# FedNoRo：通过解决类别不平衡与标签噪声异质性实现抗噪声联邦学习

**吴楠楠**[1] , **于力**[1] , **姜雪峰**[2] , **郑光廷**[3] 和 **闫增强**[1*1]华中科技大学电子信息与通信学院[2]中国科学院计算技术研究所[3]香港科技大学工学院{wnn2000, hustlyu}@hust.edu.cn, jiangxuefeng21b@ict.ac.cn, timcheng@ust.hk, z yan@hust.edu.cn

arXiv:2305.05230v2 [cs.LG] 1 Aug 2023

**摘要**

联邦噪声标签学习（FNLL）正成为一种有前景的工具，可用于隐私保护的多源去中心化学习。现有研究依赖于全局数据类别均衡这一假设，可能无法对复杂的标签噪声进行建模，尤其是在医疗场景中。本文首先提出一种新的、更符合实际的联邦标签噪声问题：全局数据存在类别不平衡，且标签噪声具有异质性；随后提出一种名为 FedNoRo[1] 的两阶段框架，用于实现抗噪声联邦学习。具体而言，在 FedNoRo 的第一阶段，采用逐类别损失指标，并结合高斯混合模型识别含噪客户端。在第二阶段，联合采用知识蒸馏和距离感知聚合函数，以实现抗噪声的联邦模型更新。在广泛使用的 ICH 和 ISIC2019 数据集上的实验结果表明，在真实世界的联邦学习场景中，FedNoRo 在解决类别不平衡与标签噪声异质性方面优于最先进的联邦噪声标签学习方法。

## 1 引言

联邦学习（FL）允许各个客户端在不共享数据的情况下协同训练深度学习模型，因此在注重隐私的场景中得到了广泛研究，例如医疗 [Kaissis 等、 2020] 金融 [Zheng 等，2021] 等应用。大多数现有的联邦学习框架都基于全监督学习范式[ McMahan等，2017；Li 等，2020b], ，其隐含假设是每个参与方的数据都被正确标注。然而，在现实的医疗场景中，构建一个带有高质量标注的完全干净数据集成本高昂，因为医疗数据标注既耗时又耗费人力，并且需要专业知识。因此，雇用非专业人员进行标注或使用自动标注技术 [Irvin 等人， 2019]时不可避免地会引入噪声标签。因此，在现实世界的联邦学习场景中，更常见的情况是，部分客户端的标签是干净的，而其他客户端的标签则不是。

由于存在噪声客户端，开发具备抗噪能力的联邦学习框架至关重要，而准确识别噪声客户端是首要且最关键的一步。现有的噪声客户端检测方法提出，计算每个客户端所有样本的平均指标（例如 损失），并将其作为该客户端的特征，再筛除特征异常的客户端，将其视为噪声客户端。这类方法假设干净客户端的特征独立同分布（IID），而噪声客户端则是遵循小损失技巧的离群值 [Han 等， 2018]。具体而言，Xu 等 [2022] 计算了每个客户端的平均 LID 值 [Houle， 2013] 并将平均 LID 值较大的客户端识别为噪声客户端。类似地，Wang 等 [2022] 用置信度分数替代 LID 值，并将置信度分数较低的客户端识别为噪声客户端。遗憾的是，由于以下观察，这些指标在应对现实联邦学习场景中的噪声客户端时可能效果不佳：

1. 从全局视角来看，数据存在严重的类别不平衡 [Wang 等, 2021; Shang 等， 2022]。在类别不平衡的情况下，对所有类别使用全局指标会对客户端的标签分布高度敏感，而这些分布可能存在巨大差异。

2. 各客户端之间的数据具有异构性 [Li 等, 2020b]。由于在联邦学习中，各客户端独立收集自己的数据，因此数据可能存在严重差异，从而影响客户端间指标的计算。

3. 各客户端之间的标签噪声具有异构性，其异构程度体现在强度和模式两方面。前者表示各客户端的噪声率不同，后者则指与客户端本地数据分布相关的各种标签噪声形式。

受类别不平衡影响，联邦学习中的联邦模型可能会偏向全局多数类别 [Kang 等，2020；Zhou 等，2020；Cao 等，2019；Menon等， 2021], 从而导致不同类别的指标差异较大，使含噪客户端的指标更难区分。例如，在临床场景中，假设有一家肿瘤专科医院 $A$和一家综合医院 $B$，$A$ 会收治更多恶性病例，而$B$ 会收治更多健康病例。因此，$A$ 更

*通讯作者[1]代码可在 https://github.com/wnn2000/FedNoRo 获取。

有可能因类别不平衡（即，健康 ≫ 恶性）而产生异常的客户端特征（例如，出现与含噪客户端类似的较大损失值）。除了数据异质性，客户端特征还会受到异质性标签噪声的影响，使含噪客户端检测更加困难。例如，在临床场景中，假设有两家患者群体构成不同的医院$C$和$D$（即，标签分布各异），一个良性病例更有可能被$C$误诊为健康病例，却被$D$误诊为恶性病例。尽管这两个标签都是错误的，但由于类别不平衡（即，健康 ≫ 恶性），$C$产生的损失值（即，用于含噪客户端检测的指标）会远小于$D$。总之，类别不平衡会使客户端特征不再仅与客户端是否含噪有关，还与其局部数据分布和标签噪声模式有关，导致上述基于全局指标的识别方法难以适用于现实的联邦学习场景。

为解决这一问题，一种直接的方法是将类别均衡 [Kang 等, 2020; Zhou 等, 2020; Cao等, 2019; Menon 等，2021] 与噪声标签学习相结合。然而，这些方法只能缓解类别先验概率上的偏差，而类别特定特征的学习难度偏差 [Yi 等，2022] 仍未解决。由于训练数据相对有限，与多数类的类别特定特征相比，少数类的类别特定特征更难学习，从而导致特征空间中的不平衡。采用特征学习 [Karthik 等, 2021; Yi 等，2022] 来缓解学习难度上的偏差时，又会受到各客户端有限的本地数据和计算资源的制约。迄今为止，如何解决联邦噪声标签学习（FNLL）中的类别不平衡问题，仍未得到充分研究。

本文首先构建一个新的 FNLL 问题，以刻画类别不平衡条件下更贴近现实的联邦学习场景，随后提出用于鲁棒学习的两阶段框架 FedNoRo。具体而言，在第一阶段，我们不使用全局指标，而是根据由 FedAvg [McMahan等人，2017]训练的预热模型计算出的客户端级每类平均损失值，识别含噪客户端。由于每个类别都是独立考量的，这种检测方法受类别不平衡和异质性的影响较小，因此检测效果更好。在第二阶段，针对干净客户端和含噪客户端分别采用不同的学习策略：干净客户端使用交叉熵损失进行训练，含噪客户端则采用知识蒸馏（KD） [Hinton 等人，2015]，以减轻噪声标签带来的负面影响。除了客户端级训练外，我们还提出一种距离感知聚合函数，以便在服务器进行全局模型更新时更好地平衡干净客户端和含噪客户端的重要性。在两个阶段的本地训练阶段中，均采用 logit 调整（LA） [Menon 等人，2021]，以应对数据异质性和类别不平衡。主要贡献总结如下：

- 一种全新的 FNLL 问题，同时考虑类别不平衡的全局数据和异质标签噪声，以模拟真实的联邦学习场景。
- 一种针对多源数据的新型标签噪声生成方法

其中，合成标签噪声具有异构性且依赖实例。

- 一种名为 FedNoRo 的两阶段联邦学习框架，旨在同时应对类别不平衡和标签噪声异质性。在 FedNoRo 中，基于异常的逐类别损失值识别含噪客户端，并采用抗噪训练策略来有效更新联邦模型。
- 在真实世界的多源医学数据集上优于最先进的FNLL方法。

## 2 相关工作

### 2.1 联邦学习

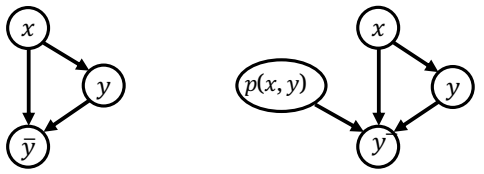
联邦学习(FL)因其在保护隐私的应用中具有潜力而受到广泛关注 [McMahan 等， 2017]。现有研究主要关注联邦学习应用面临的挑战，包括数据异质性 [Li 等，2020年b；Karimireddy 等，2020年；Li等，2021年a；Li 等，2021年b；Yan 等，2020年；Zhu 等，2021年；Li 和 Zhan， 2021], 数据质量 [Liu等，2021年；Liang 等，2022年；Jiang等，2022年a], 隐私保护 [Agarwal 等， 2018] 以及通信效率[Koneˇcn`y 等，2016年；Sattler等， 2019]。

### 2.2 噪声标签学习

由于在现实世界应用中难以收集大量高质量的干净标签，噪声标签学习至关重要。现有研究大致可分为两类：基于筛选 [Han 等，2018; Yu 等，2019; Li 等，2020a; Wei 等，2022b;Zhao 等，2022; Yi 等， 2022] 以及无筛选 [Wang 等，2019; Zhang 等，2017; Wei 等，2022a; Lukasik等，2020; Lukov 等， 2022]。前者假设干净样本和噪声样本在某个指标（例如 损失值）上表现不同，并可据此进行识别。后者则侧重于设计能够抵抗噪声标签的特殊损失函数或正则化。

### 2.3 联邦噪声标签学习

联邦噪声标签学习（FNLL）是一个新兴的研究课题。Chen 等人 [2020] 和 Yang 等人 [2021] 利用干净数据集辅助 FNLL，但这在实践中可能难以实现。RoFL [Yang 等人， 2022] 通过共享逐类质心更新本地模型，以避免噪声标签导致的决策边界不一致。FedLSR [Jiang 等人，2022年b] 在本地训练阶段引入正则化，以减轻噪声标签的影响。Fang 等人 [2022] 将 FNLL 与异质模型问题相结合，并提出了一种耐噪损失函数和一种客户端置信度重加权方案。该方法平等对待所有客户端，没有明确筛除噪声客户端。相比之下，FedCorr [Xu等人， 2022] 和 FedNoiL [Wang 等人， 2022] 是基于选择的方法：首先通过设计的指标筛选噪声客户端，然后采用适当的训练策略更新联邦模型。遗憾的是，假设全局数据类别均衡，使它们在处理类别不平衡的现实联邦学习场景时效果较差。




(a) 实例依 (b) 异质实例

图 1：标准实例依赖噪声和所提出的异质实例依赖噪声的图模型。

## 3 问题定义

### 3.1 预备知识

给定一项多类图像分类任务，其中 $x_i \in \mathcal{X}$是输入图像，$y_i \in \mathcal{Y} = [C]$ 是对应的真实标签。目标是使用大小为 $N$ 的训练数据集$D_{tr} = \{(x_i, y_i)\}_{i=1}^{N}$ 训练模型 $f(\cdot)$，用于分类；该数据集假设存在类别不平衡（即 $p_{tr}(y) \neq \frac{1}{C}$）。在联邦学习中，由于隐私问题，无法直接构建 $D_{tr}$，因此需将其划分为 $K$ 个互不重叠的子集，并分发给 $K$ 个客户端。对于客户端 $k \in [K]$中的任意私有数据集 $D_k = \{(x_i, y_i)\}_{i=1}^{N_k}$，其大小为 $N_k$，其中每个图像-标签对 $(x_i, y_i)$ 均遵循私有联合概率分布$p_k(x, y) = p_k(y)p_k(x \mid y)$。根据 [Shen 等人，2022；Xu 等人，2022]，我们假设条件分布$p_k(x \mid y)$相同，而局部类别先验 $p_k(y)$在不同客户端间有所变化（即对于 $k_1 \neq k_2$，$p_{k_1}(y) \neq p_{k_2}(y)$），这是由数据异质性导致的。为模拟噪声标签，某些客户端中的真实标签$y$ 会被替换为噪声标签$\bar{y}$。

### 3.2 噪声模型

为真实地模拟多源数据中的标签噪声，我们构建了异质标签噪声模型。在标签噪声强度方面，我们首先将全局噪声率$\rho$ 定义为含噪客户端所占的比例。然后，我们假设任意含噪客户端 $i$ 的局部噪声率 $\eta_i$（即，含噪样本所占的比例）服从均匀分布 $U(\eta^l, \eta^u)$，如[Xu 等人所示， 2022]。在标签噪声模式方面，为了更好地建模，我们提出了异质实例依赖噪声（H-IDN）。如图 1 和定义 1 所示，与标准实例依赖噪声（IDN）相比，局部数据分布是调节标签噪声的一个额外因素 [Chen等人，2021]。

**定义 1**（H-IDN）。如果噪声转移概率是局部数据分布的函数，则 *IDN* 具有异质性，即$\Pr(\bar{Y} = y \mid Y = y, X = x) = M_{y,\, y,\, x}(p_i(x, y))$，其中$M_{y,\, y,\, x}$ 表示实例 $x$的噪声转移矩阵。

通过这种方式，所提出的异质实例依赖噪声能够对更复杂且更真实的标签噪声进行建模。

### 3.3 噪声生成

为了模拟第 3.2 节中的噪声模型，我们提出了一种噪声生成器，用于按照算法 1 中的描述，从多源干净数据生成异质标签噪声。继[Chen 等人之后，使用 2021], 神经网络作为标注器，以模拟噪声客户端中的噪声标注过程。具体来说，我们首先从 $K$ 客户端中随机选择 $\rho K$ 个作为噪声客户端，并基于每个选定客户端的原始数据（即干净数据）独立训练一个神经网络。然后，使用训练好的神经网络生成属于该客户端的所有样本的分类概率。给定第 $i$ 个含噪客户端中的每个实例 $x_j$ 及其对应的分类概率 $p(Y \mid x_j) \in [0,1]^C$，其误分类概率由下式确定：

$$\widetilde{p}(x_j) = 1 - p(Y = y_j \mid x_j), \tag{1}$$

根据归一化后的误分类概率 $\widetilde{p}\,x \in (), [0,1]^{N_i}$，总共将选取 $\eta_i N_i$ 个样本作为含噪样本。根据公式，困难样本更有可能被选为含噪样本。对于每个选定的含噪样本$x_j$，从概率分布 $p(Y \mid x_j)$ 中为其分配一个不同于原始干净标签 $y$ 的含噪标签 $\bar{y}$。通过这种方式，含噪样本更有可能被赋予预测置信度较高的错误标签。由于不同客户端的噪声由不同的神经网络生成，因此符合异质噪声模式的定义。遵循上述生成过程，生成的含噪标签更加合理，也更具挑战性。

**算法 1**噪声生成。

客户端数量 $K$；干净本地数据集 $\{D_k\}_{k=1}^{K}$；全局噪声率 $\rho$；本地噪声率分布参数 $\eta^l, \eta^u$。1：$\mathcal{I} =$ 从 $[K]$中随机选择 $\rho K$ 个元素。2：对于 $i \in \mathcal{I}$，执行 3：初始化一个网络 $f_i$。4：在本地数据集 $D_i = \{(x_j, y_j)\}_{j=1}^{N_i}$上训练 $f_i$。5：计算中所有样本的 $D_i$ 分类概率 $[0,1]^{N_i \times C}$ ( $Y \mid x$)。6：根据方程，计算中每个样本的误分类概率 $\widetilde{p}\,x \in ()$。7：$\eta_i \sim U(\eta^l, \eta^u)$。8：$\mathcal{N} =$ 以概率 $\widetilde{p}(x)/\sum \widetilde{p}(x)$从 $[N_i]$ 中随机选择 $\eta_i N_i$ 个元素。▷归一化 9：对于 $t \in \mathcal{N}$，执行 10：以概率 $p(Y \mid x_t)$从 $\mathcal{Y}$ 中随机选择一个不同的标签。11：将 $y_t$ 翻转为$y_t$。▷添加标签噪声 12：结束循环 13：结束循环 输出：添加标签噪声后的本地数据集 $\{D_k\}_{k=1}^{K}$。

## 4 方法论

为解决类别不平衡和标签噪声异质性问题，我们提出了一个名为 FedNoRo 的两阶段框架，如图 2 所示。第一阶段首先基于 FedAvg 训练预热模型（即 步骤 1 和 2），并用该模型计算每个客户端的逐类别损失值（即 步骤 3）。随后，服务器收集并处理各客户端的逐类别损失值，并通过高斯混合模型（GMM）识别噪声客户端（即 步骤 4 和 5）。根据第一阶段的客户端识别结果，第二阶段分别对干净客户端和噪声客户端采用不同的训练策略。具体来说，

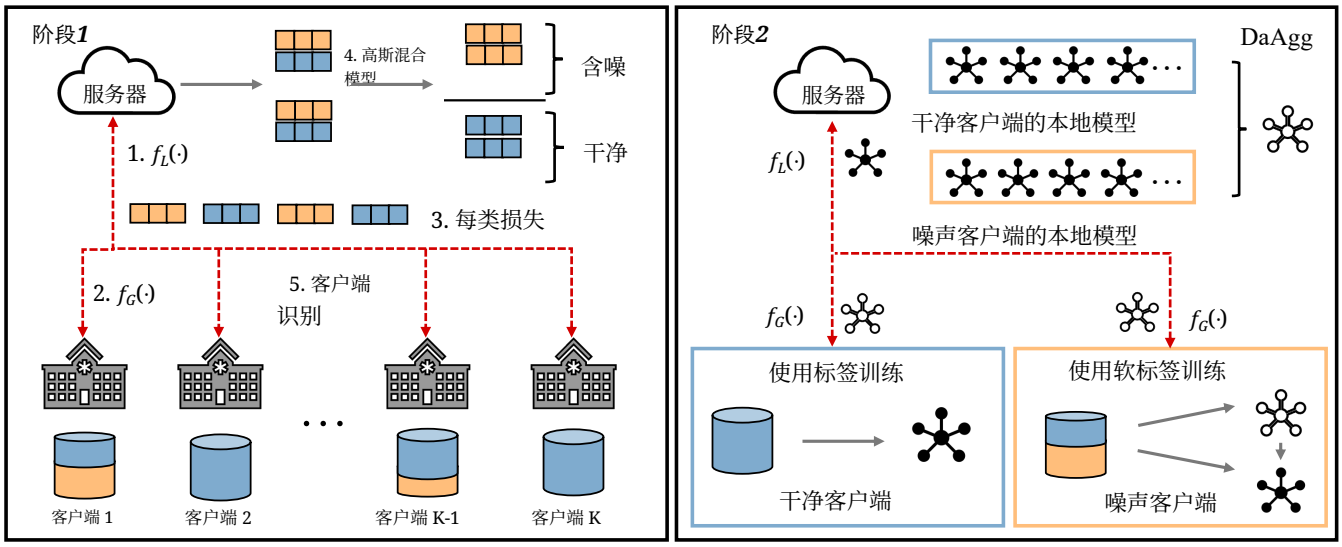



图2: 所提出的两阶段框架 Fed 概述 NoRo。

干净客户端采用标准交叉熵损失，而噪声客户端则采用知识蒸馏（KD），以更好地学习含噪标签。此外，我们定义了一个距离感知模型聚合函数，以减轻噪声客户端对全局模型更新的负面影响。需要指出的是，在两个阶段中，都会在本地训练中引入 logit 调整（LA） [Menon等， 2021] 以应对数据异质性和类别不平衡。

### 4.1 阶段1：噪声客户端检测

如第 1 节所述，类别不平衡破坏了干净客户端的客户端级平均指标（例如 损失值）服从独立同分布（IID）的规律，导致无法利用客户端级指标识别噪声客户端。为了解决这一问题，关键在于重新寻找一个适用于干净客户端的独立同分布指标，以便将噪声客户端识别为离群值。

为了识别噪声客户端，我们首先通过 FedAvg 训练一个预热模型，持续 $T_1$ 轮（将在 5.4 节讨论）。

$$w_g = \sum_{i=1}^{K} \frac{N_i}{\sum_{j=1}^{K} N_j} w_i, \tag{2}$$

其中， $w_g$ 和 $w_i$ 分别表示全局模型和第 $i$个本地模型的权重，$K$ 是参与聚合的客户端数量。随后，在本地训练阶段，采用标准交叉熵损失，并结合标签分配（LA）来训练本地模型，形式为 $f(\cdot) + \log \pi$，其中 $\pi$ 表示局部类别先验（即标签分布）。标签分配旨在使本地模型平等地对待每个类别，而不受偏置影响，不论本地数据中的类别先验是否存在偏置（即局部类别不平衡）。通过这种方式，不同本地模型的数据异质性得以缓解：无论其局部类别先验是否存在偏置，各模型都以统一的学习目标为导向，从而有助于模型聚合。此外，由于所有参与聚合的本地模型均不存在偏置，全局模型也更有可能不存在偏置。然而，研究证明，这种无偏的全局模型并不适用于通过客户端级平均指标识别含噪客户端 [Xu 等，2022年；Wang 等， 2022]。这是因为，尽管标签分配有助于减轻类别先验中的偏置，但类别特定特征学习难度方面的偏置仍未得到解决 [Yi 等， 2022]。

作为补充，本文计算每类指标，而不是使用所有类别样本的平均指标。这一方法基于我们在临床实践中的观察：尽管在类别不平衡的情况下，不同类别之间的样本并非独立同分布，但同一类别内部的样本具有更强的独立同分布特性。这是因为医学影像数据严格按照标准化流程采集。对于特定疾病，患者通常表现出相似的症状，且诊断所需的影像设备也受到严格限制，因此，即使在不同医院采集，图像也具有相同或相似的风格。换言之，通过分别关注每个客户端私有数据中的每个特定类别 $c$，即可满足独立同分布假设。因此，对于任意类别 $c$，可以有效地将类别 $c$ 中包含噪声样本的客户端识别为离群值（例如，具有较大的损失值）。此外，在标签噪声异质性的情况下，不同客户端中的噪声样本可能被分配到不同的噪声标签，仅依赖单个类别指标不足以识别噪声客户端。因此，采用每个客户端上所有类别的平均损失值 $i$，记为 $l_i = (l_i^1, l_i^2, \ldots\ldots l_i^C)^T \in \mathbb{R}^C$，用于识别。

给定 $K$每类平均损失值 $L = [l_1, \ldots, l_K]$，噪声客户端检测以无监督学习的方式实现。考虑到异构数据中的类别缺失问题，客户端 $i$中可能不存在特定类别 $c$，导致 $l_i$中缺少平均损失值 $l_i^c$。$L$ 中的此类缺失值无法用于无监督学习，必须进行填补。实际上，如果客户端 $i$中缺少类别 $c$，那么该类别在 $c$ 中必然没有噪声样本。根据小 -

损失技巧 [Han 等人， 2018], 样本越干净，就越容易学习，损失也越小。因此， $l_i^c$ 直接替换为所有客户端中类别 $c$ 的最小值（即，$\min_{i=1}^{K} l_i^c$）。考虑到不同类别的学习难度存在偏差，各类别的损失值范围可能差异很大，因此任何类别在所有客户端上的平均损失值都应统一归一化至 $[0,1]$，方法是

$$l_i^c = \frac{l_i^c - \min_i l_i^c}{\max_i l_i^c - \min_i l_i^c}. \tag{3}$$

为此，所有类别对识别噪声客户端的贡献相同，而不会被某些类别主导。随后，将双组分高斯混合模型（GMM）应用于处理后的客户端级每类损失向量，并将 $K$ 客户端划分为两个子集： $S_c$ （即干净客户端，对应于均值向量范数较小的高斯分布）和 $S_n$ （即噪声客户端，对应于均值向量范数较大的高斯分布）。需要注意的是，此噪声客户端检测过程具备隐私保护性，因为仅有平均损失值会上传至服务器。

### 4.2 阶段2：噪声鲁棒训练

检测出噪声客户端后，为实现噪声鲁棒联邦学习，分别针对干净客户端和噪声客户端提出了定制化的训练与聚合策略。与第一阶段类似，在本地训练阶段，对干净客户端和噪声客户端均引入标签分配（LA），以解决数据异质性和全局类别不平衡问题。对于干净客户端，采用标准交叉熵损失，基于干净标签训练每个本地模型。对于噪声客户端，受标签软化方法的启发——该方法在噪声标签学习中展现出巨大潜力 [Wei 等，2022a；Lukasik 等，2020；Lukov 等， 2022], ——采用一种基于知识蒸馏（KD） [Hinton等， 2015] 的训练方法。具体而言，给定任意 $x$ 及其由全局模型输出的 logit$f_G$（$x$），计算目标概率分布如下:

$$y_G = \mathrm{softmax}(\frac{f_G(x)}{T}), \tag{4}$$

其中， $T$ 是用于控制 $y_G$均匀性的温度参数，在本文中默认设为 0.8。然后，噪声客户端中每个本地模型的总训练损失写作。

$$\mathcal{L} = \lambda \mathcal{L}_{KL}(y_p, y_G) + (1-\lambda)\mathcal{L}_{CE}(y_p, \overline{y}), \tag{5}$$

其中， $y_p$ 表示本地模型的预测结果，$\mathcal{L}_{KL}$ 表示 Kullback-Leibler 散度， $\lambda$ 表示权衡系数。受半监督学习 [ Laine 和 Aila， 2017]的启发， $\lambda$ 采用高斯渐增曲线从 0 增长至 $\lambda_{max}$ （默认设为 0.8）。考虑到神经网络对含噪数据集的记忆效应 [Han 等，2018；Jiang 等，2022b], ，该策略对噪声客户端具有鲁棒性。在训练早期，每个本地模型倾向于学习简单模式（更可能对应干净样本），训练过程由硬标签 $y$ 和交叉熵损失主导。随着神经网络开始拟合复杂模式（更可能对应噪声样本），软标签 $y_G$ 逐渐主导训练过程，从而避免模型对噪声标签过拟合。

在模型聚合阶段，来自干净客户端和噪声客户端的模型具有不同的重要性。因此，受 [Liang 等的启发，提出了 2022], 距离感知模型聚合 DaAgg，其中客户端级距离度量定义为

$$d(i) = \min_{j \in S_c} \|w_i - w_j\|_2, \tag{6}$$

$w_i$ 表示第 $i$个本地模型的权重。该指标衡量模型 $w_i$ 与干净客户端的最近模型之间的距离。请注意，如果客户端$i$ 是干净的，则 $d(i)$等于0。为确保 $d(i)$有界，还需将其归一化为 $[0,1]$，即 $D(i) = {}^{d}{}_{(i)\max_j d(j)}$。然后，通过聚合本地模型来更新全局模型。

$$w_g = \sum_{i=1}^{K} \frac{N_i e^{-D(i)}}{\sum_{j=1}^{K} N_j e^{-D(j)}} w_i. \tag{7}$$

通过这种方式，任何干净客户端 $i$的聚合权重都因 $D(i) = 0$ 而保持不变，而噪声客户端的权重则会乘以 $[0,1]$中的缩放因子。此外，随着模型距离增大，权重会降低，从而在鲁棒联邦学习的全局模型更新过程中自适应地调整噪声客户端的贡献。

## 5 实验结果

### 5.1 数据集和评估指标

评估中使用了两个广泛采用的数据集 :

1. ICH 数据集 [Flanders 等人，2020] 由 67969 张脑部CT 切片组成，用于诊断五种颅内出血（ICH）亚型。该数据集按照 [Liu等人, 2021年；Jiang 等人, 2022a]进行预处理。
2. ISIC 2019数据集 [Tschandl 等人, 2018；Codella and others, 2018；Combalia等人，2019] 包含 25331 张图像，用于对八种皮肤疾病进行分类。

按照 7:3 的比例，两个数据集均被随机划分为训练集和测试集。由于两个数据集的类别分布都高度不平衡，划分后的训练集和测试集也存在类别不平衡，符合本文提出的设定。考虑到测试集类别不平衡，评估采用官方 ISIC Challenge 2019指标——平衡准确率（BACC）。

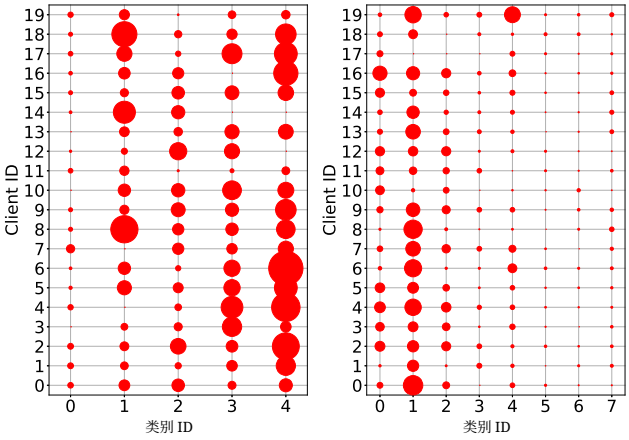
### 5.2 实验设置

**数据划分与噪声生成**

基于之前关于 FNLL [Xu 等人的工作， 2022],我们使用伯努利分布和狄利克雷分布生成异质数据划分。具体而言，对于 ICH 和 ISIC 2019 数据集，我们分别设置$\alpha_{Dir} = 2.0$， $\Phi_{ij} \sim$ 伯努利(0.9) $K = 20$和 $\alpha_{Dir}$ =1.5， $\Phi_{ij} \sim$伯努利(0.99) $K = 20$。为便于可视化，我们使用点的大小表示数据量，并绘制数据分布，如图 3 所示。为了评估不同标签噪声设置下的性能， $\rho$ 设为 0.2、0.3、0.4 和 0.6，且 ($\eta^l$， $\eta^u$) 设为 (0.3,0.5) 和 (0.5,0.7)，用于 ICH 数据集；而对于 ISIC 2019数据集， $\rho$ 设为 0.4 和 0.6，且 ($\eta^l$， $\eta^u$) 设为 (0.5,0.7)。

| 类别 | 方法 | $\rho$ | 0.0 | 0.2 | | 0.3 | | 0.4 | | 0.6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $(\eta^l, \eta^u)$ | (0.0, 0.0) | (0.3, 0.5) | (0.5, 0.7) | (0.3, 0.5) | (0.5, 0.7) | (0.3, 0.5) | (0.5, 0.7) | (0.3, 0.5) | (0.5, 0.7) |
| FL | FedAvg | Best | 69.34 | 65.81 | 64.53 | 62.49 | 60.82 | 60.52 | 58.38 | 58.46 | 54.77 |
| | | Last | 68.92 | 65.33 | 63.97 | 62.10 | 60.23 | 60.05 | 56.79 | 56.88 | 50.35 |
| | FedProx | Best | 68.16 | 64.58 | 63.64 | 61.81 | 61.91 | 60.85 | 58.50 | 60.21 | 57.57 |
| | | Last | 67.29 | 63.80 | 62.65 | 61.46 | 61.00 | 59.99 | 57.85 | 59.35 | 56.63 |
| | FedLA | Best | **73.56** | 69.45 | 69.28 | 66.84 | 64.84 | 66.60 | 62.39 | 63.90 | 58.78 |
| | | Last | **73.07** | 68.82 | 68.45 | 66.20 | 64.03 | 63.90 | 60.51 | 61.86 | 54.99 |
| 去噪联邦学习 | RoFL | Best | - | 42.57 | 40.42 | 40.95 | 39.87 | 40.35 | 35.60 | 40.18 | 35.47 |
| | | Last | - | 42.19 | 40.30 | 40.64 | 39.27 | 40.25 | 35.39 | 39.68 | 35.35 |
| | RHFL | Best | - | 57.48 | 56.91 | 56.72 | 55.74 | 55.26 | 54.30 | 54.63 | 51.00 |
| | | Last | - | 57.05 | 56.46 | 55.75 | 52.19 | 54.71 | 52.08 | 52.53 | 49.64 |
| | FedLSR | Best | - | 55.99 | 55.28 | 54.44 | 52.27 | 52.48 | 48.20 | 50.89 | 43.30 |
| | | Last | - | 55.69 | 54.77 | 53.95 | 51.70 | 51.92 | 46.77 | 49.96 | 39.81 |
| | FedCorr | Best | - | 57.90 | 56.68 | 55.02 | 54.61 | 53.62 | 50.82 | 50.89 | 47.13 |
| | | Last | - | 57.68 | 55.86 | 54.52 | 53.62 | 52.60 | 49.59 | 50.24 | 46.13 |
| 联合 | FedCorr+标签分配 | Best | - | 64.78 | 64.29 | 62.58 | 62.78 | 62.19 | 60.06 | 57.88 | 54.53 |
| | | Last | - | 63.99 | 63.55 | 61.90 | 61.87 | 61.30 | 59.58 | 57.17 | 53.92 |
| Ours | FedNoRo | Best | - | **70.59** | **70.64** | **70.14** | **69.35** | **70.69** | **69.30** | **67.55** | **63.83** |
| | | Last | - | **70.18** | **69.81** | **69.29** | **68.47** | **70.14** | **68.87** | **67.10** | **63.29** |

表 1：不同噪声率下 ICH 数据集的平衡准确率（%）定量比较结果。最佳结果以粗体标出。




(a) ICH (b) ISIC 2019

图3：两个数据集的数据分布可视化。

| 方法 | $\rho$ | 0.0 | 0.4 | 0.6 |
|---|---|---|---|---|
| | $(\eta^l, \eta^u)$ | (0.0, 0.0) | (0.5, 0.7) | (0.5, 0.7) |
| FedAvg | Best | 69.09 | 61.83 | 54.35 |
| | Last | 65.44 | 58.23 | 49.88 |
| FedProx | Best | 72.20 | 64.46 | 58.55 |
| | Last | **69.60** | 60.13 | 49.87 |
| FedLA | Best | **72.55** | 66.34 | 61.20 |
| | Last | 68.72 | 61.18 | 56.11 |
| RoFL | Best | - | 28.45 | 28.86 |
| | Last | - | 27.79 | 28.29 |
| RHFL | Best | - | 46.06 | 46.67 |
| | Last | - | 44.04 | 45.09 |
| FedLSR | Best | - | 30.15 | 27.24 |
| | Last | - | 29.11 | 26.08 |
| FedCorr | Best | - | 42.54 | 38.40 |
| | Last | - | 41.12 | 37.17 |
| FedCorr+标签分配 | Best | - | 60.38 | 55.40 |
| | Last | - | 59.16 | 54.27 |
| FedNoRo（我们的方法） | Best | - | **68.59** | **66.00** |
| | Last | - | **64.67** | **60.65** |

表 2：不同噪声率下 ISIC 2019数据集上的定量平衡准确率（%）结果。最佳结果以粗体标示。

**实现细节**

ResNet-18 [He 等人， 2016] 被用作 ICH 和 ISIC 2019数据集的骨干模型（由 ImageNet 预训练）。在联邦设置方面，ICH 数据集的最大通信轮次设为 100，本地轮数设为 5；ISIC 2019 数据集的最大通信轮次设为 300，本地轮数设为 1。两个数据集的 $T_1$ 预热训练分别设为 10 和 20。两个数据集的其他设置相同，包括采用动量项为 (0.9, 0.999) 的 Adam 优化器、权重衰减为 5e-4、恒定学习率为 3e-4，以及批量大小为 16。

## 5.3 与SOTA方法的比较

我们选取了一系列最先进的方法进行比较，包括传统联邦学习方法（即FedAvg[McMahan等， 2017], FedProx [Li 等，2020b], 和 FedLA [Menon 等， 2021]）以及去噪联邦学习方法（即RoFL [Yang 等， 2022], RHFL [Fang 和 Ye，2022],FedLSR [Jiang 等，2022b], 和FedCorr [Xu 等，2022]）。我们还将标签分配与 FedCorr 结合进行比较。

两个数据集上的定量结果汇总于表1和表2中，包括所有轮次中的最佳平衡准确率（记为最佳）以及最后10轮的平均平衡准确率（记为最后）。FedLA 通常在传统联邦学习方法中取得最佳性能，这得益于其对类别不平衡问题的缓解。与此同时，随着噪声率的增加，所有传统联邦学习方法的性能都会下降。遗憾的是，尽管四种去噪联邦学习方法专为噪声标签学习而设计，却未能实现性能提升。如果没有针对类别不平衡问题的特定解决方案，噪声标签学习将受到类别不平衡的严重影响，导致其性能甚至低于基线。这也解释了为什么将标签分配引入 FedCorr 是有益的。然而，正如前文所分析的，标签分配无法彻底消除类别不平衡问题，因此，仅将标签分配与 FedCorr 相结合并非最优方案。相比之下，FedNoRo 始终以较大优势优于其他方法。此外，随着噪声率的增加，FedNoRo 的性能表现更加稳定，充分体现了 FedNoRo 的抗噪特性。

| 指标 | LA | 每类 | 归一化 | 召回率（%） | 精确率（%） | MR（%） |
|---|---|---|---|---|---|---|
| ICH，$\rho = 0.3$，$(\eta^l, \eta^u) = (0.3, 0.5)$ | | | | | | |
| LID | ✗ | ✗ | ✗ | **100.00** | 54.54 | 0.00 |
| 置信度 | ✗ | ✗ | ✗ | 16.66 | 8.33 | 0.00 |
| Loss | ✗ | ✗ | ✗ | 94.58 | 57.12 | 0.00 |
| Loss | ✓ | ✗ | ✗ | **100.00** | 47.14 | 0.00 |
| Loss | ✗ | ✓ | ✗ | 97.19 | 90.63 | 85.77 |
| Loss | ✓ | ✓ | ✗ | 99.48 | 97.78 | 96.93 |
| Loss | ✓ | ✓ | ✓ | 99.70 | **98.76** | **98.28** |
| ICH，$\rho = 0.4$，$(\eta^l, \eta^u) = (0.3, 0.5)$ | | | | | | |
| LID | ✗ | ✗ | ✗ | 87.50 | 63.63 | 0.00 |
| 置信度 | ✗ | ✗ | ✗ | 37.50 | 25.00 | 0.00 |
| Loss | ✗ | ✗ | ✗ | 78.80 | 60.57 | 0.00 |
| Loss | ✓ | ✗ | ✗ | 89.41 | 80.46 | 0.00 |
| Loss | ✗ | ✓ | ✗ | 65.77 | **100.00** | 37.58 |
| Loss | ✓ | ✓ | ✗ | 81.10 | **100.00** | 47.09 |
| Loss | ✓ | ✓ | ✓ | **90.23** | **100.00** | **88.82** |

表 3：FedNoRo 第一阶段的消融研究。

### 5.4 消融研究

**含噪客户端检测各步骤的工作原理**

在 ICH 数据集上进行了额外实验，以分析 FedNoRo 第一阶段的每个步骤如何影响噪声客户端检测，结果如表 3 所示。这里，每类 表示分别计算每个类别的特定指标，而归一化 表示根据公式 3 对每类 指标进行最小-最大归一化。为公平比较，所有方法第一阶段的训练轮数均与第 5.2 节相同。在特定设置下，给定特定的客户端级指标后，我们反复采用高斯混合模型（使用 10000 个不同的随机种子）来识别噪声客户端。随后，使用 10000 次识别的平均召回率（即 Re）、精确率（即 Pr）和匹配率（即 MR：成功识别出的噪声客户端所占比例）来衡量该方法能否有效检测噪声客户端。

如表 3 所示， [Xu 等人、 2022] 和 [Wang 等人提出的前两种方法以及第三种方法完全忽略了类别不平衡，导致检测结果不佳。引入标签分配（LA）来解决类别不平衡问题，对含噪客户端检测有一定帮助，性能略有提升，但未能改善 MR 性能。相比之下，若独立考虑每个类别，噪声异质性和类别不平衡都能进一步缓解，从而提升 MR 性能。将标签分配（LA）与每类 方法及最小-最大归一化相结合，即 Fed NoRo，可以有效检测出大多数含噪客户端，为第二阶段奠定坚实基础。

**预热训练如何影响含噪客户端检测**

在第一阶段，全局模型先通过 FedAvg 进行 $T_1$ 轮预热，然后再进行含噪客户端检测。由于现实场景中缺乏关于含噪客户端的先验信息，因此很难准确确定

$T_1$ 以获得最佳性能。为评估 $T_1$的影响，我们在 ICH 数据集上针对 $T_1$ 的不同设置进行噪声客户端检测（$\rho = 0.4$，$(\eta^l, \eta^u)$=(0.3,0.5)），结果如图 4 所示。可以观察到，经过一定轮数的预热训练后，噪声客户端检测性能会稳定在某个范围内。换言之， $T_1$ 的设置不会显著影响

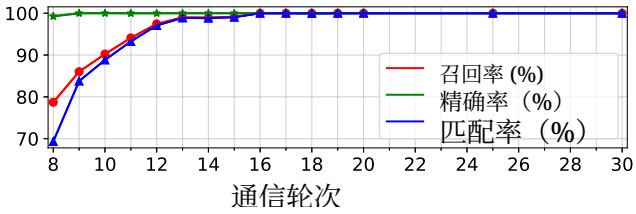



图 4： $T_1$ 预热训练的消融研究

g.

| 噪声客户端 | 去噪策略 | | Type | 平衡准确率 (%) | |
|---|---|---|---|---|---|
| | $\mathcal{L}_{KL}$ | DaAgg | | ICH | ISIC |
| ✓ | ✗ | ✗ | Best | 66.60 | 61.20 |
| | | | Last | 63.90 | 56.11 |
| ✗ | ✗ | ✗ | Best | 69.67 | 64.94 |
| | | | Last | 68.07 | 59.17 |
| ✓ | ✗ | ✓ | Best | 69.32 | 65.72 |
| | | | Last | 68.18 | 58.60 |
| ✓ | ✓ | ✗ | Best | 64.53 | 60.79 |
| | | | Last | 64.01 | 43.57 |
| ✓ | ✓ | ✓ | Best | **70.69** | **66.00** |
| | | | Last | **70.14** | **60.65** |

表4：FedNoRo 第二阶段的消融研究。设置： $\rho$=0.4 和 $(\eta^l, \eta^u)$=(0.3,0.5)，用于 ICH 数据集； $\rho$=0.6，$(\eta^l, \eta^u)$=(0.5,0.7)，用于 ISIC 数据集。

FedNoRo 的性能。

**How $\mathcal{L}_{KL}$ 和 DaAgg 如何助益噪声鲁棒训练**

FedNoRo 与多种去噪策略结合的定量结果汇总于表 4。若不进行去噪训练，低质量本地模型的噪声客户端必然会降低分类性能。尽管仅依赖干净客户端可以消除标签噪声，并在一定程度上提升分类性能，但这可能会错失其他客户端的宝贵信息，导致性能未达最优。存在噪声客户端时，以 DaAgg 替代 FedAvg 能有效降低噪声客户端在全局模型更新中的权重，从而提升性能。若存在噪声客户端，却仅在噪声客户端中采用 $L_{KL}$，效果较差，因为包含不准确软标签的噪声客户端可能会严重影响全局模型。相比之下，联合使用 $\mathcal{L}_{KL}$ 和 DaAgg，能够更好地挖掘噪声客户端中的宝贵信息，同时尽可能降低其负面影响，分类性能优于仅依赖干净客户端。

## 6 结论

本文提出了一种新的联邦标签噪声模型，其中全局数据呈类别不平衡，且不同客户端上的标签噪声具有异构性。与假设全局数据均衡的现有标签噪声模型相比，所提出的噪声模型更贴近现实，也更灵活，能够描述现实世界中的联邦学习场景。为此，我们进一步提出了一种名为 FedNoRo 的新型两阶段框架，用于实现抗噪声的联邦学习。在公开数据集上的实验结果表明，FedNoRo 的性能优于最先进的 FNLL 方法。我们相信，这种噪声模型及其解决方案都将启发未来的研究，推动开发适用于现实世界应用的联邦学习框架。

## 致谢

本研究部分得到中国国家自然科学基金项目（项目编号：62202179 和 62271220）的资助，部分得到中国湖北省自然科学基金项目（项目编号：2022CFB585）的资助，部分得到中国国家自然科学基金委员会/香港研究资助局联合研究计划项目（项目编号：N HKUST627/20）的资助，部分得到研究资助局 GRF 项目（项目编号：16203319）的资助。计算工作得到华中科技大学高性能计算平台的支持。

## 参考文献

[Agarwal等，2018] Naman Agarwal、Ananda Theertha Suresh、Felix Xinnan X Yu、Sanjiv Kumar 和 Brendan McMahan。cpSGD：通信高效且保护差分隐私的分布式 SGD。发表于*NeurIPS*，2018年。[Cao等，2019] Kaidi Cao、Colin Wei、Adrien Gaidon、Nikos Arechiga 和 Tengyu Ma。使用标签分布感知边际损失学习不平衡数据集。发表于*NeurIPS*，2019年。[Chen等，2020] Yiqiang Chen、Xiaodong Yang、Xin Qin、Han Yu、Biao Chen 和 Zhiqi Shen。FOCUS：应对联邦学习中的标签质量差异。*arXiv:2001.11359*，2020年。[Chen等，2021] Pengfei Chen、Junjie Ye、Guangyong Chen、Jingwei Zhao 和 Pheng-Ann Heng。超越类别条件假设：应对依赖实例标签噪声的初步尝试。发表于*AAAI*，第11442–11450页，2021年。[Codella 等，2018] Noel CF Codella 等。面向黑色素瘤检测的皮肤病灶分析：国际生物医学成像研讨会（ISBI）2017年挑战赛，由国际皮肤成像协作组织（ISIC）主办。发表于*ISBI*，第168–172页，2018年。[Combalia等，2019] Marc Combalia、Noel CF Codella、Veronica Rotemberg、Brian Helba、Veronica Vilaplana、Ofer Reiter、Cristina Carrera、Alicia Barreiro、Allan C Halpern、Susana Puig 等。BCN20000：真实环境中的皮肤镜病灶。*arXiv:1908.02288*，2019年。[Fang 和 Ye，2022] Xiuwen Fang 和 Mang Ye。面向含噪且异构客户端的鲁棒联邦学习。发表于*CVPR*，第10072–10081页，2022年。[Flanders等，2020] Adam E Flanders、Luciano M Prevedello、George Shih、Safwan S Halabi、Jayashree Kalpathy-Cramer、Robyn Ball、John T Mongan、Anouk Stein、Felipe C Kitamura、Matthew P Lungren 等。通过协作构建机器学习数据集：RSNA 2019 脑部 CT 出血挑战赛。*Radiology: Artificial Intelligence*，2(3)，2020年。[Han等，2018] Bo Han、Quanming Yao、Xingrui Yu、Gang Niu、Miao Xu、Weihua Hu、Ivor Tsang 和 Masashi Sugiyama。Co-teaching：利用极高噪声标签对深度神经网络进行鲁棒训练。发表于*NeurIPS*，2018年。[He等，2016] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。发表于*CVPR*，第770–778页，2016年。

[Hinton 等人，2015] Geoffrey Hinton、Oriol Vinyals、Jeff Dean 等人。《在神经网络中蒸馏知识》。arXiv:1503.02531，2015年。[Houle，2013] Michael E Houle。《维度、可区分性、密度与距离分布》。收录于 ICDMW，第 468–473 页，2013年。[Irvin 等人，2019] Jeremy Irvin 等人。《CheXpert：带不确定性标签并与专家比较的大型胸部 X 光影像数据集》。收录于 AAAI，第 590–597 页，2019年。[Jiang et al., 2022a] Meirui Jiang、Hongzheng Yang、Xiaoxiao Li、Quande Liu、Pheng-Ann Heng 和 Qi Dou。《面向类别不平衡半监督联邦图像诊断的动态银行学习》。收录于 MICCAI，第 196–206 页，2022年。[Jiang 等人，2022b] 姜雪峰、Sheng Sun、Yuwei Wang 和 Min Liu。《通过本地自正则化实现抵御噪声标签的联邦学习》。收录于 CIKM，第 862–873 页，2022年。[Kaissis 等人，2020] Georgios A Kaissis、Marcus R Makowski、Daniel R¨uckert 和 Rickmer F Braren。《医学影像中的安全、隐私保护与联邦机器学习》。Nature Machine Intelligence，2(6):305–311，2020年。[Kang 等人，2020] Bingyi Kang、Saining Xie、Marcus Rohrbach、Zhicheng Yan、Albert Gordo、Jiashi Feng 和 Yannis Kalantidis。《解耦表征与分类器以实现长尾识别》。收录于 ICLR，2020 年。[Karimireddy 等人，2020] Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。《SCAFFOLD：用于联邦学习的随机受控平均》。收录于 ICML，第 5132–5143 页，2020年。[Karthik 等人，2021] Shyamgopal Karthik、J´erome Revaud 和 Boris Chidlovskii。《从含噪标签的长尾数据中学习》。arXiv:2108.11096，2021年。[Koneˇcn`y 等人，2016] Jakub Koneˇcn`y、H Brendan McMa-han、Felix X Yu、Peter Richt´arik、Ananda Theertha Suresh 和 Dave Bacon。《联邦学习：提升通信效率的策略》。arXiv:1610.05492，2016年。[Laine 和 Aila，2017] Samuli Laine 和 Timo Aila。《用于半监督学习的时间集成》。收录于 ICLR，2017年。[Li 和 Zhan，2021] Xin-Chun Li 和 De-Chuan Zhan。《FedRS：针对标签分布非独立同分布数据的受限 softmax 联邦学习》。收录于 KDD，第 995–1005 页，2021年。[Li 等人，2020a] Junnan Li、Richard Socher 和 Steven CH Hoi。《DivideMix：将含噪标签学习作为半监督学习》。收录于 ICLR，2020年。[Li 等人，2020b] Tian Li、Anit Kumar Sahu、Manzil Zaheer、Maziar Sanjabi、Ameet Talwalkar 和 Virginia Smith。《异构网络中的联邦优化》。收录于 MLSys，第 429–450 页，2020年。

[Li 等人，2021a] Qinbin Li、Bingsheng He 和 Dawn Song。《模型对比联邦学习》。收录于 CVPR，第 10713–10722 页，2021年。[Li 等人，2021b] Xiaoxiao Li、Meirui Jiang、Xiaofei Zhang、Michael Kamp 和 Qi Dou。《FedBN：通过本地批归一化处理非独立同分布特征的联邦学习》。收录于 ICLR，2021年。[Liang 等人， 2022] Xiaoxiao Liang、Yiqun Lin、Huazhu Fu、Lei Zhu 和 Xiaomeng Li。《RSCFed：随机采样共识联邦半监督学习》。收录于 CVPR，第 10154–10163 页，2022年。[Liu 等人， 2021] Quande Liu、Hongzheng Yang、Qi Dou 和 Pheng-Ann Heng。《通过客户端间关系匹配实现联邦半监督医学图像分类》。收录于 MICCAI，第 325–335 页，2021年。[Lukasik 等人， 2020] Michal Lukasik、Srinadh Bhojana-palli、Aditya Menon 和 Sanjiv Kumar。《标签平滑能否缓解标签噪声？》。收录于 ICML，第 6448–6458 页，2020年。[Lukov 等人， 2022] Tohar Lukov、Na Zhao、Gim Hee Lee 和 Ser-Nam Lim。《通过软标签平滑训练以缓解面部表情中的噪声标签》。收录于 ECCV，第 648–665 页，2022年。[McMahan 等人， 2017] Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Ar-cas。《基于去中心化数据的深度网络通信高效学习》。收录于 AISTATS，第 1273–1282 页，2017年。[Menon 等人， 2021] Aditya Krishna Menon、Sadeep Jaya-sumana、Ankit Singh Rawat、Himanshu Jain、Andreas Veit 和 Sanjiv Kumar。《通过 logit 调整进行长尾学习》。收录于 ICLR，2021年。[Sattler 等人， 2019] Felix Sattler、Simon Wiedemann、Klaus-Robert M¨uller 和 Wojciech Samek。《基于非独立同分布数据的鲁棒且通信高效的联邦学习》。IEEE神经网络与学习系统汇刊，31(9):3400–3413，2019年。[Shang 等人， 2022] Xinyi Shang、Yang Lu、Gang Huang 和 Hanzi Wang。《通过基于联邦特征的分类器重新训练，实现异构长尾数据上的联邦学习》。收录于 IJCAI，2022年。[Shen 等人， 2022] Zebang Shen、Juan Cervino、Hamed Has-sani 和 Alejandro Ribeiro。《一种应对类别不平衡的无偏联邦学习方法》。收录于 ICLR，2022年。[Tschandl 等人， 2018] Philipp Tschandl、Cliff Rosendahl 和 Harald Kittler。《HAM10000数据集：汇集多来源常见色素性皮肤病变皮肤镜图像的大型数据集》。Scientific Data，5(1):1–9，2018年。[Wang 等人， 2019] Yisen Wang、Xingjun Ma、Zaiyi Chen、Yuan Luo、Jinfeng Yi 和 James Bailey。《用于含噪标签鲁棒学习的对称交叉熵》。收录于 ICCV，第 322–330 页，2019年。[Wang 等人， 2021] Lixu Wang、Shichao Xu、Xiao Wang 和 Qi Zhu。《解决联邦学习中的类别不平衡问题》。收录于 AAAI，第 10165–10173 页，2021年。[Wang 等人， 2022] Zhuowei Wang、Tianyi Zhou、Guodong Long、Bo Han 和 Jing Jiang。FedNoiL：一种用于含噪标签联邦学习的简单两级采样方法。arXiv:2205.10110，2022年。[Wei 等人，2022a] Jiaheng Wei、Hangyu Liu、Tongliang Liu、Gang Niu、Masashi Sugiyama 和 Yang Liu。平滑还是不平滑？当标签平滑遇到噪声标签。载于 ICML，2022年。[Wei 等人，2022b] Qi Wei、Haoliang Sun、Xiankai Lu 和 Yilong Yin。Self-Filtering：一种通过置信度惩罚进行噪声感知样本选择的方法，用于处理标签噪声。载于 ECCV，第 516–532 页，2022年。[Xu 等人， 2022] Jingyi Xu、Zihan Chen、Tony QS Quek 和 Kai Fong Ernest Chong。FedCorr：用于标签噪声校正的多阶段联邦学习。载于 CVPR，第 10184–10193 页，2022年。[Yan 等人， 2020] 闫增强、Jeffry Wicaksana、Zhiwei Wang、Xin Yang 和郑光廷。基于多源分散式医疗图像数据的变异感知联邦学习。《IEEE 生物医学与健康信息学汇刊》，25(7):2615–2628，2020年。[Yang 等人， 2021] Miao Yang、Hua Qian、Ximin Wang、Yong Zhou 和 Hongbin Zhu。含噪标签联邦学习中的客户端选择。《IEEE 车辆技术汇刊》，71(2):2193–2197，2021年。[Yang 等人， 2022] Seunghan Yang、Hyoungseob Park、Junyoung Byun 和 Changick Kim。含噪标签下的鲁棒联邦学习。《IEEE Intelligent Systems》，37(2):35–43，2022年。[Yi等人， 2022] Xuanyu Yi、Kaihua Tang、Xian-Sheng Hua、Joo-Hwee Lim 和 Hanwang Zhang。识别长尾样本分布中的难以辨别噪声。载于 ECCV，第 739–756 页，2022年。[Yu 等人， 2019] Xingrui Yu、Bo Han、Jiangchao Yao、Gang Niu、Ivor Tsang 和 Masashi Sugiyama。分歧如何帮助提升应对标签污染的泛化能力？载于 ICML，第 7164–7173 页，2019年。[Zhang 等人， 2017] Hongyi Zhang、Moustapha Cisse、Yann N Dauphin 和 David Lopez-Paz。Mixup：超越经验风险最小化。arXiv:1710.09412，2017年。[Zhao 等人， 2022] Ganlong Zhao、Guanbin Li、Yipeng Qin、Feng Liu 和 Yizhou Yu。中心性与一致性：针对依赖实例噪声标签学习的两阶段干净样本识别。载于 ECCV，第 21–37 页，2022年。[Zheng 等人， 2021] Wenbo Zheng、Lan Yan、Chao Gou 和 Fei-Yue Wang。用于欺诈信用卡检测的联邦元学习。载于 IJCAI，第 4654–4660 页，2021年。[Zhou 等人， 2020] Boyan Zhou、Quan Cui、Xiu-Shen Wei 和 Zhao-Min Chen。BBN：用于长尾视觉识别的双分支累积学习网络。载于 CVPR，第 9719–9728 页，2020年。[Zhu 等人， 2021] Zhuangdi Zhu、Junyuan Hong 和 Jiayu Zhou。面向异构联邦学习的无数据知识蒸馏。载于 ICML，第 12878–12889 页，2021年。

# 附录

## 设置比较

由于我们针对所提出的 FNLL 问题开展实验，因此与以往研究（例如 FedCorr [Xu 等人， 2022]）相比，我们的实验设置有所不同，如表 5 所示。

| | 全局类别不平衡 | 数据异质性 | 标签噪声异质性 |
|---|---|---|---|
| FedCorr | ✗ | ✓ | ✗ |
| Ours | ✓ | ✓ | ✓ |

表 5：与先前研究中的实验设置进行比较。

## B 数据划分

为模拟数据异质性，我们参照先前的 FNLL 研究 [Xu 等， 2022] 将异质数据分配给不同客户端。具体而言，首先创建一个矩阵 $\Phi \in \{0,1\}^{K\times C}$，其中每个元素 $\Phi_{i,\ j}$ 均从逐元素伯努利分布中采样，用于指示客户端 $i$ 是否拥有类别 $j$ 的样本（1 表示是）。然后，对于每个类别 $c$，使用参数 $\alpha_{Dir}$ 确定的狄利克雷分布生成概率分布，以便将类别 $c$ 的样本分配给拥有类别 $c$样本的客户端。

## C 对比方法的实现细节

我们简要介绍实验中对比方法的一些重要实现细节。

- FedAvg [McMahan 等， 2017]：本地训练阶段采用普通交叉熵损失。
- FedProx [Li 等，2020b]：近端项 $\mu_{prox}$ 在 ICH 和 ISIC2019 上分别设为 0.01 和 0.1。
- FedLA [Menon 等人， 2021]：在本地训练阶段采用 logit 调整。
- RoFL [Yang 等人， 2022]： $T_{pl}$ 设置为 10，而其他参数与开源代码相同。
- RHFL [Fang 和 Ye， 2022]：由于实验缺乏用于知识蒸馏的公开数据集，因此将最初为知识蒸馏设计的客户端置信度重加权与 FedAvg 相结合。

- FedLSR [Jiang 等人，2022b]： $t_w$ 对于 ICH 和 ISIC2019，分别设为 20 和 60。其他参数与开源代码相同。

- FedCorr [Xu 等，2022]： $T_1$， $T_2$， $T_3$ 对于 ICH 和 ISIC2019，分别设为 2、50、50 和 5、150、150。第一阶段每轮有一个在线客户端，最后两个阶段的每轮中所有客户端均在线。

- FedCorr+LA：在 FedCorr 的本地训练阶段采用 logit 调整。

为确保公平比较，对于每个数据集，所有方法均采用上述未特别说明的设置，例如，骨干模型、通信轮次、本地轮数和优化器（见第 5.2 节）。请注意，为简化实验设置，除 FedCorr 的第一阶段外，所有方法的客户端在每轮中均保持在线；这是由于 FedCorr 的特殊设计。为确保 FedCorr 得到充分训练，其第一阶段不计入总通信轮次。