客户端都会在联邦轮次期间将训练指标记录到 MLFlow 服务器。自动记录的输出包括训练损失和验证损失、类别特定性能指标（精确率、召回率、F1 分数）、混淆矩阵、基于 Shapley 值的归因（GradientShap）可解释性图、超参数、模型检查点，以及运行时间和通信开销的测量数据。

系统架构： 该实验系统采用 NVFlare 实现客户端–服务器联邦架构。每个客户端（代表一家金融机构）都使用保留在本地部署环境中的专有数据集训练本地模型。集中式 NVFlare 服务器汇总本地模型更新，形成全局模型，从而在不共享原始数据的情况下实现协作学习。

4.8 联邦系统基础设施

实验在 NVIDIA DGX Cloud 基础设施上采用了多站点联邦拓扑，如表 [4] 所示。5NVIDIA DGX Cloud 基础设施具有多项优势，对本研究的实施至关重要。首先，其全托管 GPU 集群可显著减轻运维负担，使研究人员能够专注于方法开发，而非系统管理。其次，该平台采用企业级安全和合规机制，适用于涉及敏感金融应用的研究。第三，DGX Cloud 支持按需动态缩放计算资源，便于开展规模和复杂程度各异的实验。第四，该环境基于最新一代 NVIDIA GPU 技术构建，可提供高效深度神经网络训练所需的高性能加速。最后，该平台可轻松集成 NVIDIA FLARE，并提供经过优化的网络配置，以简化联邦学习部署。

为了让所有参与方都能在 NVIDIA DGX Cloud 上运行，我们构建了一个 Docker 镜像，其中包含所有软件依赖项，例如 Python 包、驱动程序、库和配置。我们使用 NVIDIA NGC Catalog 中的 PyTorch 容器镜像作为基础镜像。在表5中总结了系统设置。

代码和数据可用性

用于数据生成以及使用 NVIDIA FLARE 进行联邦学习的代码可用[5]。

致谢

我们衷心感谢 NVIDIA 通过 DGX Cloud 提供的支持[4]，，该支持促成了本项研究。

参考文献

1. LexisNexis Risk Solutions. 2025 年欺诈真实成本研究：电子商务与零售报告——美国和加拿大版。*LexisNexisRiskSolutions* (2025)。第 15 版。**2.** Javelin Strategy & Research. 2025 年身份欺诈研究：突破创新壁垒。*JavelinStrateg. &Res.*(2025)。涵盖 2024 年数据的年度身份欺诈研究。**3.** Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. & Bontempi, G. 信用卡欺诈检测：现实建模与新颖的学习策略。 *IEEE* 神经网络与学习系统汇刊 **29**, 3784–3797 (2017)。**4.** 欧盟。2016 年 4 月 27 日欧洲议会和理事会法规 (EU) 2016/679（通用数据保护条例）。https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016)。《欧盟官方公报》。**5.** Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. 联邦学习：挑战、方法与未来方向。 *IEEE* 信号处理杂志**37**, 50–60 (2020)。

[4]https://www.nvidia.com/en-us/data-center/dgx-cloud[5]
https://nvidia.github.io/NVFlare/research/fsi-fraud-detection

6. Kairouz, P. 等。联邦学习的进展与未解问题。《机器学习基础与趋势》14, 1–210 (2021)。 7. El Hallal, T. & El Mourabit, Y. 用于信用卡欺诈检测的联邦学习：关键基础与新兴趋势。载于《2024年国际电路、系统与通信会议》(ICCSC)，1–6 (IEEE，2024)。 8. Upadhyay, S. 等。Project aikya：通过去中心化人工智能增强金融交易中的异常检测。技术报告，Kinexys by J. P. Morgan，纽约州纽约市，美国 (2025)。 9. Roth, H. R. 等。NVIDIA FLARE：从模拟到现实世界的联邦学习。《IEEE数据工程简报》，第 46 卷，第 1 期 DOI: https://doi.org/10.48550/arXiv.2210.13291 (2023)。 10. McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. 基于去中心化数据的深度网络通信高效学习。载于《人工智能与统计学》，1273–1282 (PMLR, 2017)。 11. Shapley, L. S. 等。《n人博弈的一个价值》（普林斯顿：普林斯顿大学出版社，1953年）。 12. Lundberg, S. M. & Lee, S.-I. 解释模型预测的统一方法。《神经信息处理系统进展》30 (2017)。 13. Abadi, M. 等。基于差分隐私的深度学习。载于《2016年ACM SIGSAC计算机与通信安全会议论文集》，308–318 (2016)。 14. Balle, B., Barthe, G. & Gaboardi, M. 通过子采样实现的隐私放大：基于耦合与散度的紧致分析。《神经信息处理系统进展》31 (2018)。 15. Reddi, S. 等。自适应联邦优化。arXiv 预印本 arXiv:2003.00295 (2020)。 16. Yousefpour, A. 等。Opacus：PyTorch 易用型差分隐私库。arXiv 预印本 arXiv:2109.12298 (2021)。 17. Kokhlikyan, N. 等。Captum：适用于 PyTorch 的统一通用模型可解释性库 (2020)。2009.07896。